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: pervasive translation

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

Abstract

There is accumulating evidence that some 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 can at some point acquire new functions. Here we show for the first time that such peptides do not only exist but that they are prevalent among the translation products of mouse genes that lack homologues in rat and human. The data suggests that the translation of these peptides is due to the chance occurrence of open reading frames with a favorable codon composition. Our approach combines ribosome profiling experiments, proteomics data and non-synonymous and synonymous nucleotide polymorphism analysis. We propose that effectively neutral processes involving the expression of thousands of transcripts all the way down to proteins provide a basis for *de novo* gene evolution.

The mammalian genome is pervasively 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 fraction of these transcripts is likely to translate small peptides[4–9], although the significance of this finding has remained elusive.

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[10]. 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[4,11,12]. Given enough sequence coverage the technique can uncover low-abundant small peptides that would be otherwise difficult to detect by standard proteomics approaches[13,14].

To assess the functional relevance of novel events of translation 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 homologous sequences. A more general approach that can be used in the absence of homology 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 accumulate at the same rate, whereas under purifying selection there is a deficit of non-synonymous polymorphisms because some amino acid changes disrupt the protein's function[15]. Single nucleotide polymorphism analysis can be performed on a gene-by-gene basis or in pools of sequences that share certain features[2,16].

We previously observed that, as a whole, putatively translated lncRNAs and young protein-coding genes share a number of similarities, such as small ORF size and weak selective

constraints, compared with more widely conserved genes[8]. This pointed to a link between the translation of lncRNAs and the evolution of new proteins, but it did not solve the key question of whether translation of new ORFs could occur in the absence of selection at the protein level. This is a fundamental issue because for a new protein to acquire a function it first needs to be produced in the cell at significant amounts. Here by employing a combination of ribosome profiling data, sequence analysis and single nucleotide polymorphism information we obtain strong evidence that the majority of mouse proteins that are not conserved in rat or human selection evolve in a neutral manner. This study renders visible for the first time a layer of protein expression that is not dependent on selective processes, filling a gap in our understanding of the processes underlying *de novo* gene birth.

Results

First we set 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 genes[3].

We used the RibORF program[4] to identify *bona fide* 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 genes that did not map to the gene annotations but contained translated ORFs (Fig. 1b). A widely used criterion to annotate a transcript as protein-coding is the

presence of an ORF encoding a protein of at least 100 amino acids[17]. Not surprisingly, the vast majority of ORFs translated from annotated and novel lncRNAs encoded proteins smaller than 100 amino acids (smORFs).

We hypothesized that some of the translated ORFs may evolve in a neutral manner and constitute a reservoir for the evolution of new protein-coding genes. To test this hypothesis, we first identified translated ORFs that were mouse-specific and then tested them for signatures of selection. We performed exhaustive sequence similarity searches of the 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 proteins 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 different translated ORFs that showed no homology to expressed sequences in other species (class non-conserved or NC). In general, these ORFs had lower codon usage bias than conserved ORFs, as measured by a previously described hexamer-based coding score metric[8] (Fig. 2b).

To measure the strength of selection in conserved and non-conserved translated ORFs we employed a large collection of mouse single nucleotide polymorphisms (SNPs) for the house mouse subspecies *Mus musculus castaneus* [18]. We could map a total of 324,729 SNPs to the set of translated ORFs. We grouped the ORFs into three different classes on the basis of conservation and coding score (Fig. 2b), and calculated the ratio between the number of observed non-synonymous and synonymous SNPs (PN/PS(obs)) in each class. We then 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 and 4). 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 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 (Supplementary Table 5). As expected, the PN/PS of

