Translation of neutrally evolving peptides provides a basis for *de novo* gene evolution

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: pervasive translation and de novo genes

Keywords: ribosome profiling, translation, *de novo* gene, long non-coding RNA, peptide, polymorphism, natural selection

This is a post-peer-review, pre-copyedit version of an article published in *Nature ecology & evolution*. The final authenticated version is available online at: http://dx.doi.org/10.1038/s41559-018-0506-6

Abstract

There is accumulating evidence that some protein-coding genes have originated *de novo* from previously non-coding genomic sequences. However, the processes underlying *de novo* gene birth are still enigmatic. In particular, the appearance of a new functional protein seems highly improbable unless there is already a pool of neutrally evolving peptides that are translated at significant levels and that can at some point acquire new functions. Here we use deep ribosome profiling sequencing data, together with proteomics and single nucleotide polymorphism information, to search for these peptides. We find hundreds of open reading frames that are translated, and that show no evolutionary conservation and no selective constraints. The data suggests that the translation of these neutrally evolving peptides may be facilitated by the chance occurrence of open reading frames with a favorable codon composition. We conclude that the pervasive translation of the transcriptome provides plenty of material for the evolution of new functional proteins.

Introduction

A large fraction of the genome is transcribed; this includes functional genes but also thousands of transcripts that are not conserved across species and which show weak or no signatures of natural selection[1–3]. Many of the latter transcripts are annotated as long non-coding RNAs (lncRNAs) because they lack conserved long open reading frames (ORFs). Recent studies based on the sequencing of ribosome-protected RNA fragments (ribosome profiling) have reported that a surprisingly large number of these transcripts are associated with ribosomes and may translate small proteins or peptides[4–9]. This is intriguing because only a small fraction of them are likely to encode functional micropeptides[10].

Each ribosome profiling experiment generates millions of ribosome footprints that are subsequently mapped to the genome or the transcriptome to identify open reading frames (ORFs) that are being translated[11]. The codon-by-codon movement of the ribosome along the coding sequence results in a characteristic pattern of three nucleotide periodicity of the mapped reads, which makes ribosome profiling a very useful method to detect novel events of translation[4,12,13]. Given enough sequence coverage the technique can uncover low-abundant small proteins that would be otherwise difficult to detect by standard proteomics approaches[14,15].

There is increasing evidence that some functional genes have originated *de novo* from previously non-functional parts of the genome[16–20]. There are many protein-coding genes that appear to be species-species; the encoded proteins are shorter than average and they tend to evolve under more relaxed constraints than other genes[21,22,20]. There is experimental evidence that at least some *de novo* genes are functional[23,24]. However, for a new protein to acquire a function it first needs to be produced in the cell at significant amounts. Some of the new proteins may be generated by the translation of randomly occurring ORFs in transcripts currently annotated as non-coding[7]. As these proteins are not expected to be functional at first, they are expected to evolve under no significant purifying selection, or neutrally. We previously showed that, as a whole, putatively translated lncRNAs and young annotated protein-coding genes

shared a number of similarities, such as small ORF size and weak selective constraints, compared with more widely conserved genes[8]. This suggested that the number of proteins with a recent evolutionary origin was larger than previously thought. However, it remained to be determined if there existed a pool of young proteins that were translated at significant levels and evolved in a neutral manner, as hypothesized for *de novo* gene evolution.

To assess the functional relevance of translated products one can use the ratio between the number of non-synonymous and synonymous substitutions in the putative coding sequences[4,5]. However, this method requires an alignment of at least two different homologous sequences. A more general approach that can be used in the absence of homologues in other species is the ratio between the number of non-synonymous and synonymous single nucleotide polymorphisms, compared to the one expected under neutrality. Under no selection, non-synonymous and synonymous polymorphisms persist at the same levels, whereas under purifying selection there is a deficit of non-synonymous polymorphisms because some amino acid changes disrupt the protein's function[25]. Single nucleotide polymorphism analysis can be performed on a gene-by-gene basis or in pools of sequences that share certain features[2,26].

Here we employ phylogenetic comparisons, ribosome profiling data and single nucleotide polymorphism information to search for putative precursors of *de novo* genes in mouse. The study renders visible a layer of protein expression that produces non-functional small proteins, filling a gap in our understanding of the processes underlying *de novo* gene birth.

