

Developmental dynamics of RNA translation in the human brain

Received: 21 October 2021

Accepted: 12 August 2022

Published online: 28 September 2022

 Check for updates

Erin E. Duffy ¹✉, Benjamin Finander ¹, GiHun Choi ¹, Ava C. Carter¹, Iva Pritisanac^{2,3,4,5}, Aqsa Alam², Victor Luria^{6,7,8}, Amir Karger ⁹, William Phu^{8,10,11}, Maxwell A. Sherman ¹², Elena G. Assad¹, Naomi Pajarillo¹, Alexandra Khitun¹³, Elizabeth E. Crouch^{14,15}, Sanika Ganesh¹, Jin Chen¹⁶, Bonnie Berger¹², Nenad Sestan ⁶, Anne O'Donnell-Luria ^{8,10,11}, Eric J. Huang ^{17,18,19}, Eric C. Griffith ¹, Julie D. Forman-Kay ^{3,4}, Alan M. Moses², Brian T. Kalish ^{1,20,21}✉ and Michael E. Greenberg ¹✉

The precise regulation of gene expression is fundamental to neurodevelopment, plasticity and cognitive function. Although several studies have profiled transcription in the developing human brain, there is a gap in understanding of accompanying translational regulation. In this study, we performed ribosome profiling on 73 human prenatal and adult cortex samples. We characterized the translational regulation of annotated open reading frames (ORFs) and identified thousands of previously unknown translation events, including small ORFs that give rise to human-specific and/or brain-specific microproteins, many of which we independently verified using proteomics. Ribosome profiling in stem-cell-derived human neuronal cultures corroborated these findings and revealed that several neuronal activity-induced non-coding RNAs encode previously undescribed microproteins. Physicochemical analysis of brain microproteins identified a class of proteins that contain arginine-glycine-glycine (RGG) repeats and, thus, may be regulators of RNA metabolism. This resource expands the known translational landscape of the human brain and illuminates previously unknown brain-specific protein products.

The human brain leverages extraordinary protein diversity to execute developmental programs, organize neural circuits and perform complex cognitive tasks¹. Proteomic diversity is generated through a series of transcriptional, post-transcriptional and translational mechanisms that ultimately contribute to a rich and complex 'translatome'. Although many studies have focused on transcriptional regulation in the developing human brain, much less is known regarding the complexity of translational regulation in this context, underscoring the need to study this key regulatory node in human brain development.

Deep sequencing of ribosome-protected mRNA fragments (ribosome profiling) provides a means to map genome-wide translation at nucleotide resolution². From these data, the movement of

ribosomes across codons can be determined and then used to identify protein-coding ORFs. Ribosome profiling in various cell types has revealed that the fraction of the transcriptome subject to translational regulation is far greater than previously recognized, with a single transcript often encoding many distinct protein products. Thus, RNA sequencing (RNA-seq) analysis fails to give a complete picture of the landscape of proteins produced by a cell. Indeed, ribosomal profiling studies in yeast³, as well as cardiac tissues⁴ and tumor tissues⁵, have revealed the widespread active translation of previously unknown small ORFs (sORFs) encoding microproteins ≤100 amino acids (aa). From the relatively few microproteins to be functionally studied, researchers have identified important regulators of mitochondrial metabolism,

A full list of affiliations appears at the end of the paper. ✉ e-mail: erin_duffylacy@hms.harvard.edu; brian.kalish@sickkids.ca; michael_greenberg@hms.harvard.edu

translational regulation and cell differentiation^{6–8}. To date, however, the nature and roles of similar microprotein species in the developing human brain remain almost entirely uncharacterized.

Here we describe the generation of a comprehensive translational atlas of the human prenatal and adult cortex, from 73 distinct tissue samples. In addition to cataloguing annotated gene programs that are subject to dynamic regulation at the translational level, we identify a vast array of novel sORFs and other non-canonical translation events, including many arising from previously annotated non-coding RNAs (ncRNAs). Similar findings were also obtained from ribosome profiling of stem-cell-derived human neuronal cultures, where we identified several novel microproteins translated from neuronal activity-responsive RNAs previously annotated as non-coding. We found that most sORFs in the brain are newly evolved in humans, where a subset of the sORFs arose via transposable element insertion at start codons. Although their recent evolution might be thought to suggest that the microproteins translated from these sORFs are non-functional, more than 100 of the human-specific microproteins identified in our study have been previously shown to play a key role in the viability of a non-neuronal cell type⁹. Our study, thus, markedly expands the known translational landscape of the developing brain and provides a rich resource for the study of novel human brain sORFs. This dataset is accessible via our accompanying web-based searchable database (<http://greenberg.hms.harvard.edu/project/human-brain-orf-database/>).

