Ribo-ITP expands the translatome of limited input samples

Vighnesh Ghatpande[1#], Uma Paul[1#], MacKenzie A Howard[2], Can Cenik[1*]
1 Department of Molecular Biosciences, University of Texas at Austin, Austin, Texas 78712, USA
2 Departments of Neurology and Neuroscience, Center for Learning and Memory, Dell Medical School, University of Texas at Austin, Austin, Texas 78712, USA
* Corresponding Author
# Equal contribution

Abstract:
In the last decade, an unexpectedly large number of translated regions (translons) have been discovered using ribosome profiling and proteomics. Translons can regulate mRNA translation and encode micropeptides that contribute to multiprotein complex formation, Ca^{2+} regulation in muscle, and signaling during embryonic development. However, identification of translons has been limited to cell lines or large organs due to high input requirements for conventional ribosome profiling and mass spectrometry. Here, we address this challenge using Ribo-ITP on difficult-to-collect samples like microdissected hippocampal tissues and single pre-implantation embryos. Comparative analysis of more than a thousand ribosome profiling datasets across a wide range of cell types revealed distinct sample specific expression patterns of the detected translons. To test the translational capacity of the identified translons, we engineered a translon-dependent GFP reporter system and detected expression of translons initiating at near-cognate start codons in mouse embryonic stem cells (mESCs). Mutating the translons in mESCs identified a small proportion that negatively impacted growth. Taken together, we present a proof-of-concept study to identify non-canonical translation events from low input samples which can be applied to cell and tissue types inaccessible to conventional methods.

Keywords: Ribo-ITP, translon, ribosome profiling, low input

Introduction:

In the last decade, emerging evidence strongly indicates pervasive translation in regions of the genome hitherto annotated as non-protein coding. Transcriptome-wide approaches based on ribosome profiling and its variants[1,2] initially identified non-canonical translation events at scale. Although proteome based approaches also aided peptide-oriented identification[3–5], ribosome profiling remains the primary method for discovering non-canonical translation. These translated open reading frames (ORFs) and their encoded peptides can have functions in multiple cellular processes[6–20]. To determine the potential function of the encoded peptides, a number of experimental methods have been developed, including CRISPR-Cas9 pooled knockout screens, reporter assays, phylogenetic analysis, and affinity ligation based techniques[17,21–25]. These complementary approaches revealed context dependent functions of many non-canonical translational events from embryos to the brain.

Nomenclature of non-canonical translation events is associated with the term 'ORF,' which remains a point of confusion. In this study, we adopt the term 'translon' to refer to ORFs with evidence of translation[26]. Although the annotated protein coding region in a transcript is also a translon, we will refer to these as CDS for simplicity. Importantly, the catalog of micropeptides and translons across human cell types remains incomplete. This has inspired a community-wide effort for annotation and characterization of non-canonical translation events[27].

A major challenge in identifying translons has been the lack of strategies for discovering these in limited input samples such as single cells and patient samples, limiting our knowledge of their diversity. Recently, there have been efforts to increase the efficiency and reduce the input requirements of ribosome profiling[28,29]. Among these, RIBOsome profiling by IsoTachoPhoresis (Ribo-ITP)[28] is particularly promising for profiling non-canonical translation in limited-input contexts. By leveraging an isotachophoresis-based RNA concentration method followed by size selection on a microfluidic chip, Ribo-ITP minimizes sample loss and enables efficient capture of ribosome-protected fragments from as few as a single cell, leading to high coverage of transcripts. This combination of sensitivity and scalability makes it uniquely capable of uncovering translational events in rare or heterogeneous cell populations.

Here, we present a proof-of-concept study applying Ribo-ITP to microdissected CA1 subregions of single acute hippocampal slices and single 16 and 32-cell embryo stages of mouse preimplantation development. These types of samples have been inaccessible to conventional ribosome profiling, limiting discovery of cell-type or stage-specific translons. Using Ribo-ITP, we detected thousands of previously unannotated translons, many of which initiate from non-canonical start sites and encode short peptides with distinctive amino acid compositions. Meta-analysis using a compendium of over a thousand ribosome profiling datasets revealed widespread expression of these translons across diverse cell types. Finally, functional assays in mESCs demonstrate that a subset of these translons can actively drive translation and that disruption of some translons can impair cellular fitness. Together, our results establish a framework for mapping non-canonical translation in samples previously inaccessible due to input limitations.

Results:

Ribo-ITP enables identification of non-canonical translation events from microdissected mouse hippocampus

Activity dependent translational changes in neurons are essential for synaptic plasticity[30–34]. Ribosome profiling of long term potentiation (LTP) induced acute hippocampal slices has been limited due to technical difficulties involved in scaling the laborious experimentation to meet the input tissue requirements[35]. Here, we applied Ribo-ITP to measure translation in mouse hippocampal CA1 interneurons. We performed theta burst stimulation to induce LTP in CA1 pyramidal interneurons in acute hippocampal slices and manually microdissected and flash-froze the potentiated region at 10 mins, 30 mins, and 60 mins post-LTP induction along with unstimulated control tissue (Fig1A, S1F). Based on the dimensions of the microdissected region (see Methods), we estimate that these would contain just a few thousand cells.

Ribo-ITP resulted in libraries with ribosome protected fragment (RPF) length distributions centered at 28 nucleotides (nt) and showed enrichment at annotated start and stop codons (Fig1B, C, S1A-C). Most RPFs mapped to the annotated coding regions of transcripts (Fig1D, S1D) with 77% mapping to the first frame of the CDS. The number of RPFs mapping to annotated coding regions across all 12 replicates was reproducible (Spearman correlation coefficient > 0.83; Fig S1E). These results suggest that Ribo-ITP enables isolation of RPFs from microdissected, LTP-induced CA1 neurons from single slices.

To identify translons, RPFs can be either mapped to the genome or the transcriptome. Genome mapping is the most commonly used approach as it helps identify a broader range of translation events. However, RPFs are short (~28nt) and can map to multiple regions in the genome, increasing mapping errors. Transcriptome mapping, on the other hand, can identify only a subset of translons but is less prone to mapping errors. In this study, we mapped RPFs to both the genome and the transcriptome. Pooling the deduplicated reads from the 12 replicates resulted in over 20 million genome mapping and 22 million transcriptome mapping unique molecules which were used for identifying translons (Methods).

We first used RiboTISH, a genome-based pipeline and identified a total of 19,475 translons[36]. Out of these, 12,795 were known CDSs, and the remaining 6,680 were novel translons with respect to the annotations used. Of the novel translons, 1,227 initiated upstream to the CDS, 1,954 initiated downstream to the CDS, and 69 were encoded in long intergenic non-coding RNAs (lincRNA) (Fig1F) and belong to categories overlapping to our transcriptome based approach. Of the remaining novel translons, 3,265 were truncations of CDSs, 99 were internal but out of frame with respect to the CDS, 5 from antisense transcripts, 5 from putative protein coding transcripts, 3 from small non-coding RNA transcripts like miRNA, 51 from pseudogenes, and 2 from intronic regions of genes (TableS1).

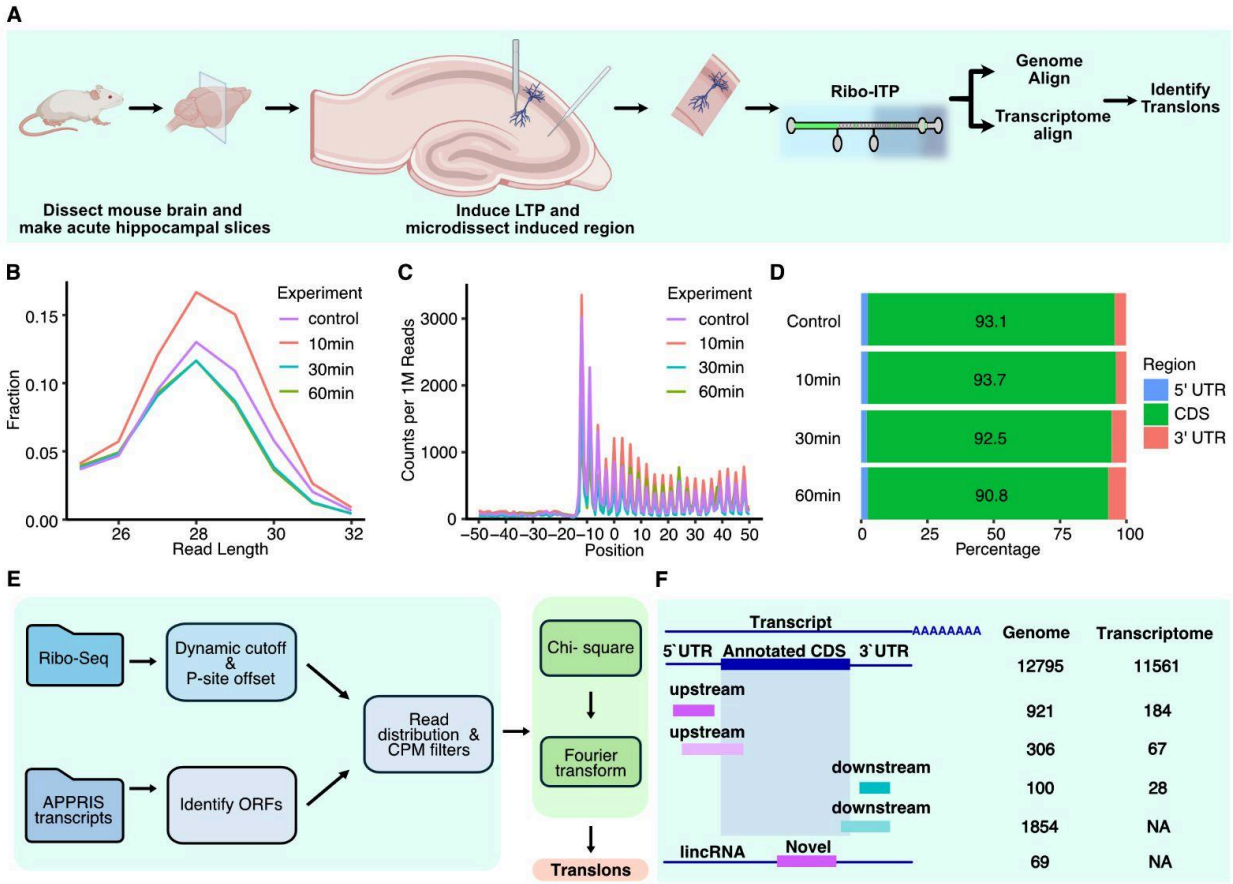


Fig1: Ribo-ITP enables identification of non-canonical translation events from low input tissue samples

A, Schematic of the experimental procedure for performing Ribo-ITP of LTP induced microdissected CA1 region from acute slices of mouse hippocampal sections. B, Normalized length distribution of the transcriptome mapping RPFs after merging replicates by time point. C, Ribosome occupancy around the translation start site after merging replicates by time points. Translation start site is denoted by the position 0. Aggregated read counts (y axis) relative to the start site are plotted using the 5'end of the read. D, Distribution of reads across transcript regions (5' UTR, CDS and 3' UTR) are shown. E, Schematic of the transcriptome based translon identification strategy developed in this study. F, Schematic showing the different types of events identified (left) and the corresponding number of ORFs identified for each event using the genome and transcriptome based approach.