Results

Many ORFs outside annotated protein-coding sequences are translated

First we set out to identify translated open reading frames (ORFs) in mouse protein-coding genes (codRNAs) and long non-coding RNAs (lncRNAs) using ribosome-profiling RNA-sequencing (Ribo-Seq) data from eight different tissues and cell lines (Supplementary Table 1 and references therein). In contrast to RNA sequencing (RNA-Seq) reads, which are expected

to cover the complete transcript, Ribo-Seq reads correspond to regions bound by ribosomes. We mapped the RNA-Seq and Ribo-Seq reads to the mouse Ensembl gene annotations and, for the sake of completeness, also to a set of previously obtained novel mouse transcripts that did not correspond to annotated protein-coding genes or lncRNAs[3].

We used the RibORF program[4] to identify translated sequences among ORFs covered by at least 10 Ribo-Seq reads in transcripts expressed in one or more tissues (Fig. 1a and Supplementary Table 1). This program calculates a score for each ORF depending on the 3-nucleotide periodicity and uniformity of the mapped reads. Using a highly stringent RibORF score cut-off of 0.7[4] we found that about 90% of the coding genes (15,020), and 20% of the annotated lncRNAs (539), were predicted to be translated in at least one sample. Additionally, we identified 286 novel, non-annotated, expressed loci that also contained translated ORFs (Fig. 1b). Annotated lncRNAs and novel transcripts mostly encoded small proteins (Fig. 1c) and the two classes were merged.

Hundreds of translated ORFs have no homologues in other species and evolve neutrally

Putative precursors of *de novo* genes should be of recent origin and evolve under no purifying selection. We performed exhaustive sequence similarity searches of the translated ORFs against high coverage transcriptomes from human and rat as well as against the annotated proteomes of 101 different eukaryotic species (Fig. 2a, Supplementary Table 2 for a list of species, see Methods for more details). For these searches we discarded any peptides shorter than 24 amino acids, as the detection of homologues may be compromised in such cases due to lack of sufficient sequence information. We identified 1,980 translated ORFs that showed no homology to expressed sequences in other species (class non-conserved or NC). This represented about 10.4% of the translated ORFs.

Next, we measured the strength of selection acting on the putative proteins using a large collection of mouse single nucleotide polymorphisms (SNPs) for the house mouse subspecies *Mus musculus castaneus*[27]. We could map a total of 324,729 SNPs to the set of translated

ORFs. We calculated the ratio between the number of observed non-synonymous and synonymous SNPs (PN/PS(obs)) in conserved and non-conserved groups of ORFs, and normalized it by the same ratio expected under neutrality (PN/PS(exp)). The expected PN/PS was estimated using a table of nucleotide mutation frequencies in *Mus musculus castaneus* and the observed codon frequencies in each set of sequences of interest (Supplementary Tables 3 to 5). This allowed us not only to compare the strength of selection across different sets of sequences, as done in a previous study of ORFs translated from lncRNAs[8], but also to discard purifying selection if the normalized PN/PS was not significantly different from 1. Specifically, we used a chi-square test that compared the number of observed and expected non-synonymous and synonymous SNPs in each sequence set. The normalized PN/PS of conserved ORFs was around 0.15 (Fig. 2b, p-value < 10^{-5}), consistent with protein functionality. One example in this group was Stannin[28,29], a highly conserved peptide that regulates neuronal cell apoptosis (Fig. 3).

In the case of the non-conserved ORFs, the normalized PN/PS depended on the codon usage bias or coding score (Supplementary Fig. 1 and 2). The coding score provides a measurement of how similar the relative codon frequencies are to those observed in functional coding sequences[20,30,21]. We divided the non-conserved ORFs in two groups: the first group, with high coding scores (≥ 0.1014, NC-H), showed a weak but significant signature of selection (Fig. 2b, p-value < 0.05); the second group, comprising the rest of ORFs, showed no significant deviation from neutrality (NC-L)(see Fig. 3 for a specific example). The fist group is likely to contain some species-specific functional genes, whereas the second group, the neutrally evolving ORFs, represents new translated peptides with no function.

Neutrally evolving ORFs were found both in lncRNAs but also in many protein-coding genes (Fig. 2b). In lncRNAs they were usually found in isolation or with other non-conserved ORFs, whereas in coding genes they were in most cases located in alternative transcript isoforms of genes already containing a conserved ORF. One possible way the secondary ORFs in protein-coding genes could become new genes would be through gene duplication and inactivation of the main ORF. The ORFs in lncRNAs and coding genes were merged into a single group,

comprising 1,291 ORFs, for subsequent analyses (neutral ORFs).