Results

Translational landscape of the human prenatal and adult brain

To characterize the human brain translational landscape at single-nucleotide resolution, we performed simultaneous RNA-seq and ribosome profiling (Ribo-seq) from human adult dorsolateral prefrontal cortex and prenatal cortex across a range of ages (Fig. 1a and Extended Data Fig. 1a). RNA-seq provides a quantitative measure of the mRNA species expressed in the brain, whereas Ribo-seq allows for a quantitative appraisal of active mRNA translation. The gestational age of prenatal cortex samples (30 total) ranged from 12 weeks to 23 weeks, whereas adult brain donors (43 total) ranged in age from 18 years to 82 years, with an average postmortem interval of 9.9 hours (Fig. 1b). Notably, across samples, Ribo-seq data exhibited the three-nucleotide periodicity characteristic of actively translating ribosomes, a key metric for confident ORF identification (Fig. 1c). Moreover, Ribo-seq reads exhibited expected fragment size distributions (Extended Data Fig. 1b) and mapped primarily to annotated gene-coding regions (Extended Data Fig. 1c), further supporting the idea that this method robustly captures RNA protected by actively translating ribosomes. Full demographic and Ribo-seq quality metrics are available in Supplementary Table 1 and Extended Data Fig. 1a–j.

High-confidence bona fide ORFs were identified based on the characteristic triplet reading frame periodicity of ribosome footprints using RibORF¹⁰. An ORF was considered high confidence if the sequences were present in two or more samples, exhibited clear start and stop codons and displayed Ribo-seq reads across the entire putative ORF region. After combining data across samples and filtering for ORF quality, we identified a total of 172,187 distinct actively translated ORFs in the human brain, mapping to 13,305 distinct genes (Fig. 1d,e and Extended Data Fig. 1d). In support of the quality of the resulting annotations, the relative proportions of each ORF type in our dataset, as well as general features such as start codon usage, were broadly consistent with previous findings in cell lines^{11,12} and other tissues¹³ (Fig. 1e,f and Extended Data Fig. 1e). Specifically, non-canonical ORFs of all types are detected in these datasets and tend to use non-AUG start codons with higher frequency compared to canonical ORFs. ORFs translated from ncRNAs were most commonly identified within previously annotated long intergenic non-coding RNAs (lincRNAs) or pseudogenes (Fig. 1g), including the recently characterized ncRNA-encoded microproteins NoBody⁸, MOXI⁷ and Cyren¹⁴. However, other highly expressed ncRNAs

that are not known to be translated, such as *XIST*, *HOTAIR* and *NEAT1*, showed no evidence of active translation in the brain, further corroborating the specificity of the identified ncRNA-associated ORFs. Taken together, these data map the translational landscape of the human cortex across development at an unprecedented level of resolution.

Transcription and translation during human brain development

Although transcriptional changes during the course of neurodevelopment have been extensively profiled^{15,16}, the contribution of translational regulation during neurodevelopment has not been analyzed in depth. Adopting previous methods that used the number of ribosomes per RNA molecule (ribosome density (RD)) as a measure of translational efficiency, we investigated the extent to which brain ORFs exhibit developmental shifts in translational efficiency¹⁷, focusing on canonical ORFs that encode proteins of known function. Comparison of our paired transcriptome and translome datasets revealed several distinct modes of developmental regulation (Fig. 2a,b and Extended Data Fig. 2a–j): buffered (change in RD that counterbalances the change in RNA level); intensified (change in RD that amplifies the change in RNA level); mRNA transcription/stability (change in the number of RNA molecules without a change in RD); or exclusively translationally regulated (a change in RD but no change in mRNA level). We found, for example, that developmental decreases in ribosomal gene RNA levels were effectively buffered by corresponding increases in translational efficiency (Fig. 2c and Extended Data Fig. 2d,e). This coordinated translational regulation likely reflects developmental changes in mTOR signaling, as these transcripts contain 5′ terminal oligopyrimidine tract (5′-TOP) motifs¹⁸, sequences at the 5′ ends of mRNAs that link their translation to the mTOR nutrient-sensing signaling pathway¹⁹. In contrast, ORFs encoding both major histocompatibility complex (MHC) components and proteins involved in complement activation display increases in both mRNA level and translational efficiency between the prenatal and adult brain (Fig. 2c and Extended Data Fig. 2d,e). Given the respective roles of these factors in developmental synapse formation and elimination^{20,21}, these findings implicate active translational regulation in the control of developmental synaptic pruning and circuit assembly.

sORFs and non-canonical translation in the human brain

Studies in other systems have shown that translational regulation is more widespread across the transcriptome than previously appreciated, often involving regions of the genome annotated as non-coding (for example, pseudogenes, lincRNAs, antisense RNAs and 5′ and 3′ untranslated regions (UTRs) of canonical protein-coding genes). To interrogate our datasets for novel human brain microproteins, we focused on Ribo-seq-identified ORFs ≤300 nucleotides (nt) (100 aa) in length that were either out of frame or did not overlap with longer ORFs. This analysis identified 38,187 actively translated sORFs originating from 8,278 genes in the prenatal or adult brain (Fig. 3a and Extended Data Fig. 3a,b). Although many of these sORFs were translated from alternative regions of canonical protein-coding transcripts, 1,705 were derived from annotated non-coding transcripts, including reported lincRNAs, pseudogenes and antisense transcripts (Fig. 3b and Extended Data Fig. 3a). Notably, although the ribosome density of sORFs was, on average, ~10-fold lower than the translation of canonical ORFs (Extended Data Fig. 3a), this low level of translation is similar to that of several previously reported microprotein-encoding RNAs with well-characterized functions, including RPL41 (ref. 22), SLN²³ and NBDY⁸, suggesting that newly described sORFs with relatively low translation compared to canonical ORFs are likely, at least in some cases, to encode functional microproteins. Like canonical ORFs, many of these sORFs were developmentally regulated via coordinated changes in RNA abundance and/or translational efficiency, which may enable the fine-tuning of sORF protein levels as the brain matures (Fig. 3c).