104 randomly selected ORFs from introns was approximately 1. Instead, the PN/PS of conserved
105 ORFs was around 0.15 (Fig. 2c, chi-square test p-value < 10^{-5}), consistent with protein
106 functionality. One example in this group was Stannin[19,20], a highly conserved peptide that
107 regulates neuronal cell apoptosis (Fig. 3).
108
109 Non-conserved ORFs with high coding scores (NC-H coding score ≥ 0.1014, Fig. 2b and c,
110 Supplementary Figure 1) had weak but significant signatures of selection (p-value < 0.05),
111 possibly because of the existence of some functional mouse-specific genes. In contrast, the
112 PN/PS ratio of the remaining non-conserved ORFs was not significantly different from 1,
113 consistent with neutral evolution. Very similar results were obtained for non-conserved genes
114 annotated as coding or lncRNA (Fig. 2c) and the two sets were merged into a single group of
115 neutrally evolving ORFs (neutral ORFs). This set comprised about two thirds of the non-
116 conserved ORFs (1,291 out of 1,980 ORFs analysed), and represented ~6.8% of the total
117 number of mouse translated ORFs.
118
119 We used proteomics data from the PRIDE database[21] to further validate the translation of this
120 latter group of proteins. Despite their small size (median 44 amino acids), a limiting factor for
121 their detection by standard proteomics-based tecniques[22], we found proteomics evidence for 32
122 of the neutral ORFs (see Methods). This represents 2.5% of the proteins in this set (compared
123 to less than 0.2% false positive rate, see Methods). This fraction is similar to the one obtained
124 for conserved proteins subsampled to have a similar size distribution as the neutral ORFs
125 (2.9%; in contrast, about 41% of all conserved ORFs have proteomics evidence). The test
126 based on the PN/PS ratio confirmed that this subset of 32 ORFs did not deviate significantly
127 from neutrality either (Supplementary Table 5).
128
129 The above analyses grouped the sequences into classes before computing the PN/PS ratio. In
130 general, ORF-by-ORF analysis was not possible because the ORFs were small and contained
131 too few SNPs. Nevertheless, 41 of the neutrally evolving ORFs contained 10 or more SNPs,
132 and we decided to compute a normalized PN/PS ratio for these individual cases. The median
133 PN/PS of these ORFs was around 1 and the distribution of PN/PS values was very different

from that of conserved ORFs (Fig. 2d, 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 1,282 ORFs that contained this type of mutation (5.6%) and in the second case 296 out of 16,892 ORFs (1.75%). 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. These analyses further support that the selective pressures acting on both kinds of ORFs are very different.

We next inspected in more detail the ribosome profiling patterns of neutral ORFs with respect to the rest of translated ORFs (hereafter called “functional”). Genes with a recent origin are usually expressed at lower levels than older genes[23,24,3], so it was not surprising to observe that neutrally evolving ORFs were associated with a lower number of Ribo-Seq reads per base than the rest of translated ORFs (median 0.193 versus 0.474, respectively, Supplementary figure 2). Consistent with translation, read periodicity in both neutral and functional ORFs was much higher than the random expectation of 0.33 (median values 0.70 and 0.80, respectively; see examples in Figure 3). Importantly, the results were highly reproducible across tissues (Figure 4a for hippocampus and embryonic stem cells; Supplementary Figure 3 hippocampus and brain), a result we would not expect in the case of spurious ribosome profiling signals. In general, the RibORF score of the translated ORFs was positively related to the number of mapped Ribo-Seq reads (Spearman correlation R=0.408), and to the size of the ORFs (Spearman correlation R=0.193). When we controlled for these two parameters, neutral and functional ORFs had equivalent distributions of RibORF score, periodicity and uniformity values (Figure 4b).

Subsequently, we compared our results to those obtained with two negative controls. The first control contained ORFs in alternative frames of annotated protein coding sequences with experimental protein evidence (“off-frame”). The second one contained randomly occurring

ORFs in small nuclear and nucleolar annotated RNA sequences ("sRNA"). The latter RNAs are sometimes detected in ribosome profiling experiments due to the formation of ribonucleoprotein particles that protect the RNA from degradation[25]. As before, we only considered ORFs with at least 10 Ribo-Seq mapped reads. As expected, the vast majority of the ORFs in these controls did not display significant 3-nucleotide read periodicity (Supplementary figure 2, see a specific example in Figure 3). We found that only 234 out of 13,596 ORFs in "off-frame", and 10 out of 304 ORFs in "sRNA", had a RibORF score ≥0.7 (the threshold employed throughout our study). This corresponds to an overall false discovery rate (FDR) of 1.75%, much lower than the fraction of neutrally evolving proteins detected in our main analysis (6.8%).