We next developed a transcriptome based approach which uses a .ribo file[37] as input and detects translons (Fig1E). We have used two different approaches to assess the periodicity of RPFs on a given ORF — based on a chi-square test or a Fourier transform analysis (see Methods). The chi-square analysis identified 1390 translons upstream and 340 downstream of annotated CDS, along with 13,795 expressed CDSs (more than 1 CPM reads). We then assessed the expression and periodicity of footprints using a database (RiboBase) of over ~2.35 billion transcriptome mapping reads from ~1600 highly periodic mouse ribosome profiling

datasets from 37 different cell and tissue types (see methods). Given the high coverage of transcripts within these datasets, we applied a Fourier transform to further assess the periodicity of the translons that passed the chi-square test (see Methods). 251 upstream ORFs and 28 downstream ORFs had evidence of periodic RPF reads in the database (Fig1F, Table S2). These results establish that thousands of non-canonical translons can be detected from microdissected CA1 regions from single hippocampal acute slices using Ribo-ITP and that many of them are detectable in other cell types as measured by conventional ribosome profiling.

Ribo-ITP enables identification of non-canonical translation events from single mouse embryos

Having identified translons in thousands of neurons from mouse hippocampal slices, we next tested if this approach would help identify translons from even fewer cells. Ribo-ITP had previously helped identify intricate translational regulation from single mouse embryos[28,38]. In zebrafish, ribosome profiling from hundreds of embryos at different stages of development had enabled discovery of a peptide called Toddler, which is essential for gastrulation and normal development[6]. Yet, the translons expressed in 16- and 32-cell stages where mice embryos undergo early asymmetric division leading to distinct cell lineages in a blastocyst[39,40] are not catalogued.

Therefore, we employed Ribo-ITP to profile single 16- and 32-cell stage preimplantation mouse embryos to identify translons (Fig2A). We treated stage-matched embryos with Harringtonine, a translation inhibitor which stalls initiating and early elongating ribosomes[41,42]. In total, we profiled 15 single embryos from 16-cell and 32-cell stages using Ribo-ITP. These libraries had RPF length distributions centered around 28-30 nt (Fig2A-D) and showed enrichment at annotated start codons (S2E-H). Most of the RPFs mapped to the annotated coding regions of transcripts (Fig2B), with 67% raw reads mapping to the first frame of the CDS. Upon Harringtonine treatment, reads mapping in the 5' UTR junction, i.e. 50 nt surrounding the annotated CDS start site, increased at least 2-fold (Fig2C, S2J,K).

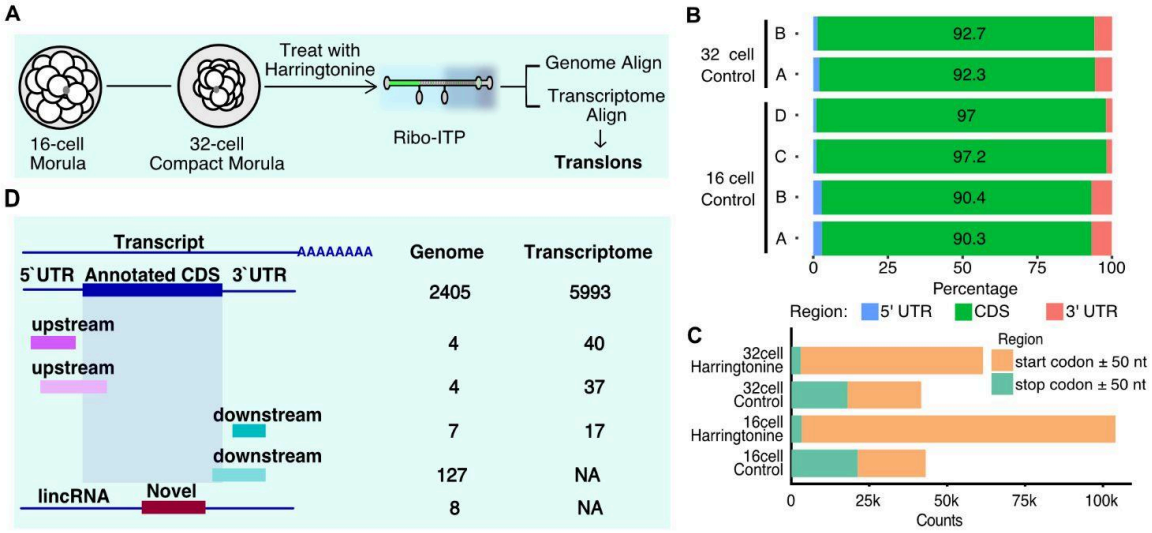


Fig2: Ribo ITP enables identification of non-canonical translation events from single mouse embryo

A, Schematic of the experimental procedure for performing Ribo-ITP of 16- and 32-cell single embryos. B, Percentage of reads mapping to the different transcriptomic regions (5'UTR, CDS and 3'UTR) in 16-cell and 32-cell control embryos. C, Distribution of reads mapping in the 50 nt region surrounding the translation start site the raw counts are merged from multiple replicates per stage. D, Schematic showing the different types of events identified (left) and a table of corresponding number of ORFs identified for each event using the genome and transcriptome based approach.

Pooling all the untreated libraries together, we mapped over half a million unique RPFs to both the genome and transcriptome. We used RiboTISH and our algorithm to identify translons (Fig 2D, Table S3). RiboTISH identified 2555 translons, of which 2405 were CDS, 8 initiated upstream to the CDS, 134 initiated downstream to the CDS, and 8 were encoded in lincRNAs. Transcriptome-based approach identified 736 translons upstream and 323 downstream of annotated CDSs (TableS4), while 5993 expressed CDSs (more than 1 CPM reads). Out of these, 66.7% upstream translons and 28.5% downstream translons had RPFs mapping around the predicted start site (±2 codons) in the Harringtonine-treated datasets. Fourier transform based analysis identified high periodicity across 77 upstream translons and 17 downstream translons across studies in RiboBase (Fig2D). Taken together, these findings suggest that Ribo-ITP can be used to identify translons from single preimplantation mouse embryos. Fourier transform analysis across multiple ribosome profiling datasets further highlighted a subset of translons with consistent periodicity. Nonetheless, remaining candidates may also represent bona fide translons, but validating their expression needs further experimentation.

Properties of identified translons

We next sought to characterize physicochemical properties of the identified translons from both the datasets. As there were few genome annotated embryonic translons we focused further analysis on the transcriptome annotated translons only. The median length of hippocampal translons was 46 amino acids (aa) and 51 aa for the transcriptome annotated (Fig3A) upstream and downstream translons respectively and 53, 150 and 65 aa for genome annotated (FigS3A) upstream, downstream and lincRNA translons respectively. The median length of the embryonic translons was 35 aa for upstream and 31 aa for downstream translons, respectively (Fig3D). In both the datasets, most of the transcriptome annotated upstream translons initiated at CTG, a near-cognate start codon, while ATG was the most common codon for most downstream translons (Fig3B,E) while ATG was the preferred initiation codon for genome annotated translons (FigS3B).

Given that the amino acid composition of a peptide affects its stability[43], we compared the amino acid composition of the identified translons with the annotated CDS (see Methods). The amino acid composition of the encoded peptides was similar for both the neuronal (see Fig3C, S3C for genome-annotated) and embryo translons (Fig3F). Specifically, in upstream translons, the amino acids arginine and proline were significantly enriched (*p-value*: hippocampal (h) $_{Arg}$ = 2.20×10^{-21}, embryonic (e) $_{Arg}$=3.93×10^{-9}, h_{Pro}=0.00326, e_{Pro}=0.00621) while aspartic acid, glutamic acid, asparagine, isoleucine, methionine, valine, tyrosine, histidine, and lysine were significantly depleted (*p-value*: h_{Asp}=1.30×10^{-16}, e_{Asp}=8.22×10^{-9}, h_{Glu}=1.77×10^{-13},

$e_{Glu}=7.13\times10^{-7}$, $h_{Asn}=4.61\times10^{-24}$, $e_{Asn}=6.20\times10^{-10}$, $h_{Ile}=3.26\times10^{-24}$, $e_{Ile}=5.77\times10^{-6}$, $h_{Met}=4.88\times10^{-8}$, $e_{Met}=0.00144$, $h_{Val}=2.91\times10^{-14}$, $e_{Val}=0.0400$, $h_{Tyr}=2.12\times10^{-32}$, $e_{Tyr}=1.52\times10^{-10}$, $h_{His}=0.00278$, $e_{His}=0.0288$, $h_{Lys}=9.84\times10^{-17}$, $e_{Lys}=1.28\times10^{-10}$). This is consistent with observations in post-mortem human brain[14] where the identified translons were enriched in aromatic amino acids, and arginines but depleted in polar amino acids and lysines. In the downstream translons, we observed a significant depletion for alanine (*p-value*: h= 0.0162, e=0.0284). Taken together, these results suggest that the identified translons have dramatically different physicochemical properties compared to annotated CDS.