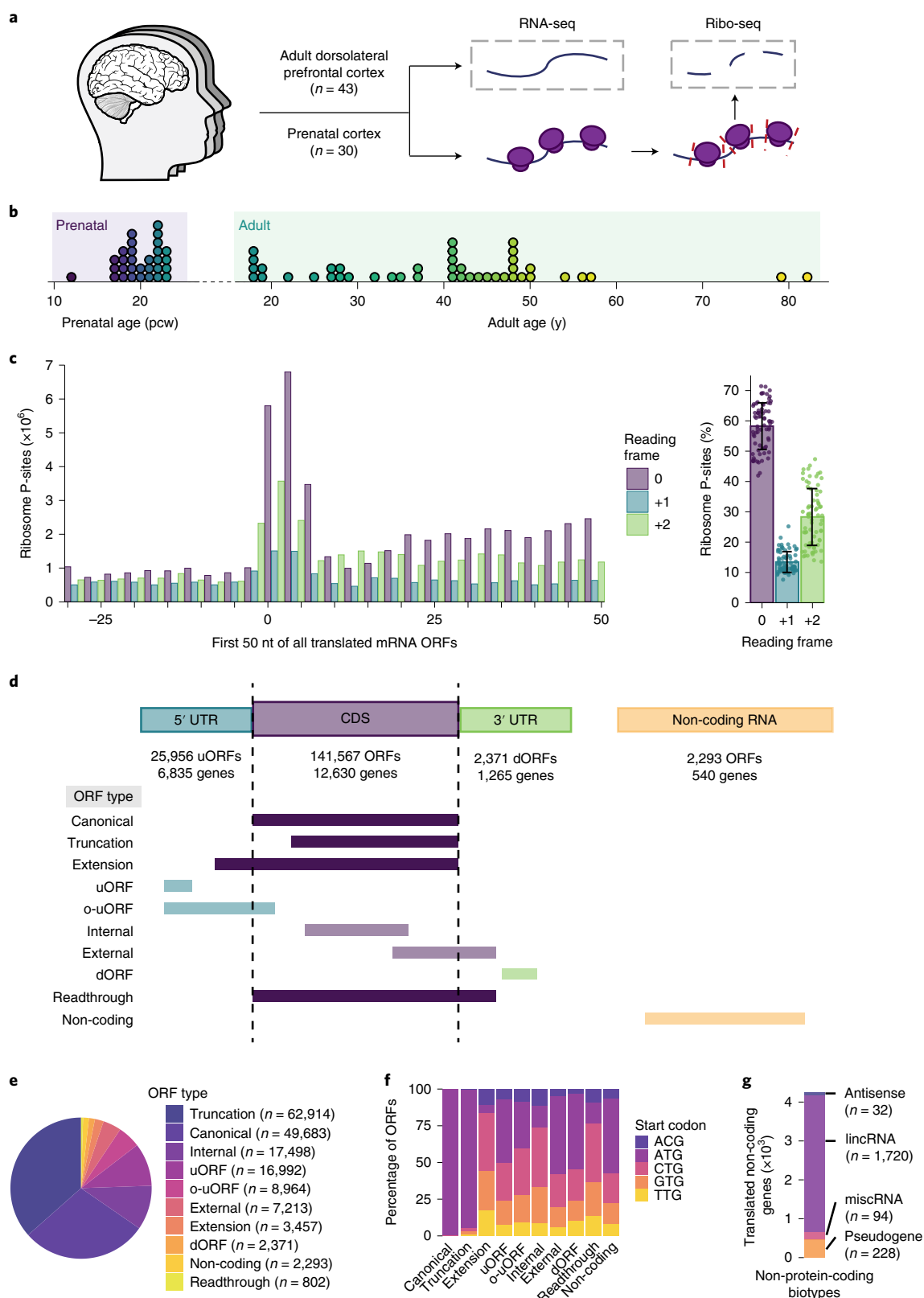


Fig. 1 | Ribosome profiling captures active translation in the human adult and prenatal brain. **a**, Overview of experimental design. **b**, Histogram depiction of patient samples included in this study. **c**, Bar plot displaying P-sites derived from offset-corrected Ribo-seq reads in the first 100 nt of annotated ORFs (left) and the percentage of footprints in each reading frame (right). Data are shown as mean \pm s.d.; $n = 73$ biologically independent tissues. **d**, Schematic overview

of ORF types detected by RibORF. CDS, coding sequence. **e**, Number of ORFs of each type identified in human adult and/or prenatal brain. **f**, Stacked bar plot of start codon usage by ORF type. **g**, Stacked bar plot of numbers and percentages of translated ncRNAs separated by transcript biotype. dORF, downstream open reading frame; miscRNA, miscellaneous RNA; o-uORF, overlapping upstream open reading frame; pcw, post-conception weeks; y, years.