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*[26] that lacked any evidence of translation in the data we analysed. We next asked which factors may influence the translation of some neutrally evolving ORFs but not of others. First, we inspected the translation initiation sequence context but did not detect any significant differences between translated and non-translated ORFs (Supplementary Figure 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. Consistent with this hypothesis, we found that the translated neutrally evolving ORFs exhibited higher coding scores than non-translated ORFs with otherwise similar characteristics (Fig. 5a, Translated versus non-translated Wilcoxon test, p-value < 10^{-5}). Importantly, we obtained a similar result after controlling for gene expression level (Fig. 5b, 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, had an effect on the translatability of the transcript (Fig. 5c).

The results suggest that the neutral ORFs that are translated are enriched in codons that are frequently found in functional protein coding sequences. This is consistent with the observation that abundant codons enhance translation elongation[27], whereas rare codons might affect the stability of the mRNA[28]. It has been previously 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[4]. However, we found evidence that some lncRNAs with nuclear functions, such as *Malat1* and *Neat1*, were translated, suggesting that the cytosolic fraction of a transcript can be translated independently of its role or preferred location.

DISCUSSION

The molecular mechanisms underlying *de novo* gene evolution are still poorly understood[29,30,31]. 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 genes. An excess of transcription was already noted in the first large-scale cDNA sequencing efforts performed in human and mouse[32], and more recent studies have found a high rate of transcriptional turnover when comparing closely related species[33]. Here we have shown that many of these transcripts are translated even if they only contain small ORFs, with the data currently available we have been able to identify 1,291 peptides in 1,132 genes that are likely to be of recent evolutionary origin and that show no signs of selection. This number is likely to be a gross underestimate because many transcripts are expressed at low levels, limiting their detection, and many cell types and tissues have not yet been sampled. According to recent estimates, the cost of transcription and translation in multicellular organisms is probably too small to overcome genetic drift[34]. Therefore, these activities may be effectively neutral. Our results indeed support that there is no barrier for the production of peptides that do not confer an immediate selective advantage.

The putative precursors of novel proteins identified here are of small size, which is consistent with observations for functional *de novo* genes identified in previous studies[23,35–37]. We have also shown that random ORFs with a more favorable, coding-like, hexamer composition are more likely to be translated than other ORFs. Codon usage bias in functional sequences is related to the abundance of different tRNAs and correlates with expression level[38,39]. Thus, it

seems logical that the translated neutral ORFs are biased towards those codons that are translated more efficiently.

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 random peptide sequences can function as secretion signals[40], 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[41]. Recent experiments performed in *E.coli* also suggest that random sequences can often affect cellular growth[42]. The pervasive translation of the transcriptome implies that *de novo* gene evolution has much more material at its disposal than previously thought.

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[43], and GSE43520[44]; 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)[45] 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)[46]. Multiple mapping to several locations in the genome was allowed unless otherwise stated. We assembled the transcriptome with Stringtie[47], 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[48].

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[49], GSE50983[50], GSE22001[51], GSE62134[52], GSE72064[53], and GSE41246. Only datasets corresponding to non-pathogenic conditions were considered. The reads from the 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*[33] to the mouse assembled transcriptome using NextGenMap[54]. As expected, the vast majority of the genes expressed in brain samples from C57BL/6 mice[49] also showed evidence of expression in *Mus musculus castaneus* brain RNA samples[33](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 so as not to compromise the detection of paralogous proteins (Supplementary Fig. 7). 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[10,12].

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 read periodicity and the read uniformity. The first one is the fraction of reads that correspond to the correct frame and the second one corresponds to the percentage of maximum entropy, a value of 1 indicates a completely even distribution of reads. For each ORF the program computes a score that depends on the values of these two parameters[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 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), and the data was subsequently integrated, taking the tissue with the highest RibORF score as representative. 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 top 5 most highly expressed proteins (Supplementary Fig. 6 and 7, respectively). For genes with no evidence of translation we selected the longest ORF across all transcripts for comparative purposes. 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}).