Validation of the set of neutrally evolving peptides

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, 41 of the ORFs in the neutral set contained 10 or more SNPs, and we decided to compute a normalized PN/PS ratio for these individual cases. The median PN/PS of these ORFs was around 1 and the distribution of PN/PS values was very different from that observed for conserved ORFs (Supplementary Fig. 3, Wilcoxon test, p-value < 10^{-5}), consistent with the previous results. Finally, we quantified the number of ORFs that contained SNPs that generated premature stop codons, truncating more than half of the ORF, in the set of neutrally evolving ORFs and in the set of conserved ORFs. In the first case we found 72 out of 1,291 ORFs that contained this type of mutation (5.6%) and in the second case 296 out of 17,043 ORFs (1.74%). Considering that neutral ORFs are in general much shorter than conserved ORFs (median protein size 44 *versus* 412 amino acids), and thus less likely to accumulate ORF-truncating mutations by chance alone, the data clearly indicates a strong excess of ORF-truncating SNPs in neutral ORFs with respect to conserved ORFs.

We next calculated which was the percentage of neutral ORFs in the transcripts that were detected as translated by our pipeline. This value was 30.5% (1,291 out of 4,232), about one order of magnitude larger than the false positive rate estimated from two negative controls. The first control was a set of ORFs derived from small nuclear and nucleolar RNAs (3.33% false positive rate, see example in Fig. 3), the second one was a set of ORFs in which the position of the reads had been randomized (4.16% false positive rate; see Methods for more details).

We used proteomics data from the PRIDE database[31] to search for additional evidence of the neutrally evolving peptides. Despite their small size (median 44 amino acids), a limiting factor for their detection by standard proteomics-based techniques[32], we found proteomics evidence for 32 of the translated ORFs in this group. This represents 2.5% of the proteins in this set

(compared to less than 0.2% false positive rate, see Methods). This fraction is similar to the one obtained for conserved proteins subsampled to have a similar size distribution as the neutral ORFs (2.9%; in contrast, about 41% of all conserved ORFs have proteomics evidence). The subset of ORFs with proteomics evidence did not deviate significantly from neutrality according to the PN/PS-based test (Supplementary Table 5).

What determines that some ORFs, but not others, are translated?

Some transcripts contained relatively long ORFs but were not translated. One example of this sort was the previously described *de novo* non-coding gene *Poldi*[33] that lacked any evidence of translation in the data we analyzed. We next asked which factors may influence the translation of neutral ORFs. First, we inspected the translation initiation sequence context but did not detect any clear differences between translated and non-translated neutral ORFs (Supplementary Fig. 4). 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 codons[34]. Consistent with this hypothesis, we found that the translated neutral ORFs exhibited higher coding scores than non-translated ORFs (Fig. 4a, Translated versus non-translated Wilcoxon test, p-value < 10^{-5}). Importantly, the differences were maintained after subsampling the genes so that the two sets had the same gene expression distribution (Fig. 4b, Wilcoxon test, p-value < 10^{-5}). This is consistent with codon composition having an effect *per se* in ORF translation. When controlling by coding score, expression level, but not ORF length, was positively related to the detection of translation (Fig. 4c).

It has been previously hypothesized that what distinguishes translated from non-translated lncRNAs is the relative amount of the lncRNA in the cytoplasm with respect to the nucleus[4]. However, we found evidence that some lncRNAs with nuclear functions, such as *Malat1* and *Neat1*, translated neutral ORFs, suggesting that the cytosolic fraction of a transcript can be translated independently of its role or preferred location. Instead, we found that neutral ORFs enriched in codons that are abundant in functional protein-coding genes could be translated

more efficiently than other ORFs. This is consistent with the observation that abundant codons enhance translation elongation[35], whereas rare codons might affect the stability of the mRNA[36]. These differences could result in compositional biases in newly emerged proteins.

Discussion

The molecular mechanisms underlying *de novo* gene evolution are still poorly understood[17,18,37]. The sudden appearance of a new protein-coding gene from a genomic segment seems *a priori* highly improbable, but the process becomes much more likely if the genome is already being pervasively transcribed and translated outside functional protein-coding sequences. An excess of transcription was already noted in the first large-scale cDNA sequencing efforts performed in human and mouse[38], and more recent studies have found a high rate of genome transcriptional turnover when comparing closely related species[39]. Other works have found that many open reading frames outside annotated protein-coding sequences can be translated[4,5,7], raising the possibility that they are an important source of *de novo* genes[9,20,8].