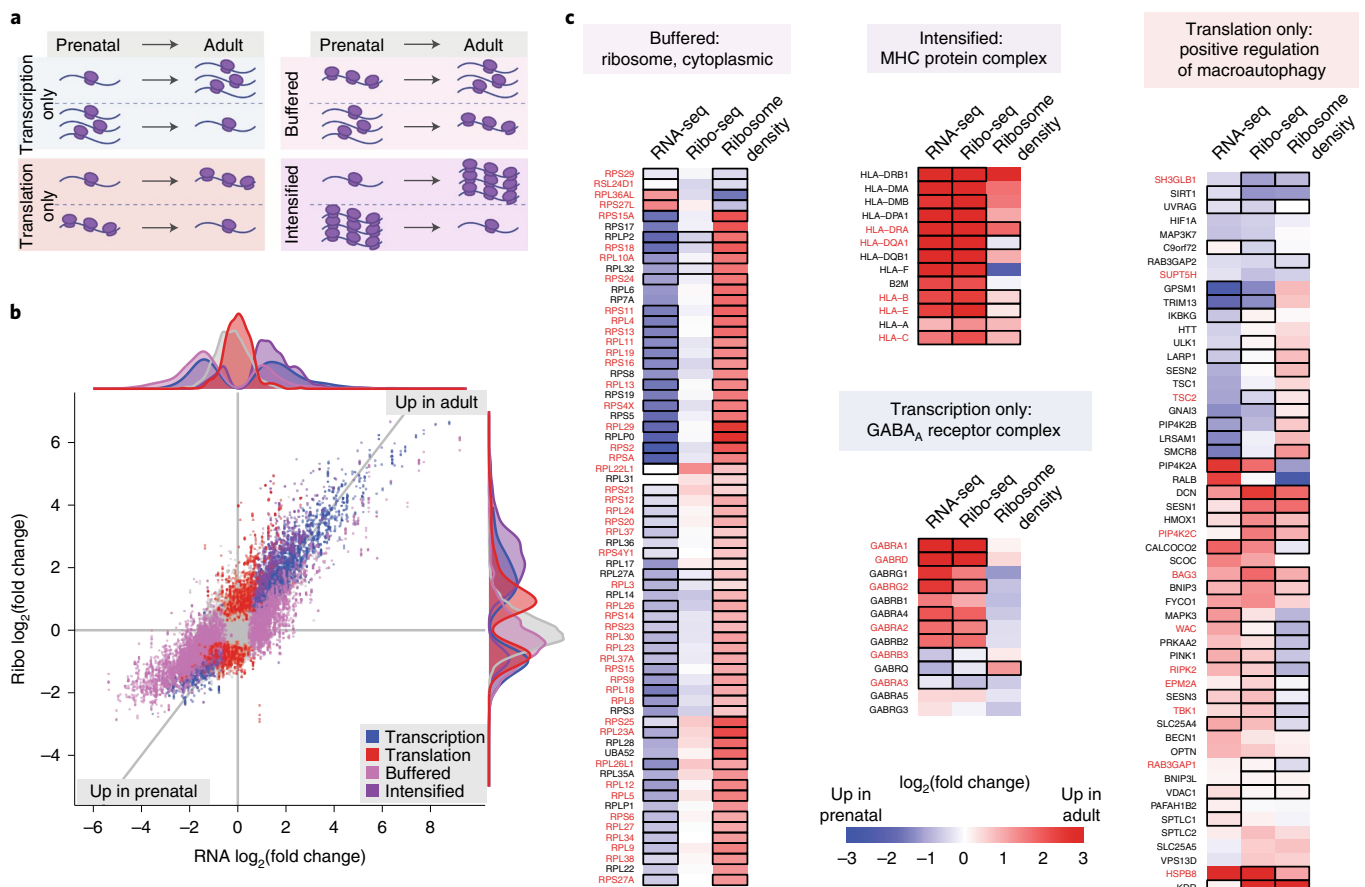


Fig. 2 | Transcriptional and translational regulation across human brain development. **a**, Classification of genes based on RNA-seq, Ribo-seq and RD measurements. **b**, Scatter plot of fold changes between adult and prenatal brain for all canonical ORFs in Ribo-seq data and the corresponding gene in RNA-seq data. Positive values indicate enrichment in the adult brain, whereas negative values indicate enrichment in the prenatal brain. Transcriptionally regulated genes (blue; change in transcription with no change in RD); translationally

regulated genes (red; change in RD with no change in transcription); buffered genes (light purple; change in RD that counterbalances the change in mRNA transcription); and intensified genes (dark purple; change in RD that amplifies the change in mRNA) are highlighted. **c**, Heat map of genes associated with the top GO term in each regulatory category identified in **a**. Black outlines indicate DESeq2 $P_{\text{adj}} < 0.05$; gene names in red indicate inclusion in a given regulatory category.

While recognizing the difficulties associated with proteomic microprotein detection, we sought to independently corroborate our Ribo-seq findings at the protein level. Toward this end, we performed size-selected mass spectrometry-based proteomics for enhanced detection of protein species less than 20 kDa²⁴. To facilitate the identification of proteins not annotated by UniProt, this analysis incorporated a proteogenomic approach, whereby all peptides detected by mass spectrometry were matched to a custom database constructed from our Ribo-seq data. To further increase our ability to detect rare microproteins, we also re-analyzed published mass spectrometry data from 50 human adult brain tissue samples to search for signatures of sORF-derived microproteins²⁵.