We generated a negative control set by combining out-of-frame ORFs in mouse coding genes with experimental protein evidence according to Uniprot (“off-frame”) and 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. The total number of sequences in the negative control was 13,900.

We also generated a positive control set composed of 2,163 randomly taken annotated mouse coding sequences with protein evidence in Uniprot. With these controls we estimated a false positive rate of 1.75% and a false negative rate of 2.54% for the above mentioned RibORF score cut-off.

Sequence conservation

We searched for mouse translated ORF homologues in the human and rat transcriptomes using TBLASTN (limited to one strand, e-value < 10^{-4})[55]. 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[48], 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. Translated 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[56]. 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.

Single nucleotide polymorphism analysis

We obtained single nucleotide polymorphism (SNP) data from 20 individuals of the house

mouse subspecies *Mus musculus castaneus*[18]. 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 2 to 5). 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[57]. We tested for purifying selection by the number of observed and expected non-synonymous and synonymous SNPs using a chi-square test with one degree of freedom. 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[21] 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 with a false positive rate < 0.2%[48].

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 accessed online (http:// http://evolutionarygenomics.upf.edu/cipher).

374 **Statistical tests and plots**

375

376 The generation of plots and statistical tests was performed with the R package[58].

377

378 **Data availability**

379

380 Transcript assemblies from mouse, human and rat, as well as the mouse open reading frames
381 (ORFs) predicted to be translated have been deposited at figshare
382 (http://dx.doi.org/10.6084/m9.figshare.4702375). The code and executable file to calculate the
383 coding score can be accessed at https://github.com/jorruior/CIPHER.
384 The C program to calculated the PN/PS expected under neutrality is available at
385 https://figshare.com/articles/computePNPS_c/5085706. Supplementary file 1 contains
386 supplementary tables and figures.

387

REFERENCES

388 1. Kutter, C. *et al.* Rapid turnover of long noncoding RNAs and the evolution of gene
389 expression. *PLoS Genet.* **8,** e1002841 (2012).
390 2. Wiberg, R. A. W. *et al.* Assessing Recent Selection and Functionality at Long Noncoding
391 RNA Loci in the Mouse Genome. *Genome Biol. Evol.* **7,** 2432–44 (2015).
392 3. Ruiz-Orera, J. *et al.* Origins of De Novo Genes in Human and Chimpanzee. *PLOS*
393 *Genet.* **11,** e1005721 (2015).
394 4. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are
395 translated and some are likely to express functional proteins. *Elife* **4,** e08890 (2015).
396 5. Raj, A. *et al.* Thousands of novel translated open reading frames in humans inferred by
397 ribosome footprint profiling. *Elife* **5,** (2016).
398 6. Ingolia, N. T., Lareau, L. F. & Weissman, J. S. Ribosome profiling of mouse embryonic
399 stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* **147,**
400 789–802 (2011).
401 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. 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).

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

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

13. Aspden, J. L. *et al.* Extensive translation of small ORFs revealed by Poly-Ribo-Seq. *Elife* e03528 (2014). doi:10.7554/eLife.03528

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

15. 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).

16. Gayà-Vidal, M. & Albà, M. M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15,** 599 (2014).

17. Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.* **4,** e1000176 (2008).

18. 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).

19. 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).

20. Pueyo, J. I. *et al.* Hemotin, a Regulator of Phagocytosis Encoded by a Small ORF and

Conserved across Metazoans. *PLoS Biol* **14,** e1002395 (2016).

21. Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44,** D447–D456 (2016).

22. Slavoff, S. A. *et al.* Peptidomic discovery of short open reading frame-encoded peptides in human cells. *Nat. Chem. Biol.* **9,** 59–64 (2013).

23. Toll-Riera, M. *et al.* Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26,** 603–12 (2009).

24. Xie, C. *et al.* Hominoid-specific de novo protein-coding genes originating from long non-coding RNAs. *PLoS Genet.* **8,** e1002942 (2012).

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

26. 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).

27. 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).