The precursors of new functional proteins are expected to be translated at relatively high levels and evolve under no purifying selection. Here we have detected hundreds of small proteins that appear to be of recent evolutionary origin and that fulfill these properties. Their real number could be much higher considering that many cell types and tissues have not yet been sampled. According to recent estimations, the cost of transcription and translation in multicellular organisms is probably too small to overcome genetic drift[40]. Therefore, if the peptides do not affect the fitness *per se*, these activities may be effectively neutral. Our results are in line with this prediction.

It has been proposed that nascent proteins are pre-adapted in the sense that only peptides that do not have a harmful effect will be expressed at significant levels[9], and that this may explain the observed compositional differences between young proteins and intergenic sequences[41]. Another possible source of bias is the codon composition of the sequence. According to our results, ORFs that have a high coding score may be more efficiently translated than those that

have a lower coding score. As a result, new proteins may have a more more coding-like composition than the average random or intergenic sequence.

The process of *de novo* gene origination involves the gain of a useful function by a previously non-functional sequence. The rate at which this happens remains to be determined but it has been observed that many artificially generated peptides can function as secretion signals[42], and selection for ATP-binding activity in a library of randomly generated 80 amino acid polypeptides successfully identified several candidates capable of binding to ATP[43]. Recent experiments performed in *E.coli* also suggest that random expressed sequences can often affect cellular growth[44].

We have shown that the transcriptome is pervasively translated, and that this generates novel peptides that evolve under no constraints. This set of translated peptides, and their sequences, are expected to change rapidly during evolution, generating a myriad of opportunities for the birth of new functional genes.

Methods

Transcript assembly

We used strand-specific polyA+ RNA sequencing data (RNA-Seq) data from different mouse and human tissues to assembly the species transcriptomes (Gene Expression Omnibus mouse GSE69241[3], GSE43721[45], and GSE43520[46]; human GSE69241[3]). The mouse RNA samples were extracted from strain Balb/C. RNA-Seq reads were filtered by length (> 25 nucleotides) and by quality using Condetri (v.2.2)[47] with the following settings: -hq = 30 –lq = 10. We aligned the reads to the corresponding reference species genome with Tophat (v. 2.0.8, –N 3, -a 5 and –m 1)[48]. Multiple mapping to several locations in the genome was allowed unless otherwise stated. We assembled the transcriptome with Stringtie[49], 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 excluded lncRNAs that overlapped annotated pseudogenes or that showed significant sequence similarity to known protein-coding sequences (BLASTP, e-value < 10^{-4}). In the case of rat we employed a previously generated transcript assembly[50].

Ribosome profiling data

We used ribosome profiling data (Ribo-Seq) from 8 different mouse tissues or cell lines (see Supplementary Table 1), obtained from Gene Expression Omnibus under accession numbers GSE51424[51], GSE50983[52], GSE22001[53], GSE62134[54], GSE72064[55], and GSE41246. Only datasets corresponding to non-pathogenic conditions were considered. The reads from experimental replicates were merged before using RibORF to increase the resolution of the read periodicity, as done in the original RibORF paper[4]. For all analyses we considered only genes expressed at significant levels in at least one sample (RNA-Seq fragments per kilobase per Million mapped reads (FPKM) > 0.2). The expression of the genes detected in these samples is expected to be highly representative of the *Mus musculus* species as a whole. We mapped several brain RNA-Seq datasets from *Mus musculus castaneus*[39] to the mouse assembled transcriptome using NextGenMap[56]. As expected, the vast majority of the genes expressed in brain samples from C57BL/6 mice[51] also showed evidence of expression in *Mus musculus castaneus* brain RNA samples[39](Supplementary Table 6).

We discarded anomalous reads (length < 26 or > 33 nt) and reads that mapped to annotated rRNAs and tRNAs in mouse from the Ribo-Seq sequencing datasets. Next, reads were mapped to the assembled mouse genome (mm10) with Bowtie (v. 0.12.7, parameters -k 1 -m 20 -n 1 --best --strata). Considering that the ORFs had to be extensively covered by reads to be considered translated (high uniformity), we decided to include multiple mapped reads. This is expected to increase the sensitivity to detect translation events, especially in gene families, but also the number of false positives. We observed that using multiple mapped reads increased the periodicity and uniformity values of a number of protein-coding genes, but that it had little effect on translated ORFs from lncRNAs (Supplementary Fig. 5). 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 (second binding site for tRNA in the ribosome) for reads of different size, as previously described[11,13].