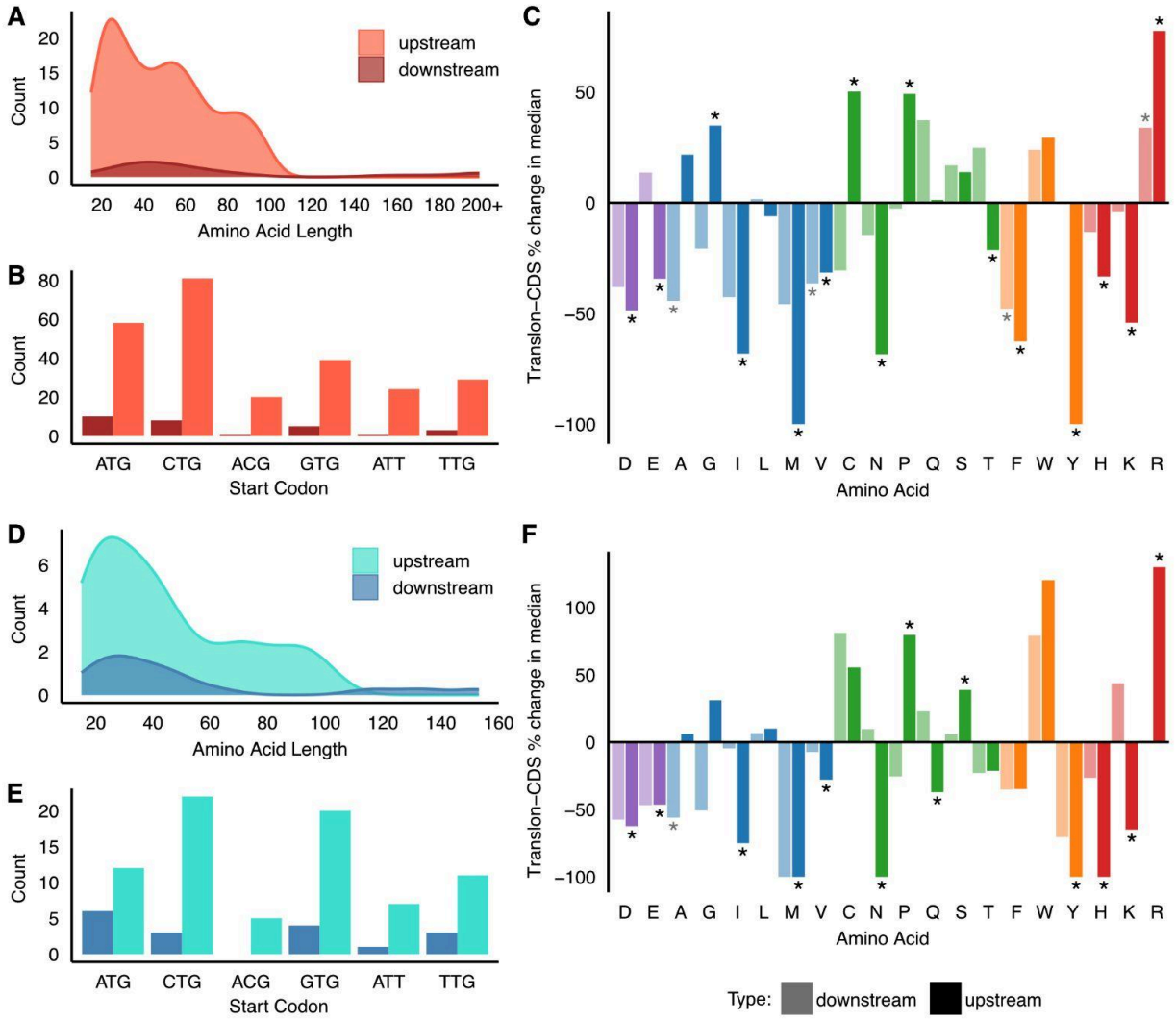


Fig3: Properties of embryonic and hippocampal translons.
*A, Density plots showing the amino acid length distribution of the peptides encoded by the translons identified in the hippocampal dataset. B, The number of translons identified with cognate and near-cognate start codons in the hippocampal dataset. C, Bar graph showing the fractional difference in the upstream (dark) or downstream (light) hippocampal translons with the CDS for the individual amino acids. Colors indicate amino acid type. * indicates Wilcoxon's rank-sum test p-values < 0.05. D, Density plots showing the amino acid length distribution of the peptides encoded by the translons identified in the embryo dataset. E, The number of translons*

*identified with cognate and near-cognate start codons in the embryo dataset. F, Bar graph showing the fractional difference in the embryonic upstream (dark) or downstream (light) translons with the CDS for the individual amino acids. Colors indicate amino acid type. * indicates Wilcoxon's rank-sum test p-values < 0.05.*

Expression patterns of non-canonical translons

Given that translons identified in both the datasets have significantly different properties, we next asked if these also have unique expression patterns. To address this, we first focused on translons identified in the 16-cell and 32-cell stages of mouse embryonic development. Embryonic upstream translons had robust expression (>10 CPM median reads) across multiple tissue and cell types in contrast to embryonic downstream translons (<1 CPM median reads in all) (Fig4A, Table S5). These observations suggest that the embryonic upstream translons are not specific to preimplantation development and instead are broadly expressed in a wide range of cell types.

Upstream and downstream translons can influence the translation of the CDS [18,19,23,44]. Therefore, we assessed the relationship between ribosome occupancy of these translons with that of the CDS. Specifically, we calculated the partial correlation between the RNA-abundance-matched ribosome occupancy of the CDS and the translon. For the embryonic translons, we observed that there was a weak but positive partial correlation between the ribosome occupancy on the upstream translon and the CDS (mean Spearman correlation coefficient = 0.25, Wilcoxon *p-value* = 1.81×10^{-13}; Fig4B). This observation is consistent with previous results related to upstream translons identified in postmortem human brains and in zebrafish embryonic development [14,18]. The downstream translon expression was not correlated with the CDS (mean Spearman correlation coefficient = 0.11, Wilcoxon *p-value* = 7.63×10^{-4}; Fig4C). The overall low expression of most downstream translons may mask their expression relation with the CDS in contrast to their activating role [44]. Using reporter assays Wu et al., showed that the presence of downstream translons but not the number or sequence positively modulates translation efficiency of the CDS.

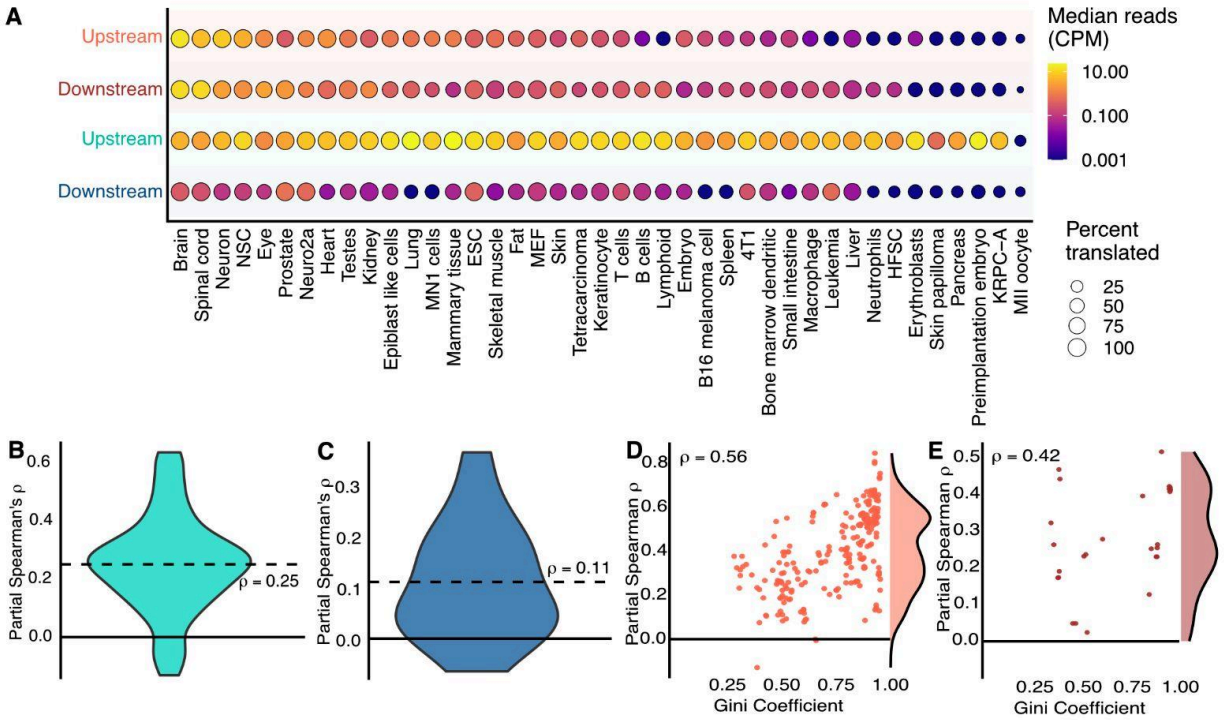


Fig4: Expression patterns of embryonic non-canonical ORFs
A, The average expression of all the identified translons across the different cell types in RiboBase is plotted. The color coding of the dots represents the median expression of all the translons aggregated for each cell type and the size of the dot represents the percentage of expressed translons. The cooler color scheme on the y-axis represents the hippocampal translons while the warmer color scheme represents embryonic translons. B, Spearman's R values for the partial correlation between the RNA normalized embryonic upstream translons and their corresponding CDS ribosome occupancy across all RiboBase studies. C, Spearman's R values for the partial correlation between the RNA normalized embryonic downstream translons and their corresponding CDS ribosome occupancy across all RiboBase studies. D, Correlation between the Spearman's R values for the partial correlation between the RNA normalized neuronal upstream translons and CDS ribosome occupancy across all RiboBase studies with Gini coefficient. Distribution of Spearman's R value for partial coefficients is shown on the right. E, Correlation between the Spearman's R values for the partial correlation between the RNA normalized neuronal downstream translons and CDS ribosome occupancy across all RiboBase studies with Gini coefficient. Distribution of Spearman's R value for partial coefficients is shown on the right.

We next looked at the expression patterns of the hippocampal translons. Surprisingly, the upstream and downstream translons were robustly expressed (>1 CPM median reads) only in brain, spinal cord, neurons, neural stem cells, and heart (Fig4A, TableS5). This suggests that the hippocampal upstream translons might have functions specific to neuronal tissue types unlike embryonic ones. As we had over ~40X more reads in the hippocampal dataset, we might have detected more cell type specific translons than single embryos where most highly expressed and general events could be detected. We next examined the partial correlations

between the RNA-normalized ribosome occupancies of these translons with the CDS. The mean partial correlation was higher for hippocampal upstream and downstream translons and the CDS (Fig4D,E [distribution on the right]) than the embryonic translons, with a Spearman correlation coefficient of 0.4 (Wilcoxon $p = 1.69\times10^{-41}$) and 0.23 (Wilcoxon $p = 1.49\times10^{-8}$), respectively. Further, these correlations showed a bimodal distribution. We hypothesize that this may relate to the tissue-specific expression of these translons. Indeed, upstream translons with higher tissue specificity (higher Gini coefficients) exhibited stronger positive correlation with CDS translation (ρ vs. Gini, *p-value* < 2.2×10^{-16}). For the downstream translons, this association was weaker but still significant (*p-value* = 0.02). Taken together, these results suggest that translational coordination between the upstream and downstream translons with the CDS is influenced by their expression specificity, with the hippocampal translons having functions in the nervous system.

Peptidomic evidence can provide further support for stability of the translon encoded peptides. Therefore we mined two proteomic datasets focusing on non-canonical translation in mice brains and identified peptide evidence for a total of 47 translons. This included 45 translons from Wang et al.,[45] who profiled various embryonic and adult mouse organs and 2 translons in Budamgunta et al.,[46] who profiled adult mouse striatum. This evidence from independent studies strengthens our identification approach (Table S6).

Non-canonical translons can drive translation of a reporter and affect cellular proliferation

Given the broad expression pattern of embryonic translons, we next explored their functional capacity, which remains challenging to establish. As a first step, we asked if these sequences have sufficiently strong activity to robustly drive the expression of a reporter. We selected seven embryonic upstream translons and cloned the sequence from the start of the transcript until the stop codon of the translon ORF in frame with no start-GFP (Fig5A). Consequently, we expect to detect fluorescence only if the ORF is robustly translated. We had three controls, 1. a positive control i.e. GFP with ATG (start-GFP), 2. a negative control i.e. GFP with no ATG (nostart-GFP) and 3. insulin transcript (5` UTR + CDS), which codes for a 110 amino acid pro-peptide in frame with nostart-GFP. Upon overexpression in mESCs, the positive control had robust GFP expression (Fig. 5B), the negative control did not have GFP expression (Fig. 5C) while the insulin control showed weak GFP expression (Fig5D), which may be attributed to factors like pro-peptide cleavage and tissue-specific expression of insulin in pancreatic cells.