28. Presnyak, V. *et al.* Codon optimality is a major determinant of mRNA stability. *Cell* **160,** 1111–24 (2015).

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

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

31. 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).

32. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420,** 563–73 (2002).

33. Neme, R. & Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* **5,** (2016).

34. Lynch, M. & Marinov, G. K. The bioenergetic costs of a gene. *Proc. Natl. Acad. Sci. U. S. A.* **112,** 15690–5 (2015).

35. Carvunis, A.-R. *et al.* Proto-genes and de novo gene birth. *Nature* **487,** 370–4 (2012).

462 36. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in
463 Drosophila melanogaster populations. *Science* **343,** 769–72 (2014).
464 37. Wilson, B. A., Foy, S. G., Neme, R. & Masel, J. Young Genes are Highly Disordered as
465 Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. *Nat. Ecol. Evol.* **1,**
466 0146–146 (2017).
467 38. Plotkin, J. B. & Kudla, G. Synonymous but not the same: the causes and consequences
468 of codon bias. *Nat. Rev. Genet.* (2011). doi:10.1038/nrg2899
469 39. Dana, A. & Tuller, T. The effect of tRNA levels on decoding times of mRNA codons.
470 *Nucleic Acids Res.* **42,** 9171–81 (2014).
471 40. Kaiser, C. A., Preuss, D., Grisafi, P. & Botstein, D. Many random sequences functionally
472 replace the secretion signal sequence of yeast invertase. *Science* **235,** 312–7 (1987).
473 41. Keefe, A. D. & Szostak, J. W. Functional proteins from a random-sequence library.
474 *Nature* **410,** 715–718 (2001).
475 42. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are
476 an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* (2017).
477 doi:10.1038/s41559-017-0127
478 43. Soumillon, M. *et al.* Cellular source and mechanisms of high transcriptome complexity in
479 the mammalian testis. *Cell Rep.* **3,** 2179–90 (2013).
480 44. Necsulea, A. *et al.* The evolution of lncRNA repertoires and expression patterns in
481 tetrapods. *Nature* **505,** 635–40 (2014).
482 45. Smeds, L. & Künstner, A. ConDeTri--a content dependent read trimmer for Illumina data.
483 *PLoS One* **6,** e26314 (2011).
484 46. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of
485 insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
486 47. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from
487 RNA-seq reads. *Nat Biotech* **33,** 290–295 (2015).
488 48. Luis Villanueva-Cañas, J. *et al.* New Genes and Functional Innovation in Mammals.
489 *Genome Biol. Evol.* **9,** 1886–1900 (2017).
490 49. Gonzalez, C. *et al.* Ribosome profiling reveals a cell-type-specific translational landscape
491 in brain tumors. *J. Neurosci.* **34,** 10924–36 (2014).

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

51. 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).

52. 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).

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

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

55. 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).

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

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

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

ACKNOWLEDGEMENTS

We are grateful for valuable discussions with many colleagues during this study. The work was funded by grants BFU2012-36820 and BFU2015-65235-P 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.

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.** Number of translated and non-translated expressed genes belonging to different classes after integrating data from eight different mouse tissues (Supplementary Table 1). **c.** Number of translated ORFs belonging to different classes. The translated ORFs have been divided into small ORFs (smORF, < 100 aa) and long ORFs (≥ 100 aa), depending on their length.

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). **b.** Coding score in 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}. Non-conserved ORFs with a high coding score value (≥ 0.1014) were classified as NC-H, and the rest were classified as NC-L. **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; 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. Many 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. **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 analysed. The differences between C and NC-L are significant (Wilcoxon test, p-value <10^{-5}).

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 the frame bias is randomly distributed in the negative control (SNORA18). The exon/intron structure and the amino acid sequence for translated ORFs is also shown.