Identification of translated ORFs

We predicted all translated ORFs (ATG to STOP) with a minimum length of 9 amino acids in the transcripts with RibORF (v.0.1)[4]. Only ORFs with a minimum of 10 mapped Ribo-Seq reads were considered. The RibORF classifier is based on a support vector machine algorithm, originally applied to human transcripts. The input parameters are the 3-nucleotide periodicity and the uniformity of the reads. The periodicity refers to the distribution of the reads in the three different possible frames; two values are computed, f1, which is the fraction of reads that correspond to the correct frame, and f2, which is the fraction of reads in position +1 of the correct frame. The uniformity corresponds to the percentage of maximum entropy, a value of 1 indicates a completely even distribution of reads across regions. For each ORF the program computes a score that depends on the periodicity and the uniformity[4]. We used the same score cut-off as in the original paper (≥ 0.7), which had a reported false positive rate of 0.67% and a false negative rate of 2.5%.

We eliminated any redundancy in the translated ORFs by taking the longest ORF when several overlapping translated ORFs were detected in the same gene. The identification of translated ORFs was done separately for the different tissues (Supplementary Table 1). Differences in the number of translated ORFs in different tissues were related to the depth of sequencing and the number of reads that mapped to the most highly expressed proteins (Supplementary Fig. 6 and 7, respectively). When translation of the same ORF was detected in several tissues, we kept the information for the tissue with the highest RibORF score.

Sequence conservation

We searched for mouse ORF homologues in the human and rat transcriptomes using TBLASTN

(limited to one strand, e-value < 10^{-4})[57]. We also performed sequence similarity searches against the annotated proteomes of 67 mammalian-species and 34 non-mammalian eukaryotes from a diverse range of groups compiled in a previous study[50], using BLASTP (e-value < 10^{-4}). For these searches we only considered query proteins of size 24 amino acids or longer, as shorter proteins may not contain sufficient information to perform homology searches. Mouse ORFs that did not have any homology hits in other species were classified as non-conserved, the rest as conserved. Non-conserved ORFs located upstream or downstream of another longer ORF in a conserved transcript (uORFs and dORFs) were excluded from this analysis.

We inspected the rat genomic syntenic regions of translated ORFs using LiftOver[58]. We classified the ORFs in two groups depending on whether the ORF was truncated in rat or not (the truncation had to affect more than half of the protein). For neutrally evolving ORFs the number of cases in which the ORF was truncated was similar to the number of cases in which it was not truncated, and in both cases the polymorphism patterns were consistent with neutrality (Supplementary Table 5). This indicated that, for this group, the presence of a similar ORF in rat does not imply functional conservation of the ORF. Therefore, we did not use information on rat genomic synteny to classify the genes as conserved/non-conserved.

Coding score

We used a previously described metric based on hexamer frequencies to calculate the coding score of the sequences[8]. The method uses a table of pre-calculated hexamer scores that measure the relative frequency of each hexamer in coding versus non-coding sequences. These scores are then used to evaluate the coding propensity of a sequence based on its hexamer composition. The method has been implemented in a computational program called CIPHER that can be used to predict putative translated ORFs and can be accessed online (http://evolutionarygenomics.upf.edu/cipher).

Comparison of the coding score in translated and non-translated ORFs

For comparative purposes we generated a set of non-translated ORFs that was completely independent from the translated ORFs (Fig. 4). We selected expressed loci that did not contain any translated ORF and selected the longest predicted ORF sequence. An ORF was considered not to be translated if it had less than 10 Ribo-Seq mapped reads, or a RibORF score < 0.7. Selecting the longest ORF was justified by the fact that, 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 also generated a set of 4,013 randomly taken ORFs from introns, after discarding ORFs that showed significant sequence similarity to known proteins from the same species (BLASTP, e-value < 10^{-4}).

In order to investigate the influence of gene expression level in our capacity to detect translation of an ORF we subsampled the set of translated neutral ORFs and the set of non-translated ORFs so that the two sets had equivalent gene expression value distributions. As gene expression units we used Fragments Per Kilobase per Million mapped reads (FPKM), and we selected the highest FPKM value across the different tissues examined. We applied the same procedure to control for coding score values when evaluating the effect of gene expression level and ORF length.