Four out of the seven tested embryonic translons drove at least some GFP expression when transfected in mESCs. Translons on mRNAs that encode Eif5 and Clpx drove strong GFP expression in mESCs (Fig5E,F), while translons on Npm1 and Hdgf mRNAs drove weak GFP expression in mESCs (Fig5G-H). Translons on Rps12, Olfr1217 and Top1 mRNAs did not show any GFP expression (Fig5I,S). The Eif5 translon is the most highly expressed upstream translon in 16-cell and 32-cell embryos; moreover a uORF in human Eif5 transcript is a known regulatory translon[36,47]. The upstream translon in the Clpx transcript has not been previously reported to the best of our knowledge. Clpx translon initiates at a CTG codon with no in-frame ATG, suggesting that translons with non-canonical initiation codons can drive robust translation.

We reasoned that if a translon is essential for cellular growth then mutating it might result in slower growth in a model system such as mESCs. To test this, we designed a pooled CRISPR-Cas9 screening strategy to introduce indels in 67 translons identified by the chi-squared filter (see Methods). We aimed to introduce mutations targeting the region of the translons that is non-overlapping with the CDS. To detect even subtle growth defects, mESCs were grown for approximately 20 doublings (day 10) after selecting for cells expressing gRNA and Cas9 (Fig5L). MAGeCK analysis of the guide RNA abundances on day 1 and day 10 revealed that perturbations in 4 out of the 66 translons led to reduced growth ($p<0.05$, FDR<0.25) (Fig5M, Table S7).

A wide range of mutation types are generated in pooled CRISPR-Cas9 screens. The resulting heterogenous allele populations likely do not affect growth equally. To orthogonally validate results from the screen, we generated three clonal populations harboring unique versions of the Phgdh upstream translon, which had the most significant effect: (+/+), which did not have any mutations; (+/-), which had an inframe stop codon introduced in a single allele; and (-/-), which had mutations in both the alleles, resulting in a frameshift in one allele while a truncation on the other allele. Growth curve analysis of these clones suggested that (-/-) (both the Phgdh upstream translon alleles targeted) had the most pronounced reduction ($p = 0.00014$) in cellular growth as compared to parental mESCs (Fig5N). This further supports the observation that the embryonic upstream translons identified in the mESC perturbation screen likely have subtle effects on cell growth.

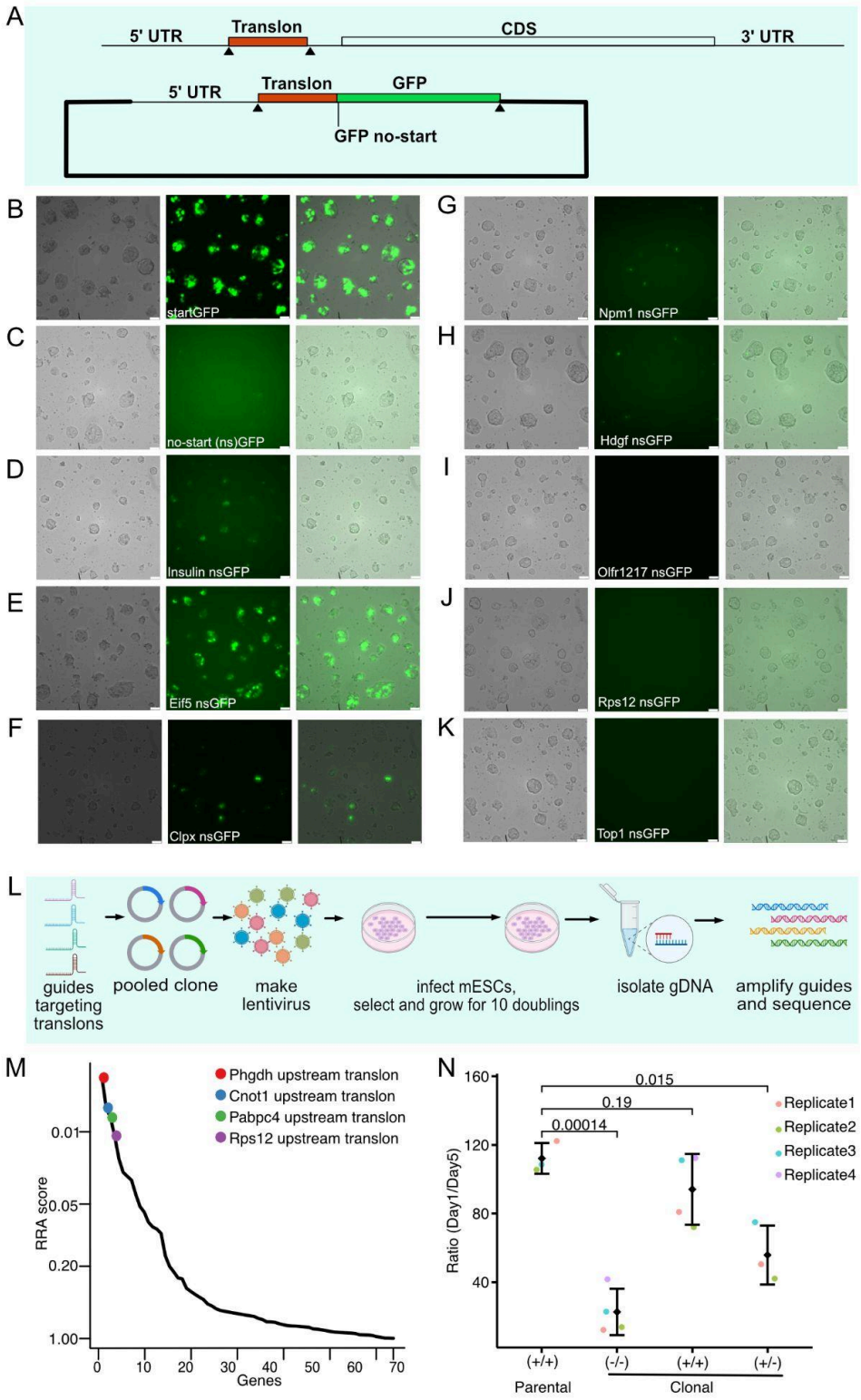


Fig5: Non-canonical ORFs contribute to cell proliferation.

A, Schematic of the reporter design for GFP based translational capacity reporter. B-G, Fluorescence, DIC and merged images of mESCs transfected with the 5`UTRs containing the

candidate ORFs cloned in the GFP reporter. Scale bar width is adjusted for effective visualization purpose, scale bar = 50µm. H, Schematic of the CRISPR screen experimental setup. I, Histogram of the RRA scores for uORFs tested. The highlighted ORFs have a low RRA score with a FDR less than 25%. J, Table of the mutations introduced in the clones generated for validation. K, Graph showing the ratio of live cells on Day1 to Day5 for each Phgdh clone tested and parental mESCs. Two-sided Student's-t test p values are annotated on the plots.

Discussion:

We present proof-of-concept experiments demonstrating that Ribo-ITP enables the detection of translons in samples that are inaccessible to conventional ribosome profiling or proteomics due to challenges in collecting sufficient quantities of material. This method yields high-quality, reproducible ribosome occupancy profiles with strong 3-nt periodicity, essential for identifying translons. Despite conventional ribosome profiling using several orders-of-magnitude more input, of the 1625 previously published ribosome profiling studies, only 70% achieve a similar or better 3-nt periodicity as we have with all our datasets. While recent efforts proposed to build atlases of micropeptides and translons using conventional approaches[27], such strategies will remain limited to abundant cell types and large tissues. The diversity of translons in rare, biologically important cell populations would therefore remain inaccessible. Our study provides a path forward for capturing non-canonical translation in these contexts.

In this study, we used Ribo-ITP to measure translation from the microdissected CA1 region of a single hippocampal slice, which is an orders-of-magnitude reduction in input compared to previous ribosome profiling studies which pooled ~40-50 slices from multiple animals[30,31,35,48]. This reduced input lowers both experimental burden and cost. Similarly, previous studies in zebrafish required 400-600 embryos per replicate to identify micropeptides [6,49]. Such large-scale collection is either impractical or currently impossible in many systems, including *C. elegans*[50] and mammals[28].

Our transcriptome pipeline allowed us to leverage a large compendium of uniformly processed ribosome profiling data (RiboBase) [51]. This rich resource contains paired data of 995 mouse Ribo-seq experiments from different cell and tissue types, enabling expression analysis of upstream and downstream translons across diverse contexts. Most translons detected in embryos exhibited broad expression across cell and tissue types, while those identified in the CA1 region of the hippocampus were largely specific to the nervous system. This analysis suggests that translons may support both housekeeping functions and cell type-specific roles, particularly in the nervous system. These observations raise questions about the regulation and function of tissue-specific translons: What mechanisms underlie their specificity? Do they contribute to cellular homeostasis? Expanding the compendium of ribosome profiling datasets will likely enable uncovering the full diversity and biological significance of these translons.

The peptides that may be encoded by the identified translons were enriched for proline and arginine and depleted in methionine, aspartic acid, histidine and lysine. Noncanonical proteins and canonical protein isoforms starting with non-methionine amino acids are enriched for degradation-related features like high C-term hydrophobicity and C-term degrons compared to all canonical proteins[43]. Arginine in the C-terminal can act as degron leading to degradation of the protein[52]. Stop codon read-through leads to translation into the 3'UTR which results in protein with 3' UTR encoded peptides or poly-lysine tracks which destabilize the protein[53,54].

Therefore it is likely that some of the translon-encoded peptides with non-canonical properties degrade quickly after their translation.

Stable non-canonical proteins show phenotypic effects when knocked-out in cell lines[5,21]. Our CRISPR-Cas9 screen identified that mutations in 4 out of the 66 tested upstream translons subtly hampered mESC growth. Peptides like Toddler, Elabella and Pri are encoded by non-canonical translons and regulate embryonic development in zebrafish and Drosophila [6,55–57]. Therefore it is possible that the embryonic non-canonical translational events we identified can affect mouse embryonic development as well. Identifying the molecular mechanisms of these events will be a future direction of our study.

Upstream translons are known to repress the translation of the CDS by controlling ribosome loading on the CDS, ribosome stalling on the uORF, or activation of NMD [15,18–20,58]. We observed that embryonic upstream translon expression was poorly but positively correlated with the CDS expression. Although a poor but positive correlation with CDS expression was reported by others the reason for this surprising observation is not yet known[14,18,59].

Finally, we want to highlight some important limitations of our study. First, while we observed subtle growth phenotypes upon disruption of select upstream translons, these effects are modest and suggest that many non-canonical translation events act in concert with other regulatory mechanisms rather than functioning as primary effectors. Given that CRISPR-Cas9 screens generate a heterogeneous pool of allelic series that may modulate the function of the CDS in addition to the targeted translon, disentangling functional signals remains a technical and conceptual challenge. Second, although many of the identified translons are expressed during early development, establishing their precise roles in embryogenesis remains highly challenging especially given the limitations of multiplex phenotypic assays. In this study, we have decided to not focus on activity-dependent changes associated with LTP. While technically reproducible, we observed highly variable expression patterns across timepoints, possibly due to variability in microdissection of the LTP-activated region. Comprehensive profiling of LTP-associated translation would likely require more precise anatomical targeting and increased temporal and biological replicates.

Materials and Methods:

Animals

All procedures were approved by the Institutional Animal Care and Use Committee at the University of Texas at Austin. C57BL6/J mice (Jackson Labs, RRID:IMSR_JAX:000664) were used for ribosome profiling experiments.