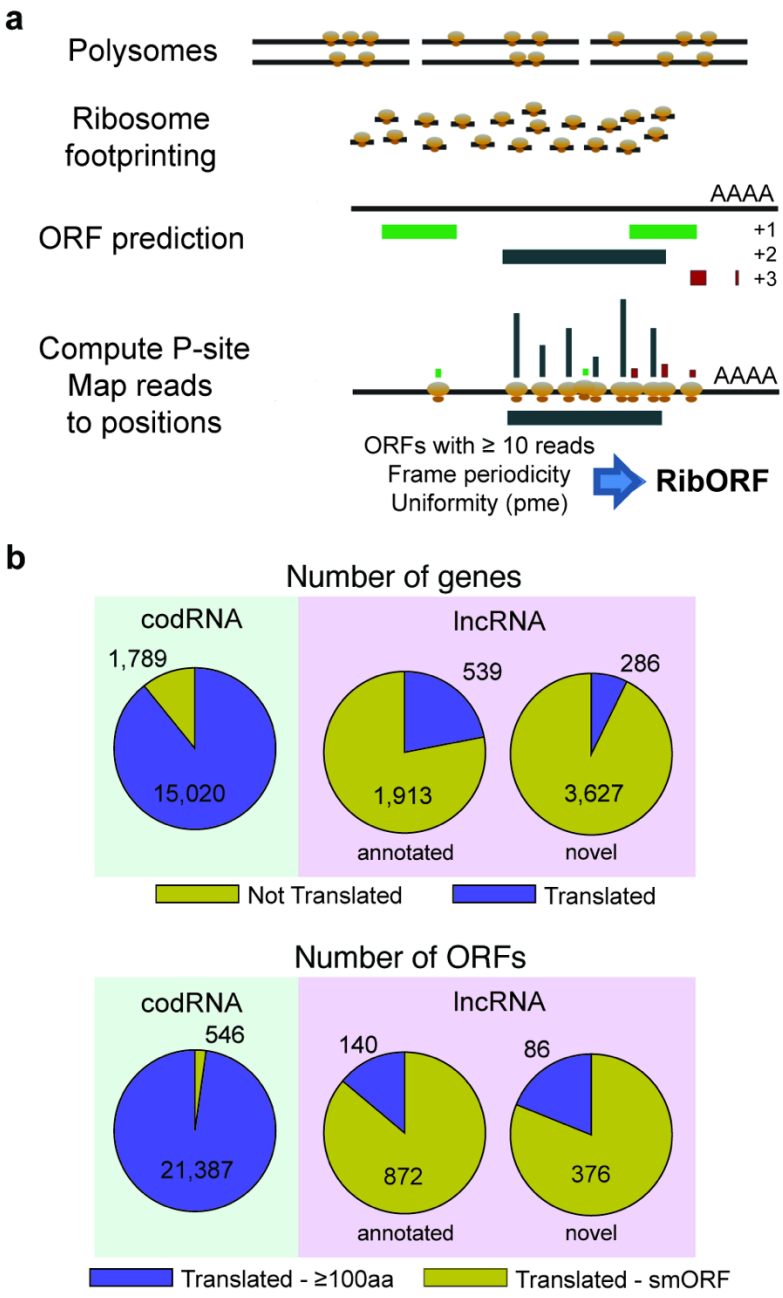
Figure 4. Properties of neutrally evolving ORFs. a. Relationship between the percentage of reads falling in the correct frame in neural embryonic stem cells cells and hippocampus samples, for neutral and functional ORFs having at least 10 reads in both samples and being translated in at least one sample. Spearman correlation coefficient is R=0.4224 for the neutrat set (p-value < 10^{-5}) and R=0.4360 for the functional set (p-value < 10^{-5}). **b.** Distribution of read periodicity, read uniformity and RibORF scores in neutral and functional translated ORFs after controlling for the number of Ribo-Seq reads and size of ORFs. The 'functional normalized' set is a randomly taken subset of the functional ORFs that has the same number of mapped Ribo-Seq reads and ORF size distribution as the set of neutrally evolving ORFs (n=900). Data is represented as box-plots for different number of read intervals; the box contains 50% of the data, horizontal line is the median value.

Figure 5. 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 controlling for gene expression values, the two sets have comparable maximum FPKM gene expression (median FPKM value = 11.10). Translated ORFs showed significantly higher coding score values than non-translated ORFs; (Wilcoxon test p-value < 10^{-5}). **c.** Influence of maximum FPKM gene expression and ORF length in the translatability of neutral ORFs normalized by coding score (median coding score value = -0.0052). Translated ORFs showed significantly higher FPKM values than non-translated ORFs (Wilcoxon test p-value < 10^{-5}); differences in length were not significant.