Collectively, these analyses identified peptides corresponding to 4,104 unique ORFs (Fig. 3d and Extended Data Fig. 3d–g), including 199 sORFs, 39 uORFs and four non-coding ORFs. To highlight one such example, this analysis confirmed the presence of a novel microprotein encoded by an upstream ORF (uORF) in *GLUD1* (glutamate dehydrogenase 1), a gene critically involved in glutamate metabolism (Fig. 3e)^{26,27}. Notably, this *GLUD1* uORF contains a translation initiator of short 5' UTR (TISU) motif, which is known to enable uninterrupted translation under conditions of energy stress²⁸, suggesting that this microprotein might contribute to neuronal responses to acute metabolic demands. A full list of proteomically detected sORF species is available in Supplementary Table 2. Given the sensitivity limitations of unbiased

spectrometry, it is not surprising that the overall proportion of sORFs validated by proteomics in this context is relatively low. However, these proteogenomic datasets serve to validate the presence of a sizeable number of non-canonical ORF-derived microproteins expressed in the brain at levels similar to functionally validated microproteins identified in other tissues. In total, the identified microproteins represent a substantial expansion of the known brain translome, with potential relevance for human development and disease.

Regulated sORF translation in human neurons

To complement these tissue-based studies, we also characterized the translational landscape in human embryonic stem cell (hESC)-derived neuronal cultures. To this end, we employed an engineered hESC line harboring an integrated doxycycline-inducible NGN2 construct. We adapted a previously described protocol²⁹ in which doxycycline induction of NGN2 was combined with SMAD and WNT inhibition to induce patterning toward a forebrain phenotype (Fig. 4a). The resulting cultures (hereafter, NGN2 neurons) have transcriptional signatures that are similar to those of well-differentiated glutamatergic neurons (Extended Data Fig. 4a).

For these studies, we also exposed the differentiated NGN2 neurons to elevated levels of potassium chloride (KCl), a treatment that is known to induce acute, synchronous membrane depolarization and to promote activity-dependent changes in RNA transcription in these