Collection of mouse brain tissue for ribosome profiling

Mice were deeply anesthetized with a mix of ketamine (90 mg/kg; TW Medical, Decatur, GA) and xylazine (10 mg/kg; Animal Health International, Loveland, CO) intraperitoneally. Mice were transcardially perfused with ice-cold, oxygenated cutting solution, containing: 205mM sucrose, 25mM sodium bicarbonate, 2.5mM KCl, 1.25mM sodium phosphate, 7mM MgCl2, 7mM D-glucose, 3mM sodium pyruvate, 1.3mM ascorbic acid, and 0.5mM CaCl2. Brains were removed and hemisected, and the dorsal portion of the brain was removed using a scalpel blade, angled at 20°. The dorsal cut side was then mounted in the slicing chamber of a Leica

VT1200 vibratome using superglue and submerged in ice-cold, oxygenated cutting solution. 250 mm oblique-horizontal hippocampal slices were made. Slices were incubated for 30 min at 37°C in holding solution containing: 125mM NaCl, 25mM sodium bicarbonate, 2.5mM KCl, 1.25mM sodium phosphate, 12.5mM D-glucose, 2mM MgCl2, 2mM CaCl2, 1.3mM ascorbic acid, and 3mM sodium pyruvate. Slices were then stored for at least 30 min in the holding solution at room temperature before being used for recordings. Reagents for electrophysiology solutions were purchased from Sigma Aldrich. Only slices from the middle of the dorsal-ventral axis of the hippocampus were used for recordings. Glass recording electrodes of 3–6 MOhms were pulled from thin-walled borosilicate filamented glass (Sutter BF150-110-10) using a P-1000 puller (Sutter). Electrodes were filled with artificial cerebrospinal fluid (ACSF) containing: 125mM NaCl, 25mM NaHCO3, 2.5mM KCl, 1.25mM NaH2PO4, 12.5mM D-glucose, 1mM MgCl2, and 2mM CaCl2. Data was acquired with a MultiClamp 700B amplifier (Molecular Devices) interfaced with a Digidata 1550B acquisition system and pClamp 11.1 software (Molecular Devices, San Jose, CA).

Slices were transferred to the recording chamber and continuously bathed in warm (32.5°C), oxygenated ACSF. Slices were visualized using a Zeiss Axio Examiner with Dodt contrast optics and epifluorescence, and a Zeiss Axiocam 503 infrared digital camera. Prior to recording, a small cut was made through the stratum radiatum at the border between CA3 and CA2 to reduce activation of recurrent connections. The recording electrode was placed in the stratum radiatum of CA1. Synaptic stimulation was delivered by a tungsten bipolar stimulating electrode (WE3ST30.1ASS, Microprobes) placed in the stratum radiatum near the CA2/CA1 border to stimulate the Schaffer collateral (SC) axons, and a constant current stimulus isolator (World Precision Instruments). First, slices were stimulated at 0.1 Hz, and field excitatory postsynaptic potentials (fEPSPs) were recorded for 10-30 min to ensure that responses were stable, indicating good slice health. Next, the stimulus intensity was adjusted to evoke minimal and maximal responses and an input/output curve. The stimulus intensity was then set at the mid-point of this curve such that either increases or decreases in response could be recorded. Baseline fEPSPs were then stimulated and recorded at 0.1 Hz for 10 min. Slices were then stimulated with a high frequency stimulus designed to induce LTP: 100 Hz for 1 s, repeated 4 times with a 20 s interval. Slices were then stimulated at 0.1 Hz for 10 mins, 30 mins or 60 mins, after which the slices were transferred to a new dish and the CA1 region was dissected under a dissection microscope. A slice that was approximately 1mm X 0.5mm was dissected using a sharp scalpel, transferred to a 1.5mL Eppendorf tube with minimal buffer carryover, flash-frozen in liquid nitrogen, and stored at -80°C.

The approximate number of neurons in the microdissected regions was estimated based on the volume of the microdissected region (1mm X 0.5mm X 0.2mm) and the neuronal density (80,000cells/μm^3) in CA1 region of mouse hippocampus [60].

Mouse embryo isolation and treatments

Eight-week-old C57BL/6 female mice were injected with 5 U Pregnant Mare Gonadotropin (PMSG) (Ilex Life Sciences). Forty-eight hours later, they were injected with 5 U Human Chorionic Gonadotropin (HCG) (Millipore Sigma) and immediately paired with C57BL/6J male mice. Embryos were flushed from the oviducts on day 2.5 post-fertilization and cultured in KSOM medium (K0113, Cytospring). Approximately half of the embryos at the 16-cell stage

were removed and remaining were cultured to the 32-cell stage. Embryos in both stages were separated equally into two experimental groups: KSOM + 2 µg/ml Harringtonine and KSOM + DMSO (vehicle). The embryo groups were cultured for 30 minutes, then rinsed through three drops of 20 mg/ml BSA. Single embryos from each group were placed in thin-walled PCR tubes (VWR 20170-012) with minimal liquid volume using a pulled Pasteur embryo transfer pipette and flash-frozen in liquid nitrogen. The tubes were stored at -80°C.

Ribo-ITP of microdissected mouse brain tissue and single embryos

Ribo-ITP was performed as described in Rao et al. [61]. The brain tissue and single embryos were lysed differently as mentioned; the flash-frozen hippocampal tissue was thawed on ice, and 10uL of lysis buffer (1% TritonX-100, 50mM MgCl2, 0.1 mg/mL cycloheximide) was added. The tissue was homogenized with a pestle until the solution was visibly clear. 4uL of the lysate was transferred to a 0.2mL PCR tube containing 1uL of 5X ITP lysis buffer (100 mM BisTris pH 7.2, 5.0% Triton X-100, 25mM MgCl2, 12.5 mM CaCl2, 500mM NaCl, 5 mM DTT, 0.5 mg/ml cycloheximide). For embryo lysis, flash-frozen single embryos were thawed on ice, 5µL of ITP lysis buffer (20 mM BisTris pH 7.2, 1.0% Triton-X100, 5mM MgCl2, 2.5 mM CaCl2, 100mM NaCl, 1 mM DTT, 0.1 mg/ml cycloheximide) was added to the droplet containing the embryo on the wall of the tube. For both the lysates, 1µL of RNaseI (Ambion) diluted 1:300 in the ITP lysis buffer was added and incubated at 37°C for 30 mins. 1µL of 0.7% SDS was added to stop RNaseI, followed by running the lysate on an ITP chip. Ribosome footprints were eluted by collecting the RNA fragments between 19nt and 36nt markers on the ITP chip. NGS libraries were made using Diagenode small RNA library preparation kit with the following modifications: a. 0.5µL PNK enzyme (NEB) was added along with the dephosphorylation reagent, b. The TSO was diluted 1:5, c. The cDNA was amplified for 16 cycles. The libraries were pooled and cleaned up with AmpureXP beads (1.5X) and purified on a 10% PAGE. The libraries were sequenced on a Novaseq 6000 (Illumina) instrument using the PE150 sequencing chemistry.

Identification of translated ORFs

The ribosome profiling libraries were processed using the Riboflow pipeline[28]. The RiboR package was used for making metagene plots, region coverage plots and read length distribution and RiboGraph for data QC and visualization [37,62]. For identification of translated ORFs from these transcriptome aligned files, all possible candidate ORFs were initially identified from the APPRIS release M25 (GRCm38.p6) transcriptome by extracting all in-frame start-to-stop codon combinations. Only non-overlapping dORFs and non-overlapping and overlapping uORFs were considered. ORFs in the 5'UTR were limited to a length of 11 to 100 amino acids. For ORFs overlapping the CDS, only the portion within the 5′ UTR was considered for downstream analysis. RPFs were aligned and corrected for the ribosome P-site using an offset determined from the peak of the start site metagene profile within a range of 11–16 nucleotides. A dynamic cutoff algorithm [51] was applied to include RPF lengths in which 85% of total CDS reads were contained. ORFs with fewer than 10 ORF reads per million experiment reads across all experiments, as well as those lacking significant 3-nucleotide periodicity based on chi-squared tests with false discovery rate (FDR) correction ($p > 0.05$), were excluded. To refine the set of in-frame ORFs within the same gene that share a stop codon, an additional filtering step was applied: longer ORFs were removed if all their reads were fully contained

within a shorter in-frame ORF. For each group of overlapping in-frame ORFs, the maximum number of reads across all experiments was computed. Only ORFs containing more than 75% of this maximum read count in every experiment were retained, and if none met this threshold, all ORFs in the group were kept. Among the retained candidates, ORFs with non-canonical start codons were removed unless all ORFs in the overlapping group had non-canonical starts, in which case all were kept.

For genome alignment, the trimmed and mouse rRNA and tRNA filtered .fastq files obtained from Riboflow pipeline were aligned to the mouse genome (GRCm38/ mm10) using STAR (2.5.2) aligner with *end-to-end* (no soft-clip) and *--outSAMattributes All* mode. The aligned reads with a mapping quality cutoff of 10 were deduplicated using UMI tools version 1.1.1 with "--read-length" parameter. The genome-aligned sorted and indexed bam files were used to run RiboTISH 0.2.7[36]. First, the read distribution around the start and stop site, p-site offset, and the 3nt periodicity of the data was checked using *ribotish quality* with the regular riboseq option. These files were then used to identify all translons using the *ribotish predict* command with *--longest, --seq*, and *--aaseq* options. For the single embryo dataset, `*--altcodons CTG,GTG,TTG,ATT,ACG*` and `--harr` options were additionally used, while for the microdissected hippocampal dataset, only default options with *--longest* option was used.

Amino acid composition

The amino acid composition between translons and CDS regions was compared with a Wilcoxon rank-sum test. The CDSs included were those of the same genes as the translons. The amino acid frequency was calculated for each translon and CDS, defined as the number of residues of each amino acid divided by the total number of residues. For each amino acid, a Wilcoxon rank-sum test was run, comparing the distribution of frequencies of the translon vs. CDS frequencies for the upstream and downstream translons. P-values were adjusted using Bonferroni correction. For visualization, the median frequency for each amino acid was calculated for the upstream and downstream translons, along with the CDSs. The percent change was calculated as the difference between the translon and CDS median divided by the CDS median for each amino acid.

Translon expression across ribosome profiling studies

Uniformly processed .ribo files were obtained from RiboBase[51], a curated repository of high-quality ribosome profiling datasets. All candidate ORFs were analyzed across available mouse experiments in RiboBase. For each Ribo-seq dataset, P-site offsets and read length cutoffs were dynamically determined as described above. Datasets were excluded if fewer than 85% of coding region reads mapped within the selected read length range.

To ensure sufficient depth for detecting cell line–specific signals, datasets with fewer than 100,000 reads in the read length range were removed. To evaluate translation in specific cellular contexts, reads(CPM) were calculated for each ORF in each experiment and averaged across all experiments annotated to the same cell line.

To assess 3-nucleotide periodicity, datasets were further filtered by quantifying the number of CDS-mapped reads in each reading frame for each read length of an experiment [51]. Experiments were removed if no read length had ≥75% of CDS-mapped reads aligned to a single reading frame. For each ORF, P-site coverage was aggregated across the remaining datasets. ORFs with a normalized read density below 0.1 reads per nucleotide were excluded.