578 **FIGURES**

579 **Figure 1**

580

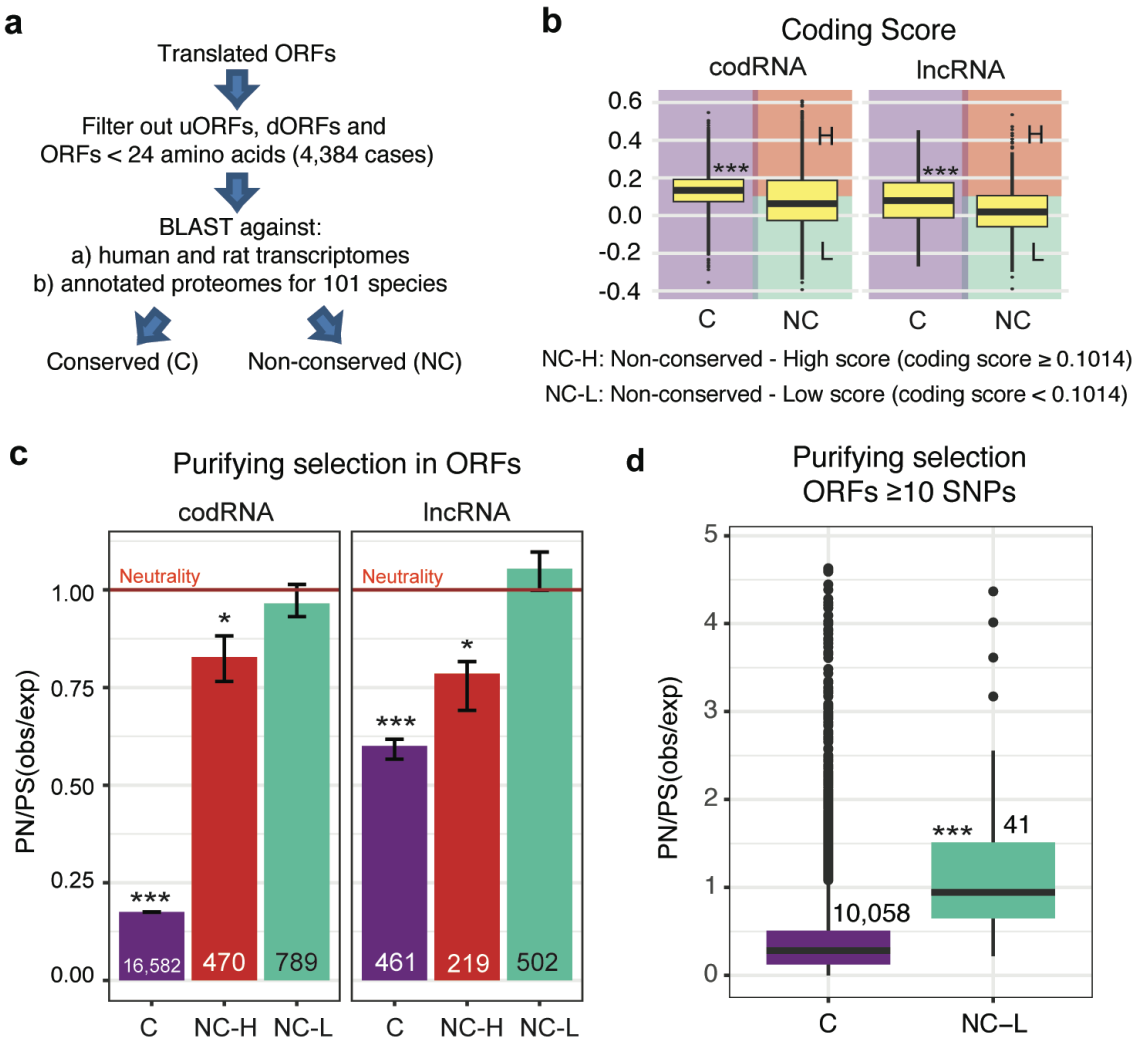


581

582

583 **Figure 2**

584

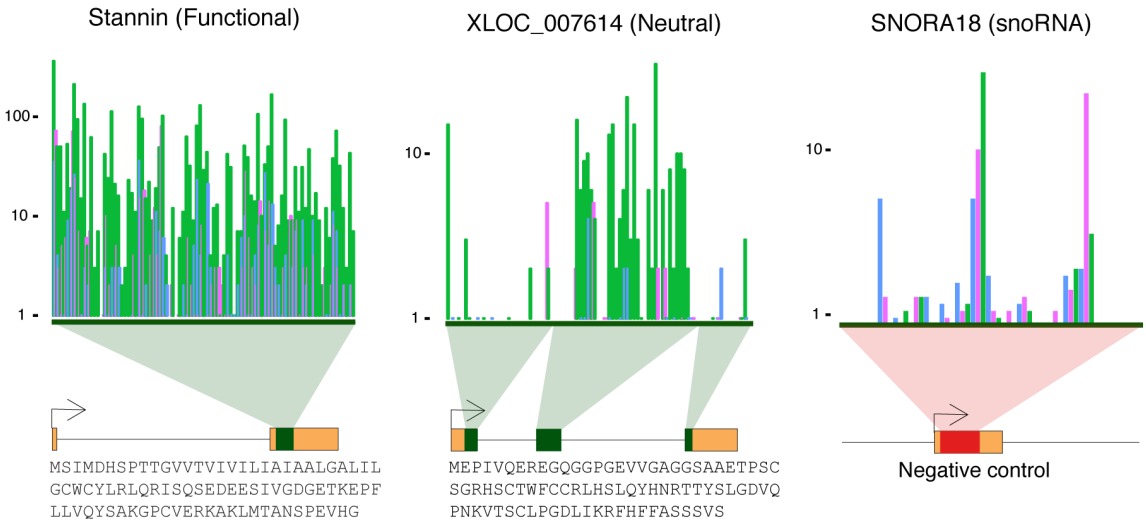