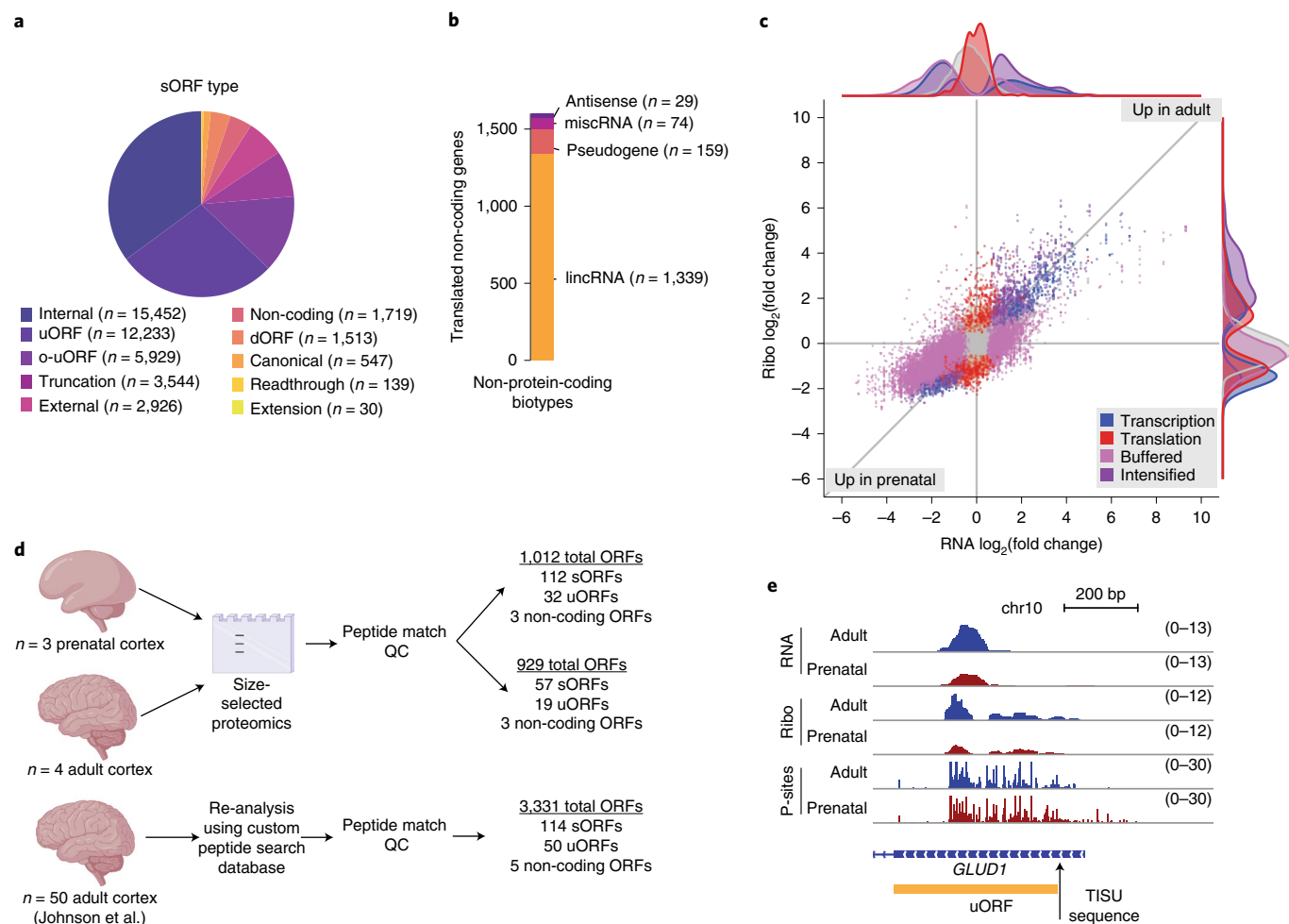


Fig. 3 | Microprotein expression and validation across brain development. a, Number of sORFs of each type identified in human adult and/or prenatal brain. **b**, Stacked bar plot of numbers and percentages of translated ncRNAs containing at least one sORF, separated by transcript biotype. **c**, Scatter plot of fold changes between adult and prenatal brain for all sORFs in Ribo-seq data and the corresponding gene in RNA-seq data. Positive values indicate enrichment in the adult brain, whereas negative values indicate enrichment in the prenatal brain. Genes regulated by transcription (blue), translation (red), buffered (light purple) and intensified mechanisms (dark purple) are highlighted. **d**, Number and type of ORFs identified by size-selection proteomics in the adult and prenatal brain or by Johnson et al.²⁵. **e**, Genomic locus of *GLUD1*. Tracks represent merged and depth-normalized reads across all adult versus prenatal samples for RNA-seq, Ribo-seq as well as P-site positions. The sORF identified by RibORF is shown in gold, and the TISU sequence is indicated with an arrow. dORF, downstream open reading frame; miscRNA, miscellaneous RNA; o-uORF, overlapping upstream open reading frame; QC, quality control.

neurons³⁰. Thus, day 28 cultures from three independent differentiation cohorts were harvested for combined RNA-seq and ribosome profiling with or without 6 hours of membrane depolarization with 55 mM KCl. The resulting datasets passed key quality control metrics, with clear three-nucleotide periodicity observed in the Ribo-seq data (Fig. 4h) and high data correlation between separate differentiation cohorts (Extended Data Fig. 4b,c). Moreover, robust induction of known activity-responsive loci was observed in all depolarized samples (Extended Data Fig. 4d). Collectively, this analysis identified a total of 124,613 actively translated ORFs in NGN2 neurons (Fig. 4b,c), more than 60% of which (78,965/124,613) were also observed in the human brain tissue samples (Extended Data Fig. 4e,f). Principal component analysis (PCA) plots showed that NGN2 samples cluster more with the fetal rather than the adult transcriptome (Fig. 4d), as might be expected for hESC-derived neurons.

In addition to being useful for the study of human neuronal activity-dependent changes, these cultured human neurons are amenable to other manipulations such as translational inhibition with harringtonine, a small-molecule drug that immobilizes ribosomes immediately after translation initiation and results in ribosome

footprint accumulation at initiation sites¹¹. The use of harringtonine is a key validation of Ribo-seq experiments not available in postmortem tissue. Therefore, we performed Ribo-seq on membrane-depolarized NGN2 neurons treated with harringtonine or a vehicle control. We used Ribo-TISH to predict ORFs from harringtonine data, as this approach has recently been shown to be superior in identifying non-canonical and lowly expressed ORFs³¹. Compared to vehicle-treated control neurons, we observed the expected accrual of ribosome footprints at translational initiation sites in harringtonine-treated samples (Fig. 4i and Extended Data Fig. 4h). We validated the start codons of 61,400 ORFs, including 8,881 sORFs, 5,258 uORFs and 2,005 ORFs translated from ncRNAs (Fig. 4j, Extended Data Fig. 4i and Supplementary Table 3). Canonical ORFs showed the highest concordance (74.4%) between the harringtonine experiments and the original dataset analyzed using RibORF. Notably, we also confirmed the start codons of non-canonical ORFs using harringtonine treatment, including 43.1% of uORFs and 41.3% of non-coding ORFs. These ORFs represent the ‘highest confidence’ ORFs from the NGN2 dataset.

We next characterized the novel sORFs found to be translated from previously annotated ncRNAs. In this regard, we observed active

translation of novel sORFs in 128 of 706 activity-dependent ncRNAs detected in NGN2 cultures (Fig. 4e and Extended Data Fig. 4j–m), many of which (101) were also detected by Ribo-seq in our human brain datasets. Notably, several of the predicted protein products from these ncRNAs were verified biochemically using size-selected proteomics (Supplementary Table 2). Among these translated ncRNAs was *LINCO0473* (Fig. 4f, g), a previously characterized primate-specific and activity-dependent lincRNA³² that has been implicated as a sex-specific driver of stress resilience when expressed ectopically in the mouse prefrontal cortex³³. Thus, *LINCO0473* and many other previously annotated neuronal activity-dependent non-coding transcripts are translated to produce microproteins that may modulate key neuronal responses to activity. Together, these studies complement our analysis of human postmortem brain tissue and highlight the utility of NGN2 cultured neurons as a relatively homogeneous human neuronal population that is amenable to genetic and chemical manipulations.