A one-dimensional discrete Fourier transform (DFT) was applied to the aggregated, P-site–adjusted coverage vector for each ORF. The frequency spectrum was computed using the *scipy.fft.fft* function, which calculates the DFT using the fast Fourier transform algorithm. The five frequency components with the highest amplitudes were identified, and their indices were normalized by the length of the frequency spectrum to yield relative frequencies ranging from 0 to 1. ORFs were retained if at least one of these five dominant frequency peaks fell within the 0.32–0.34 range, corresponding to 3-nucleotide periodicity.

This method determines 3-nucleotide periodicity by combining two complementary strategies: filtering based on global frame mapping along with a frequency-domain analysis of periodicity across the full read coverage profile.

Correlation of ribosome occupancy of translons vs. CDS

To evaluate the relationship between translon and main CDS ribosome occupancy, we analyzed CPM-normalized paired RiboSeq and RNASeq read counts from RiboBase. A partial Spearman correlation was performed using the R *ppcor* package to compare translon CPM and CDS CPM values, while controlling for gene-level RNA-seq expression to account for differences in transcript abundance. Correlations were calculated separately for hippocampal and embryonic datasets, and for uORFs and dORFs within each. To assess whether the resulting correlation coefficients (Spearman's ρ) were significantly different from zero, we performed a Wilcoxon signed-rank test on the ρ values in each group.

The CPM values of each translon averaged across all datasets within each cell line were used to compute a Gini coefficient which determines the cell-line specificity of each translon. The Gini coefficient was calculated with the R *DescTools* package[63]. to assess whether cell-line specificity was associated with translon-CDS translation relationship, Spearman correlation was calculated between the Gini coefficients and the partial Spearman correlation coefficients for upstream and downstream translons. Translons containing no RPF reads in the CDS region in all datasets were excluded.

Cell culture

HEK293T cells were cultured in Dulbecco's Modified Eagle Medium (DMEM) with 10% fetal bovine serum (FBS) and 1% penicillin-streptomycin. C57/BL6 mouse embryonic stem cells were obtained from ATCC, acclimatized to non-feeder culture conditions, and later grown in 2i medium (Knockout DMEM with 15 % FBS, 100U/mL penicillin-streptomycin, 2mM glutamine, 1X non-essential amino acids, 0.15mM 2-mercaptoethanol, 100U/mL Lif, 3μM CHIR99021, 1μM PD0325901)

Cloning and expressing of GFP reporter

The GFP sequence with (startGFP) and without (nostartGFP) the start codon was amplified and inserted into the pLVX vector (Addgene #141367) between the BamHI and MluI restriction sites using In-Fusion® Snap Assembly (Takara). In the noStartGFP plasmid, all 5'UTRs were cloned in-frame with the nostartGFP (between the EcoRI and BamHI sites) using In-Fusion® Snap Assembly. Insulin 5'UTR along with the CDS was cloned as a positive control. The 5' UTR sequences of the seven transcripts harbouring translons were cloned until the stop codon of the translon to test their translation capacity. Actin 5'UTR, which did not have any predicted translon, was cloned in both StartGFP and noStartGFP plasmids. Sequence-confirmed individual plasmids were reverse transfected in a single cell suspension of mESCs with Lipofectamine 2000 (Invitrogen). Live cells were imaged using a STELLARIS 8 fluorescence microscope (Leica) 24 hours post-transfection. The brightness and contrast of the images were adjusted post-imaging for effective visualization.

Translon Pooled Knockout Screen

The genomic locations of the identified translons were used to design guide RNAs against the targets using CHOPCHOP v3 [64] and IDT's CRISPR-Cas9 guide RNA design tool. The following criteria were used for choosing the best 2 guides per translon: high on-target score, low mismatch score, and position on the predicted uORF. In case of overlapping ORFs, we chose guides which cut as far as possible from the annotated ORF. The list of guide sequences used is mentioned in Supplemental table S8. These guides were cloned into lentiCRISPR v2 (Addgene #52961) as a pool, as described for the GeCKO Lentiviral CRISPR Toolbox [65]. A list of the guides and oligos cloned for this screen is provided in Supplemental Table S8. Briefly, first the lentiCRISPRv2 vector plasmid (Addgene # 52961) was digested with BsmBI-v2 (NEB) for 1 hr at 37°C and heat-inactivated at 80°C for 10 mins. The digested product was separated on a TBE-Agarose gel, and a 8kb band was cut and purified using a gel extraction kit (Macherey-Nagel). The forward and reverse oligos for each guide sequence with flanking BsmBI sites (IDT) were phosphorylated with T4 PNK (NEB) and hybridized as individual pairs by incubating at 37°C for 30 mins and inactivation at 95°C for 10 mins, followed by ramping down the temperature to 4°C at a rate of 5°C/min. Equimolar quantities of all the hybridized guides were pooled together and ligated with 1µg linearised lentiCRISPR v2 plasmid at room temperature for 30 mins. Ligated product was transformed in 250µL of chemically competent NEB Stbl3 cells by heat shock at 40°C for 1 min followed by incubation on ice for 2 mins and recovery in 10 ml warmed SOC outgrowth medium (NEB) for 1 hr. All cells were plated on a bioassay plate and incubated at 37°C overnight. Transformants were collected by scraping all the colonies off the plate in LB+Amp, and plasmid was isolated using a plasmid maxiprep kit (Zymo) including endotoxin removal steps. This pooled guide library was then used for generating lentivirus. HEK293T cells were grown in T175 flasks and transfected with equimolar ratio of psPAX2, pMD2.G, and the pLentiCRISPRv2_pooled library plasmids using Lipofectamine 3000 reagent (Invitrogen) according to manufacturer's protocol. Viral titer was calculated by infecting mES cells with serial dilutions of virus followed by puromycin selection and colony counting.

1.3 x 10^5 mESCs were infected in a T25 flask with the pooled sgRNA lentivirus library at a multiplicity of infection of 0.3 such that there was >100X representation of each guide. Next day,

the media was replaced with 2i ES growth media, and cells were allowed to grow for 24 hours. Infected cells were selected using 2i ES growth media + 0.3µg/mL puromycin for the next 48 hours. The surviving cells were seeded in a T75 flask (considered as Day 1 for the screen) and kept on selection for 8 days followed by recovery in 2i medium for 2 days. A portion of Day 1 cells was stored at -20℃ as input. After 10 days of growth (approximately 20 cell division cycles), cells were harvested and genomic DNA was isolated using Zymo mini DNA isolation kit. Genomic regions with the guide sequences were amplified from 5µg gDNA in a 100µL PCR reaction divided into four 25µL reactions for 22 cycles using custom made Illumina indexing primers to generate amplicon libraries (Supplemental table S8). PCR products were cleaned with AmpureXP beads and quantified on a bioanalyzer. Guide amplicon libraries were sequenced using NovaSeq 6000 (Illumina) with a PE150 chemistry.

Analysis of uORF Pooled Knockout Screen

NGS reads were trimmed to remove vector-specific PCR handles using Cutadapt[66] v1.18 with the following parameters:
"-g ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTGGAAAGGACGAAACACCG -a GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCCAATTCCCACTCCTTTCAAGACC". Trimmed reads were aligned to all guide RNA sequences used in the screen using bowtie2[67] v2.3.4.3. Count table was made using a custom script which was used to calculate a ratio of guide abundance on day 1 to guide abundance on day 10. Further, MAGeCK analysis was also performed to identify significantly under and over represented guides using the count table with the following parameters: --paired --adjust-method fdr and --sort-criteria negative.

Preparation of Phgdh 5` UTR translon knock-out clones and cell lines

To make the Phgdh knock-out single clones, both sgRNAs used in the knockout screen targeting Phgdh were ordered individually from Synthego (Supplemental table S8). Actively growing mESCs were trypsinized and $2x10^6$ cells were pelleted. 600pM of the guide was mixed with 300pM Cas9 protein (NEB) in Cas9 buffer (20mM HEPES-KOH pH 7.5, 150mM KCl, 1mM MgCl2, 10% glycerol [v/v]) and incubated for 30 mins at room temperature to form the RNP complex. 65 µL of Lonza SF Cell Line Nucleofector Solution was added to the RNP mix to make a total volume of 104 µL. This mixture was added to the mESC pellet, mixed gently, transferred to a Lonza 4D XL nucleofection cuvette, and electroporated using the K562 cell protocol in a 4D Nucleofector. Cells were allowed to recover in the cuvette for 8 mins and cultured in complete growth medium. After 72 hrs, genomic DNA was isolated from 10,000 cells, Phgdh locus was amplified by PCR using Phgdh_mut_chk primers (Supplemental table S8) for 35 cycles and sequenced by Sanger sequencing. To check the perturbation efficiency, Synthego ICE CRISPR Analysis Tool was used. To isolate single clones, 1000 cells were serially diluted in a 96 well plate such that at least 8 wells had single cells. These were allowed to grow until a single colony was visible. Multiple wells with single colonies were selected and scaled up. gDNA was isolated and mutants were identified by Sanger sequencing.

Growth curve of single clones

Phgdh knock-out clonal mESCs were plated in 24 well plates at a density of 2000 cells/well in 5 replicates. Cells were incubated at 37℃ for 5 days. At every 24-hour interval, cells from one well

were trypsinized, trypan blue viability staining was performed, and cells were counted using a Luna automated cell counter to find the number of live cells. The ratio of the number of live cells on Day1/Day5 was calculated and plotted.

Data Availability

All ribosome profiling raw data sets and the processed .ribo files generated in this study have been submitted to the Gene Expression Omnibus GSE300869, GSE300832. All custom scripts used to perform bioinformatics analyses are available at GitHub (https://github.com/up114/translons) and as Supplemental Code.

Competing Interest Statement

The authors declare no competing interests.

Acknowledgements

We thank W. Shawlot and the UT Austin Mouse Genetic Engineering Facility for their contributions. This work was supported by the National Institutes of Health grants (R35GM150667; HD110096), as well as a Welch Foundation grant (F-2027-20230405) (C.C.). Figures were generated using BioRender.com under a publication license (JO26GB978U). All original text in this paper was authored by the researchers. Additionally, we acknowledge the assistance of a Large Language Model (OpenAI) for suggesting edits aimed at improving clarity and grammar.

Author contributions

V.G. wrote the original manuscript. U.P., M.H., and C.C. participated in reviewing and editing the manuscript. V.G. and C.C. conceptualized the study design. V.G. and U.P. generated the figures for the manuscript. V.G. performed all the ribosome profiling experiments, quality control analysis of the libraries, GFP screen and Crispr screen. M.H. performed electrophysiology recordings, U.P. developed the transcriptomic analysis pipeline and RiboBase analysis workflows. C.C. provided study oversight and acquired funding. All authors approved the final manuscript.