Ribosome profiling controls

We inspected the ribosome profiling patterns of neutral ORFs with respect to the rest of translated ORFs ("functional"). The number of Ribo-Seq reads per base was in general lower in neutral ORFs than in functional ORFs, although there was a significant overlap in the two distributions (Supplementary Fig. 8). Translated sequences were characterized by strong 3-nucleotide periodicity and this property showed a significant correlation across pairs of tissues, both for neutral and functional ORFs (Supplementary Fig. 8 and 9). After subsampling the functional ORFs to match the neutral ORFs with regards to ORF size and number of Ribo-Seq reads, the two sets showed similar RibORF score, periodicity and uniformity value distributions (Supplementary Fig. 10).

We generated a negative control set by taking randomly occurring ORFs in mouse small nuclear and nucleolar RNAs ("sRNAs"). These ORFs were required to have at least 10 Ribo-Seq mapped reads and were processed in the same manner as the main set of ORFs under study. Only 10 out of 304 ORFs in "sRNA" had a RibORF score equal or higher than 0.7 (3.33% false positive rate). This control was characterized by low uniformity and low read periodicity (Supplementary Fig. 8). A second negative control in which all parameters were equal to those observed in neutral ORFs, but the reads were randomly distributed in the three frames, resulted in 176 out of 4,232 ORFs classified as translated ("Random", 4.16% false positive rate).

We used a similar strategy to estimate the fraction of translated non-conserved ORFs (NC) over all non-conserved ORF and the fraction of translated neutral ORFs over all neutral ORFs (NC-L). Non-translated ORFs were those covered by 10 or more Ribo-Seq reads but with a RibORF score lower than 0.7. We followed the same procedure described for the translated ORFs (Figure 2a) to classify them into classes. We detected translation for 33.1% of the non-conserved ORFs (1,980 out of 5,975) and 30.5% of the neutral ORFs (1,291 out of 4,232). These values are much higher than the estimated false positive rate.

We also generated a positive control set composed of 2,163 randomly taken annotated mouse coding sequences with protein evidence in Uniprot. With this control we estimated a false negative rate of 2.54%.

Single nucleotide polymorphism analysis

We obtained single nucleotide polymorphism (SNP) data from 20 individuals of the house mouse subspecies *Mus musculus castaneus*[27]. 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 ((PN/PS)obs). 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

(Supplementary Tables 3 and 4). The observed transition to transversion ratio was 4.42, very similar to the 4.26 value obtained in early observations based on mouse-rat divergence data[59]. We tested for purifying selection using a chi-square test with one degree of freedom (Supplementary Table 5). Positively selected mutations are rapidly fixed in the population and their effect is expected to be negligible when using SNP data.

Proteomics data

We used the proteomics database PRIDE[31] 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. Under these conditions the false positive rate was estimated to be below 0.2% in a previous study[50]. The estimation was based on the results obtained for ORFs in which the amino acid sequence was randomized and for ORFs in intronic sequences.

Statistical tests and plots

The generation of plots and statistical tests was performed with the R package[60].

Code availability

Code for calculating the coding score is available at https://github.com/jorruior/CIPHER and for calculating the expected PN/PS at https://figshare.com/articles/computePNPS_c/5085706.

Data availability

The datasets generated during the current study are available from figshare repository, http://dx.doi.org/10.6084/m9.figshare.4702375.

Figure legends

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 with length ≥ 30 nucleotides 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 and read uniformity are evaluated by RiboORF. **b-c.** Number of translated and non-translated expressed loci **(b)** and ORFs **(c)** using a RibORF cut-off of 0.7. The original data come from eight different mouse tissues (Supplementary Table 1). The translated ORFs have been divided into small ORFs (smORF, < 100 aa) and long ORFs (≥ 100 aa), depending on their length. “Novel” corresponds to expressed loci that do not correspond to annotated protein-coding genes or lncRNAs, they are merged with lncRNAs as they mostly contain short non-conserved ORFs.

Figure 2. Identification of selection signatures. a. Workflow to identify conserved and non-conserved ORFs. Translated ORFs shorter than 24 amino acids, as well as non-conserved upstream and downstream ORF in conserved transcripts (uORFs and dORFs, see Methods), were filtered out. Any ORF with at least one BLAST match in another species was classified as conserved (C), otherwise it was classified as non-conserved (NC). Non-conserved ORFs with a high coding score value (≥ 0.1014) were classified as NC-H, and the rest were classified as NC-L. **b**. 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; a value of 1 is expected in the absence of selection at the protein level. Conserved and NC-H ORFs showed significant purifying selection signatures. In contrast, NC-L ORFs did not show evidence of purifying selection at the protein level. Some conserved ORFs in lncRNAs are likely to encode functional micropeptides. Differences between observed and expected PN/PS were assessed with a chi-square test, * p- value < 0.05, *** p-value < 10^{-5}. Error bars indicate the standard error of the sample proportion. Numbers of ORFs for the different categories are also displayed.