585

586

587 **Figure 3**

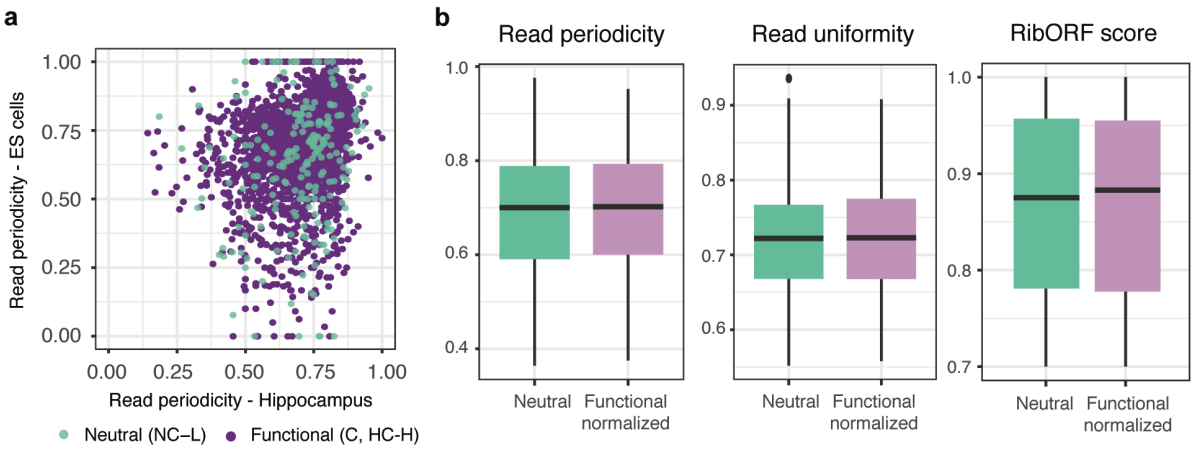
588



589

590

591 **Figure 4**



592

593

Figure 5