References:

1. Gao, X. *et al.* Quantitative profiling of initiating ribosomes in vivo. *Nat. Methods* **12**, 147–153 (2015).
2. 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).
3. Vieira de Souza, E. *et al.* Rp3: Ribosome profiling-assisted proteogenomics improves coverage and confidence during microprotein discovery. *Nat. Commun.* **15**, 6839 (2024).
4. Ruiz Cuevas, M. V. *et al.* Most non-canonical proteins uniquely populate the proteome or immunopeptidome. *Cell Rep.* **34**, 108815 (2021).
5. Yang, H., Li, Q., Stroup, E. K., Wang, S. & Ji, Z. Widespread stable noncanonical peptides identified by integrated analyses of ribosome profiling and ORF features. *Nat. Commun.* **15**, 1932 (2024).
6. Pauli, A. *et al.* Toddler: an embryonic signal that promotes cell movement via Apelin receptors. *Science* **343**, 1248636 (2014).
7. Na, Z. *et al.* The NBDY microprotein regulates cellular RNA decapping. *Biochemistry* **59**, 4131–4142 (2020).
8. Chen, Y. *et al.* Unannotated microprotein EMBOW regulates the interactome and chromatin and mitotic functions of WDR5. *Cell Rep.* **42**, 113145 (2023).
9. Cloutier, P. *et al.* Upstream ORF-encoded ASDURF is a novel prefoldin-like subunit of the PAQosome. *J. Proteome Res.* **19**, 18–27 (2020).
10. Rocha, A. L. *et al.* An inner mitochondrial membrane microprotein from the SLC35A4 upstream ORF regulates cellular metabolism. *J. Mol. Biol.* **436**, 168559 (2024).
11. Senís, E. *et al.* TUNAR lncRNA encodes a microprotein that regulates neural differentiation and neurite formation by modulating calcium dynamics. *Front. Cell Dev. Biol.* **9**, 747667

(2021).

12. Bencun, M. *et al.* A novel uORF regulates folliculin to promote cell growth and lysosomal biogenesis during cardiac stress. *Sci. Rep.* **15**, 3319 (2025).
13. Anderson, D. M. *et al.* Widespread control of calcium signaling by a family of SERCA-inhibiting micropeptides. *Sci. Signal.* **9**, ra119–ra119 (2016).
14. Duffy, E. E. *et al.* Developmental dynamics of RNA translation in the human brain. *Nat. Neurosci.* **25**, 1353–1365 (2022).
15. Lee, D. S. M. *et al.* Disrupting upstream translation in mRNAs is associated with human disease. *Nat. Commun.* **12**, 1515 (2021).
16. Martinez, T. F. *et al.* Profiling mouse brown and white adipocytes to identify metabolically relevant small ORFs and functional microproteins. *Cell Metab.* **35**, 166–183.e11 (2023).
17. Hofman, D. A. *et al.* Translation of non-canonical open reading frames as a cancer cell survival mechanism in childhood medulloblastoma. *Mol. Cell* **84**, 261–276.e18 (2024).
18. Chew, G.-L., Pauli, A. & Schier, A. F. Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. *Nat. Commun.* **7**, 11663 (2016).
19. Johnstone, T. G., Bazzini, A. A. & Giraldez, A. J. Upstream ORFs are prevalent translational repressors in vertebrates. *EMBO J.* **35**, 706–723 (2016).
20. McGeachy, A. M. & Ingolia, N. T. Starting too soon: upstream reading frames repress downstream translation. *EMBO J.* **35**, 699–700 (2016).
21. Chen, J. *et al.* Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
22. Zheng, C. *et al.* CRISPR-Cas9-based functional interrogation of unconventional translatome reveals human cancer dependency on cryptic non-canonical open reading frames. *Nat. Struct. Mol. Biol.* **30**, 1878–1892 (2023).
23. Hacisuleyman, E. *et al.* Neuronal activity rapidly reprograms dendritic translation via eIF4G2:uORF binding. *Nat. Neurosci.* **27**, 822–835 (2024).

24. Sandmann, C.-L. *et al.* Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. *Mol. Cell* **83**, 994–1011.e18 (2023).
25. May, G. E. & McManus, C. J. Multiplexed analysis of human uORF regulatory functions during the ISR using PoLib-seq. *Methods Mol. Biol.* **2428**, 41–62 (2022).
26. Goel, S. C. Transcription unit. *Nature* **245**, 397–397 (1973).
27. Mudge, J. M. *et al.* Standardized annotation of translated open reading frames. *Nat. Biotechnol.* **40**, 994–999 (2022).
28. Ozadam, H. *et al.* Single-cell quantification of ribosome occupancy in early mouse development. *Nature* **618**, 1057–1064 (2023).
29. VanInsberghe, M., van den Berg, J., Andersson-Rolf, A., Clevers, H. & van Oudenaarden, A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. *Nature* **597**, 561–565 (2021).
30. Glock, C. *et al.* The translatome of neuronal cell bodies, dendrites, and axons. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2113929118 (2021).
31. Hien, A., Molinaro, G., Liu, B., Huber, K. M. & Richter, J. D. Ribosome profiling in mouse hippocampus: plasticity-induced regulation and bidirectional control by TSC2 and FMRP. *Mol. Autism* **11**, 78 (2020).
32. Fernandez-Moya, S. M., Bauer, K. E. & Kiebler, M. A. Meet the players: local translation at the synapse. *Front Mol Neurosci* **7**, 84 (2014).
33. van Oostrum, M. & Schuman, E. M. Understanding the molecular diversity of synapses. *Nat Rev Neurosci* **26**, 65–81 (2025).
34. Sun, C. *et al.* The prevalence and specificity of local protein synthesis during neuronal synaptic plasticity. *Sci Adv* **7**, eabj0790 (2021).
35. Liu, B. *et al.* Optimization of ribosome profiling using low-input brain tissue from fragile X syndrome model mice. *Nucleic Acids Res.* **47**, e25 (2019).

36. Zhang, P. *et al.* Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.* **8**, 1749 (2017).

37. Ozadam, H., Geng, M. & Cenik, C. RiboFlow, RiboR and RiboPy: an ecosystem for analyzing ribosome profiling data at read length resolution. *Bioinformatics* **36**, 2929–2931 (2020).

38. Zhang, C., Wang, M., Li, Y. & Zhang, Y. Profiling and functional characterization of maternal mRNA translation during mouse maternal-to-zygotic transition. *Sci. Adv.* **8**, eabj3967 (2022).

39. Johnson, M. H. & Ziomek, C. A. The foundation of two distinct cell lineages within the mouse morula. *Cell* **24**, 71–80 (1981).

40. Sutherland, A. E., Speed, T. P. & Calarco, P. G. Inner cell allocation in the mouse morula: the role of oriented division during fourth cleavage. *Dev. Biol.* **137**, 13–25 (1990).

41. Huang, M. T. Harringtonine, an inhibitor of initiation of protein biosynthesis. *Mol. Pharmacol.* **11**, 511–519 (1975).

42. Fresno, M., Jiménez, A. & Vázquez, D. Inhibition of translation in eukaryotic systems by harringtonine. *Eur. J. Biochem.* **72**, 323–330 (1977).

43. Casola, C., Owoyemi, A. & Vakirlis, N. Degradation determinants are abundant in human noncanonical proteins. *bioRxiv* (2024) doi:10.1101/2024.05.01.592071.

44. Wu, Q. *et al.* Translation of small downstream ORFs enhances translation of canonical main open reading frames. *EMBO J.* **39**, e104763 (2020).

45. Wang, H. *et al.* Tissue- and stage-specific landscape of the mouse translatome. *Nucleic Acids Res.* **49**, 6165–6180 (2021).

46. Budamgunta, H. *et al.* Comprehensive peptide analysis of mouse brain striatum identifies novel sORF-encoded polypeptides. *Proteomics* **18**, e1700218 (2018).

47. Loughran, G., Sachs, M. S., Atkins, J. F. & Ivanov, I. P. Stringency of start codon selection modulates autoregulation of translation initiation factor eIF5. *Nucleic Acids Res.* **40**,

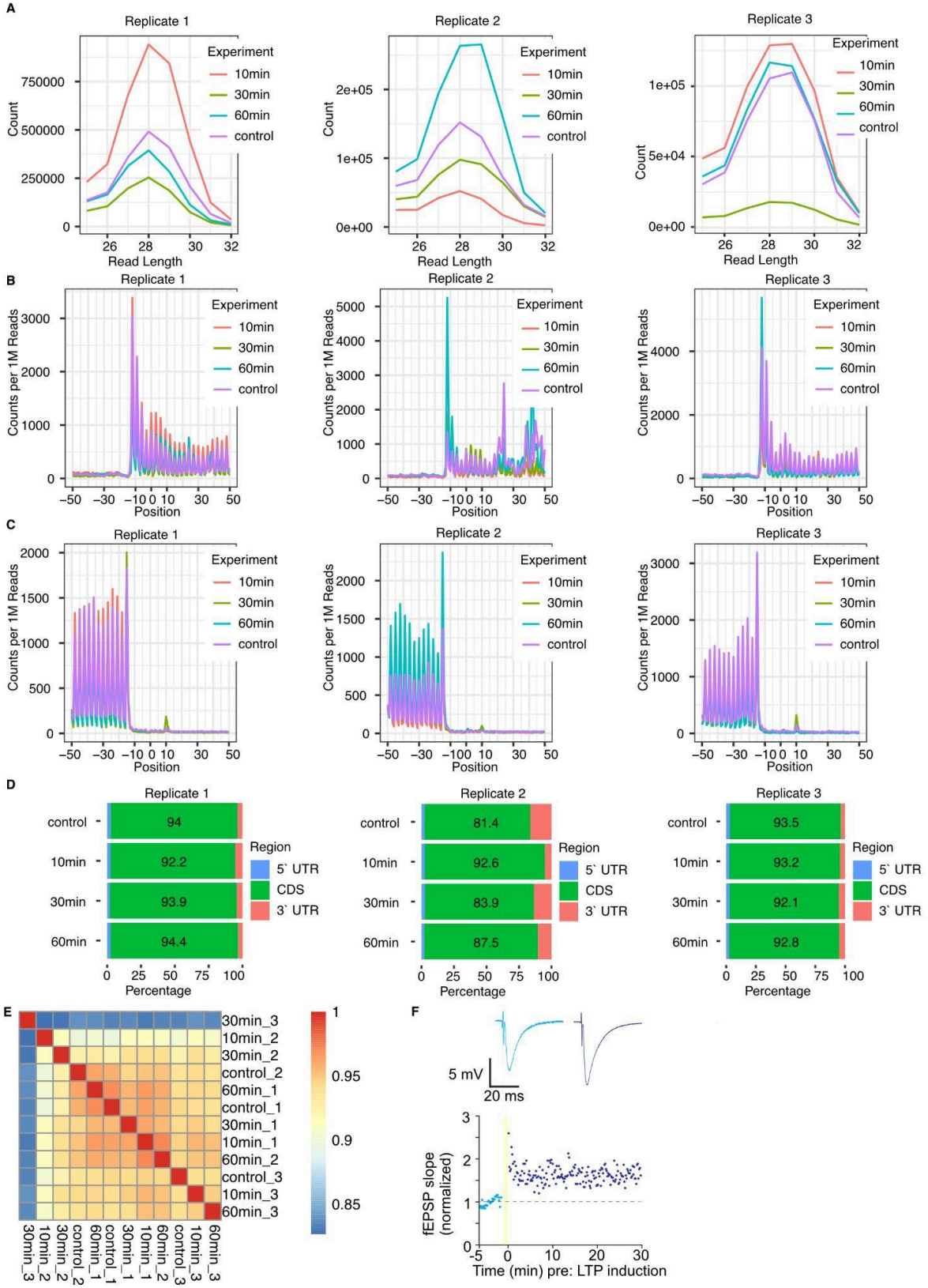
2898–2906 (2012).