Evolutionary conservation of sORFs

As a first step for prioritizing human brain sORFs for future study, we analyzed the evolutionary origins of brain sORFs using genomic phylostratigraphy—an approach that dates the origin of individual genes by examining the presence or absence of homologs across species³⁴. Determination of the minimal evolutionary age for human brain sORFs revealed that, compared to most annotated protein-coding genes, most sORFs are human-specific (67% of sORFs versus 12% of annotated protein-coding genes (Fig. 5a and Extended Data Fig. 5a)), consistent with the low levels of sORF sequence conservation observed in other tissues^{4,5,35}. Our analysis further revealed that more recently evolved sORFs are shorter, contain fewer splice junctions and exhibit lower ribosome density compared to their more evolutionarily ancient counterparts (Fig. 5b–d). Microproteins encoded by the more evolutionarily ancient sORFs are also more likely to be detectable by proteomics, perhaps as a consequence of their higher overall levels of expression (Extended Data Fig. 5b). These features are consistent with the classic view that the more evolutionarily conserved regions of the genome are more likely to be translated⁵, thus nominating the highly conserved sORFs as promising candidates for future functional studies. However, the rapid evolution of human-specific sORFs also suggests that these sequences may represent evolutionary experiments. These regions may gain translation capacity in a given species that is not necessarily conserved during further evolution. To begin to test whether a subset of these newly evolved sORFs is functional, we overlapped sORFs that show evidence of translation in the human brain with a recently published dataset of CRISPR–Cas9 perturbation of sORFs in K562 or human induced pluripotent stem cells (iPSCs), selecting for those sORFs that show a significant growth phenotype when knocked out (Mann–Whitney *U*-test $P < 0.05$)⁹. Of the 124 sORFs that satisfy these criteria, a striking 101 are

human-specific (chi-squared test $P = 0.0004752$), lending support to the hypothesis that these newly evolved microproteins have acquired important functions. It is also notable that, relative to all sORFs, those derived from brain-enriched transcripts are significantly more likely to be specific to humans (Kolmogorov–Smirnov test, $P < 2.2 \times 10^{-16}$; Extended Data Fig. 5c), consistent with the idea that these protein products contribute to human-specific aspects of brain development.

Recently, Playfoot and colleagues provided evidence for transposable element (TE) involvement in new ORF formation³⁶. We directly explored this as a possible mechanism of sORF generation in the brain, finding that, compared to canonical protein-coding ORFs, ncRNA-associated sORFs have a significantly increased overlap with TE insertions (10% versus 4%, respectively, $P < 2.2 \times 10^{-16}$ by two-proportion z-test; Fig. 5e and Extended Data Fig. 5d,e). This TE enrichment within ncRNAs has been previously noted and suggested to contribute new non-coding sequences for RNA-mediated ncRNA function^{37–39}. Our findings, however, provide evidence that TEs might also play an important role in the generation of new protein-coding ORFs within these annotated non-coding regions. Notably, different classes of TEs were also found to be associated with distinct ORF types (Fig. 5f); however, the functional importance of this observation requires further investigation.

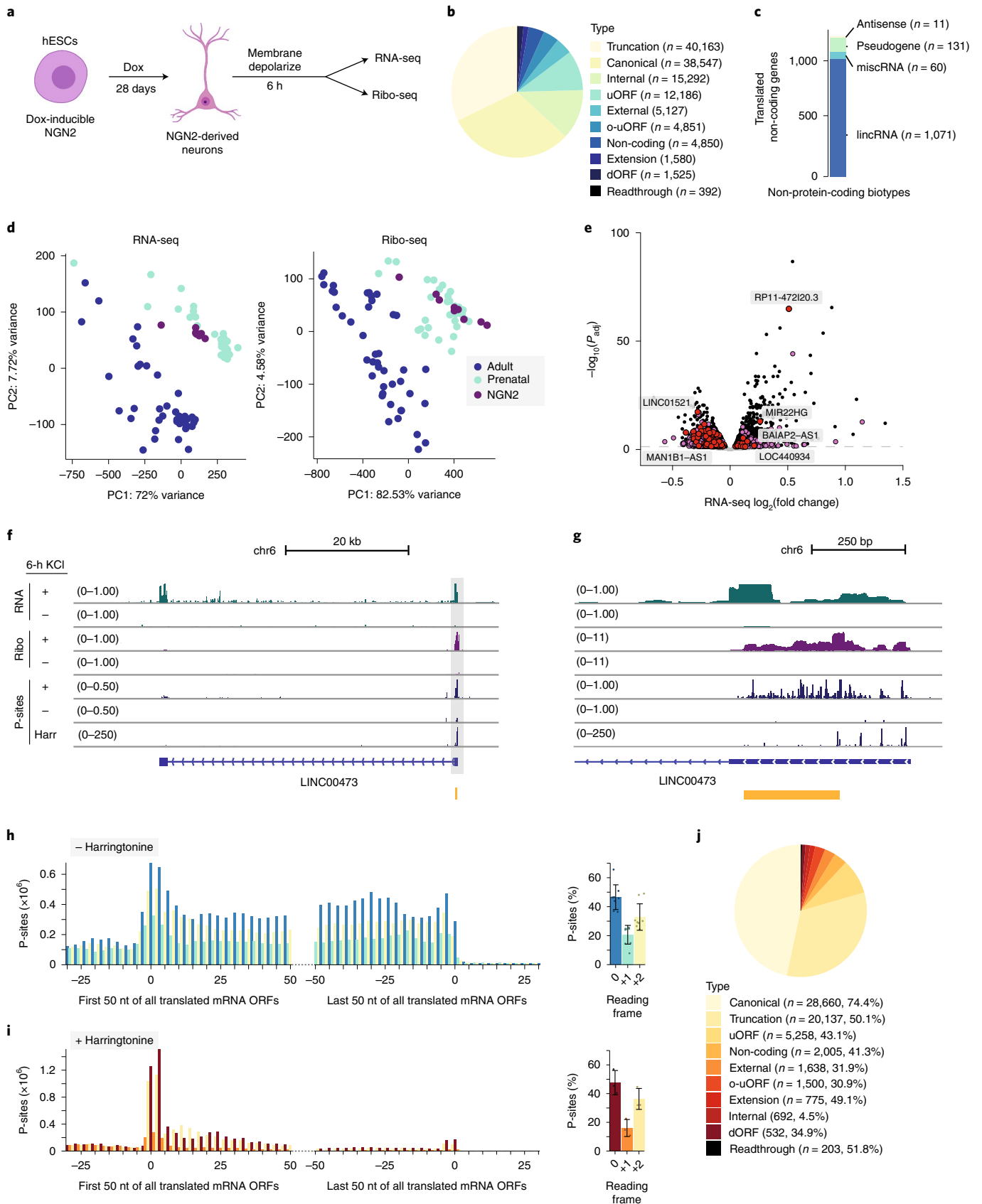
uORF regulation of canonical protein translation