Figure 3. Three nucleotide periodicity of translated ORFs. The mapping of Ribo-Seq reads

on different types of ORFs is shown. The Y axis represents the log-number of reads, the X axis the positions in the ORF. The reads show strong frame bias in the functional (conserved) and the neutral (NC-L) examples, with a preponderance of in-frame reads (green) versus off-frame reads (red and blue), while there is no consistent frame bias in the negative control (SNORA18). The exon/intron structure and the amino acid sequence for translated ORFs is also shown.

Figure 4. Factors influencing the translation of neutrally evolving ORFs. a. Influence of coding score in the translatability of neutrally evolving ORFs. Translated ORFs showed significantly higher coding score than non-translated ORFs, both sets had significantly higher coding scores than introns (Wilcoxon test p-value < 10^{-5}, indicated by ***). **b.** Influence of coding score in the translatability of ORFs after controlling for gene expression values. Translated ORFs showed significantly higher coding score values than non-translated ORFs (Wilcoxon test p-value < 10^{-5}). **c.** Influence of gene expression level and ORF length in the translatability of neutral ORFs after controlling for coding score. Translated ORFs showed significantly higher FPKM values than non-translated ORFs (Wilcoxon test p-value < 10^{-5}); differences in length were not significant.

References

1. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene expression. *PLoS Genet.* **8,** e1002841 (2012).
2. Wiberg, R. A. W. *et al.* Assessing Recent Selection and Functionality at Long Noncoding RNA Loci in the Mouse Genome. *Genome Biol. Evol.* **7,** 2432–44 (2015).
3. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLOS Genet.* **11,** e1005721 (2015).
4. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *Elife* **4,** e08890 (2015).
5. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by ribosome footprint profiling. *Elife* **5,** e13328 (2016).
6. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic

stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,** 789–802 (2011).

7. Ingolia, N. T. *et al.* Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Rep.* **8,** 1365–79 (2014).

8. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Alba, M. M. Long non-coding RNAs as a source of new peptides. *Elife* **3,** e03523 (2014).

9. Wilson, B. A. & Masel, J. Putatively noncoding transcripts show extensive association with ribosomes. *Genome Biol. Evol.* **3,** 1245–52 (2011).

10. Couso, J.-P. & Patraquim, P. Classification and function of small open reading frames. *Nat. Rev. Mol. Cell Biol.* **18,** 575–589 (2017).

11. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S. & Weissman, J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* **324,** 218–23 (2009).

12. Bazzini, A. A. *et al.* Identification of small ORFs in vertebrates using ribosome footprinting and evolutionary conservation. *EMBO J.* **33,** 981–93 (2014).

13. Calviello, L. *et al.* Detecting actively translated open reading frames in ribosome profiling data. *Nat Meth* **13,** 165–170 (2016).

14. Aspden, J. L. *et al.* Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *Elife* e03528 (2014).

15. Mackowiak, S. D. *et al.* Extensive identification and analysis of conserved small ORFs in animals. *Genome Biol.* **16,** 1–21 (2015).

16. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de Novo Evolution of Testis-Expressed Genes in the Drosophila yakuba/Drosophila erecta Clade. *Genetics* **176,** 1131–1137 (2006).

17. Tautz, D. & Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12,** 692–702 (2011).

18. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17,** 567–578 (2016).

19. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in Drosophila melanogaster populations. *Science* **343,** 769–72 (2014).

20. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487,** 370–4 (2012).
21. Toll-Riera, M. *et al.* Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26,** 603–12 (2009).
22. Cai, J. J. & Petrov, D. A. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Genome Biol. Evol.* **2,** 393–409 (2010).
23. Chen, S., Zhang, Y. E. & Long, M. New Genes in Drosophila Quickly Become Essential. *Science (80-.).* **330,** 1682–1685 (2010).
24. Reinhardt, J. A. *et al.* De novo ORFs in Drosophila are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet.* **9,** e1003860 (2013).
25. Sunyaev, S., Kondrashov, F. A., Bork, P. & Ramensky, V. Impact of selection, mutation rate and genetic drift on human genetic variation. *Hum. Mol. Genet.* **12,** 3325–3330 (2003).
26. Gayà-Vidal, M. & Albà, M. M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15,** 599 (2014).
27. Harr, B. *et al.* Genomic resources for wild populations of the house mouse, Mus musculus and its close relative Mus spretus. *Sci. Data* **3,** 160075 (2016).
28. Buck-Koehntop, B. A., Mascioni, A., Buffy, J. J. & Veglia, G. Structure, dynamics, and membrane topology of stannin: a mediator of neuronal cell apoptosis induced by trimethyltin chloride. *J. Mol. Biol.* **354,** 652–65 (2005).
29. Pueyo, J. I. *et al.* Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and Conserved across Metazoans. *PLoS Biol* **14,** e1002395 (2016).
30. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* (2011). doi:10.1038/nrg2899
31. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456 (2016).
32. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).
33. Heinen, T. J. A. J., Staubach, F., Häming, D. & Tautz, D. Emergence of a new gene from an intergenic region. *Curr. Biol.* **19,** 1527–31 (2009).

34. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons. *Nucleic Acids Res.* **42,** 9171–81 (2014).
35. Yu, C. *et al.* Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell* **59,** 744–754 (2015).
36. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160,** 1111–24 (2015).
37. Schlötterer, C. Genes from scratch – the evolutionary fate of de novo genes. *Trends Genet.* (2015).
38. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420,** 563–73 (2002).
39. Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* **5,** e09977 (2016).
40. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15690–5 (2015).
41. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nat. Ecol. Evol.* **1,** 146 (2017).
42. Kaiser, C. A., Preuss, D., Grisafi, P. & Botstein, D. Many random sequences functionally replace the secretion signal sequence of yeast invertase. *Science* **235,** 312–7 (1987).
43. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library. *Nature* **410,** 715–718 (2001).
44. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 217 (2017).
45. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.* **3,** 2179–90 (2013).
46. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505,** 635–40 (2014).
47. Smeds, L. & Künstner, A. ConDeTri--a content dependent read trimmer for Illumina data. *PLoS One* **6,** e26314 (2011).
48. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of

insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

49. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotech* **33,** 290–295 (2015).

50. Luis Villanueva-Cañas, J. *et al.* New Genes and Functional Innovation in Mammals. *Genome Biol. Evol.* **9,** 1886–1900 (2017).

51. Gonzalez, C. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape in brain tumors. *J. Neurosci.* **34,** 10924–36 (2014).

52. Castañeda, J. *et al.* Reduced pachytene piRNAs and translation underlie spermiogenic arrest in Maelstrom mutant mice. *EMBO J.* **33,** 1999–2019 (2014).

53. Guo, H., Ingolia, N. T., Weissman, J. S. & Bartel, D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466,** 835–40 (2010).

54. Diaz-Munoz, M. D. *et al.* The RNA-binding protein HuR is essential for the B cell antibody response. *Nat Immunol* **16,** 415–425 (2015).

55. Cho, J. *et al.* Multiple repressive mechanisms in the hippocampus during memory formation. *Science* **350,** 82–87 (2015).

56. Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* **29,** 2790–2791 (2013).

57. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

58. Karolchik, D. *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.* **42,** D764–D770 (2014).

59. Rosenberg, M. S., Subramanian, S. & Kumar, S. Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. *Mol. Biol. Evol.* **20,** 988–993 (2003).

60. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016).

Acknowledgements

We are grateful for valuable discussions with many colleagues during this study. The work was funded by grants BFU2012-36820, BFU2015-65235-P and TIN2015-69175-C4-3-R from

Ministerio de Economía e Innovación (Spanish Government) and co-funded by FEDER (EC). We also received funding from Agència de Gestió d'Ajuts Universitaris i de Recerca Generalitat de Catalunya (AGAUR), grant number 2014SGR1121.

Author contributions

J.R-O. and M.M.A. conceived the study, interpreted the data and wrote the paper. J.R-O. performed most of the analyses including the transcript assemblies, identification of translated open reading frames, BLAST searches, SNP mapping and generation of controls. J.R-O., P.V-G. and J.L.V-C wrote code and performed analyses on the coding score. X.M. wrote code to calculate the expected SNP frequencies. M.M.A. coordinated the study.

Competing interests

The authors declare no competing financial interests.