48. Harnett, D. *et al.* A critical period of translational control during brain development at codon resolution. *Nat. Struct. Mol. Biol.* **29**, 1277–1290 (2022).
49. Chew, G.-L. *et al.* Ribosome profiling reveals resemblance between long non-coding RNAs and 5' leaders of coding RNAs. *Development* **140**, 2828–2834 (2013).
50. Shukla, Y., Ghatpande, V., Hu, C. F., Dickinson, D. J. & Cenik, C. Landscape and regulation of mRNA translation in the early C. elegans embryo. *Developmental Biology* (2024).
51. Liu, Y. *et al.* Translation efficiency covariation across cell types is a conserved organizing principle of mammalian transcriptomes. *bioRxivorg* (2024) doi:10.1101/2024.08.11.607360.
52. Chen, X. *et al.* Molecular basis for arginine C-terminal degron recognition by Cul2FEM1 E3 ligase. *Nat. Chem. Biol.* **17**, 254–262 (2021).
53. Arribere, J. A. *et al.* Translation readthrough mitigation. *Nature* **534**, 719–723 (2016).
54. Lang, W.-H., Calloni, G. & Vabulas, R. M. Polylysine is a Proteostasis Network-Engaging Structural Determinant. *J. Proteome Res.* **17**, 1967–1977 (2018).
55. Chng, S. C., Ho, L., Tian, J. & Reversade, B. ELABELA: A hormone essential for heart development signals via the apelin receptor. *Dev. Cell* **27**, 672–680 (2013).
56. Dib, A. *et al.* Pri smORF peptides are wide mediators of ecdysone signaling, contributing to shape spatiotemporal responses. *Front. Genet.* **12**, 714152 (2021).
57. Gallois, M. *et al.* Pri peptides temporally coordinate transcriptional programs during epidermal differentiation. *Sci. Adv.* **10**, eadg8816 (2024).
58. May, G. E. *et al.* Unraveling the influences of sequence and position on yeast uORF activity using massively parallel reporter systems and machine learning. *Elife* **12**, (2023).
59. Brar, G. A. *et al.* High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557 (2012).
60. Attili, S. M., Silva, M. F. M., Nguyen, T.-V. & Ascoli, G. A. Cell numbers, distribution, shape, and regional variation throughout the murine hippocampal formation from the adult brain

Allen Reference Atlas. *Brain Struct. Funct.* **224**, 2883–2897 (2019).

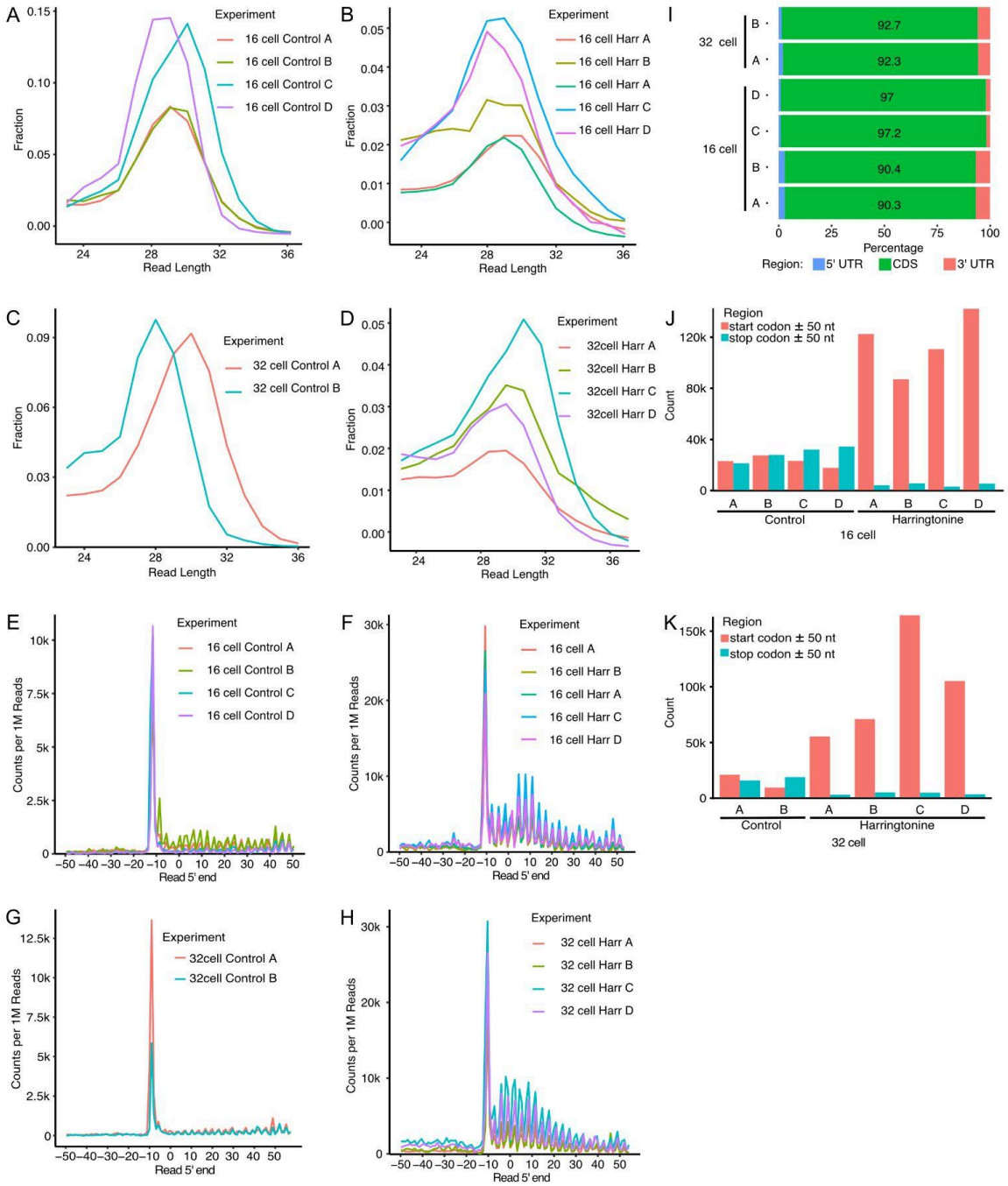
61. Rao, S., Tonn, T. & Cenik, C. Ribo-ITP: A method for ribosome profiling in single cells and low-input samples. *Methods Mol. Biol.* **2923**, 197–214 (2025).
62. Chacko, J., Ozadam, H. & Cenik, C. RiboGraph: an interactive visualization system for ribosome profiling data at read length resolution. *Bioinformatics* **40**, (2024).
63. Wright Muelas, M., Mughal, F., O'Hagan, S., Day, P. J. & Kell, D. B. The role and robustness of the Gini coefficient as an unbiased tool for the selection of Gini genes for normalising expression profiling data. *Sci. Rep.* **9**, 17960 (2019).
64. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).
65. Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
66. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10 (2011).
67. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

Ribo-ITP expands the translatome of limited input samples
Supplemental Figures:



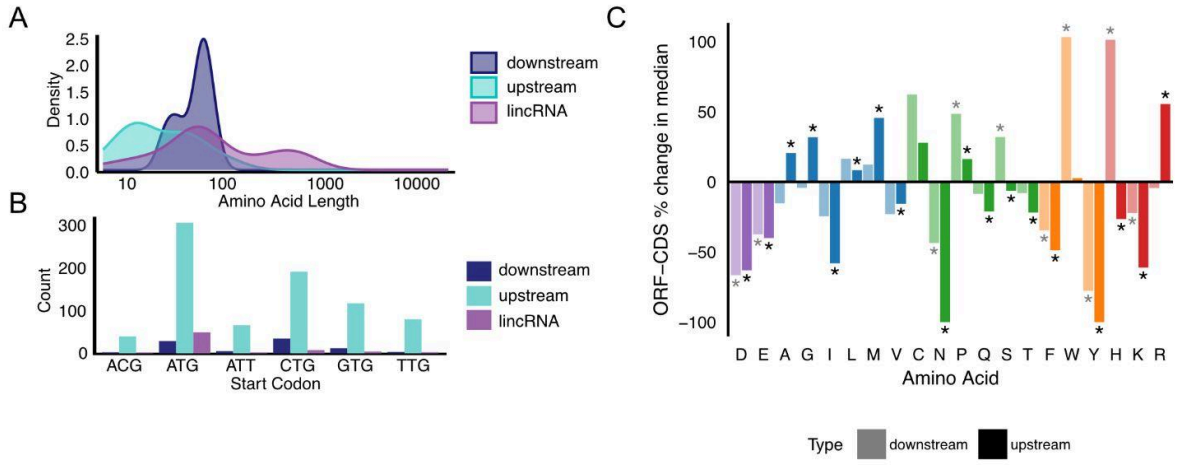
FigS1: QC plots for the individual hippocampal RiboITP replicates:

A, Read length distribution plots of the transcriptome mapping RPFs from individual biological replicates. The raw read counts for each read length are plotted. B, The Ribosome occupancy around the translation start site for individual replicates. Translation start site is denoted by the position 0. Aggregated read counts (y-axis) relative to the start site are plotted without P-site offset correction. C, The Ribosome occupancy around the translation stop site for individual replicates. Translation stop site is denoted by the position 0. Aggregated read counts (y-axis) relative to the stop site are plotted without P-site offset correction. D, Read coverage across different regions of the transcript. Percent mapping reads in the 5'UTR, CDS and 3'UTR are plotted. E, Replicate similarity between the four time points across the three biological replicates. Heatmap of the Spearman's correlation R values for all the hippocampal datasets. F, Voltage traces of LTP induction in hippocampal slices. Plots on the top show traces recorded before (light blue) and after (dark blue) the presentation of the LTP stimulus. The plot on the bottom shows the slope of the response as a function of time. The yellow shaded region is the time during which the LTP stimulus is being presented.



FigS2: QC plots for the individual embryo RiboITP replicates:

A-D, Read length distribution plots of the transcriptome mapping RPFs from individual embryos. Normalized read counts for each read length are plotted. E-H, Ribosome occupancy around the translation start site for individual embryos. Translation start site is denoted by the position 0. Aggregated read counts (y-axis) relative to the start site are plotted without P-site offset correction. I, Read distribution plot showing the percent reads mapping to the different regions in a transcript (5'UTR, CDS and 3'UTR). J-K, Distribution of raw reads mapping in the 50 nt region surrounding the translation start site and stop site for individual embryos.



FigS3: Properties of translated non-canonical ORFs identified by RiboTISH
*A, Density plots showing the amino acid length distribution of the peptides encoded by the translons identified in the hippocampal dataset identified by RiboTISH. B, The number of translons identified with cognate and near-cognate start codons in the hippocampal dataset identified by RiboTISH. C, Bar graph showing the fractional difference in the upstream (dark) or downstream (light) hippocampal translons identified by RiboTISH with their corresponding CDS for the individual amino acids. Colors indicate amino acid type. * indicates Wilcoxon's rank-sum test p values < 0.05.*