Of the actively translated sORFs identified from the human brain, 8,239 (22%) were translated from brain-enriched or brain-specific transcripts⁴⁰, suggesting that, in many cases, their functions may be unique to the brain. To more directly investigate sORF function, we first focused on uORFs, a category of ORFs commonly thought to negatively regulate downstream translation of canonical ORFs through a variety of mechanisms^{41–43}. Somewhat surprisingly, but consistent with more recent findings^{4,42,43,44}, we found that uORF translation was not generally anti-correlated with translation of the corresponding canonical ORF (Fig. 6a–c and Extended Data Fig. 6a–c). Notwithstanding this general finding, we still identified several individual uORFs that were strongly anti-correlated with translation of their canonical downstream ORFs. One such example involved a uORF in *DLGAP1*, which encodes an important brain-enriched post-synaptic scaffolding protein⁴⁵ (Fig. 6d–f). In this case, translation of the *DLGAP1* uORF was strongly enriched in prenatal samples through the preferential use of an alternative transcriptional start site (TSS) (Fig. 6d,e and Extended Data Fig. 6d) and was associated with a reduction in translation of the canonical *DLGAP1* protein (Fig. 3f). Together, these data point toward a mechanism in which the use of an alternative TSS in the prenatal, but not the mature, brain leads to specific translational repression of the canonical *DLGAP1* protein. Notably, *DLGAP1* is a known autism-associated gene⁴⁶, raising the possibility that this

Fig. 4 | Activity-dependent translation in hESC-derived neurons. a,

Schematic of Ribo-seq and RNA-seq from NGN2 neurons after 6-hour membrane depolarization. **b**, Breakdown of translated ORFs of each type identified in NGN2 neurons. **c**, Stacked bar plot of numbers and percentages of translated ncRNAs separated by transcript biotype. **d**, PCA analysis based on RNA-seq and Ribo-seq reads mapping to annotated genes in primary adult and prenatal brain tissue and NGN2 neurons. **e**, Volcano plot of $-\log_{10}(P_{\text{adj}})$ versus $\log_2(\text{fold change})$ in RNA-seq expression between membrane-depolarized and unstimulated NGN2 neurons. Black indicates $DEseq2 P_{\text{adj}} < 0.05$; purple indicates activity-dependent ncRNAs with no evidence of translation in human brain or NGN2 neurons; and red indicates activity-dependent ncRNAs with evidence of translation in human brain and/or NGN2 neurons. **f**, Genomic locus of *LINCO0473* in NGN2 neurons. Tracks represent merged and depth-normalized reads across three biological replicates of membrane-depolarized (6-hour KCl) and unstimulated neurons for RNA-seq, Ribo-seq as well as P-site positions for Ribo-seq and harringtonine-treated Ribo-seq. The sORF identified by RibORF is shown in gold. **g**, High-resolution depiction of genomic locus of the ORF encoded by *LINCO0473* in NGN2 neurons.

Tracks represent merged and depth-normalized reads across three biological replicates of membrane-depolarized (6-hour KCl) and unstimulated neurons for RNA-seq, Ribo-seq as well as P-site positions for Ribo-seq and harringtonine-treated Ribo-seq. The sORF identified by RibORF is shown in gold. **h**, Bar plot displaying P-sites derived from offset-corrected Ribo-seq reads from NGN2 neurons treated with vehicle control. The first 50 nt (left) and the last 50 nt (right) of annotated ORFs are shown. Data are shown as mean \pm s.d.; $n = 6$ independent cell differentiations examined over two independent experiments. **i**, Bar plot displaying P-sites derived from offset-corrected Ribo-seq reads from NGN2 neurons treated with harringtonine. The first 50 nt (left) and the last 50 nt (right) of annotated ORFs are shown. Data are shown as mean \pm s.d.; $n = 3$ independent cell differentiations. **j**, Number of ORFs of each type identified in NGN2 neurons treated with harringtonine. Absolute number (n) and percentage of overlap with ORFs identified from NGN2 neurons treated with cycloheximide alone are noted in parentheses. dORF, downstream open reading frame; h, hours; miscRNA, miscellaneous RNA; o-uORF, overlapping upstream open reading frame; PC, principal component.



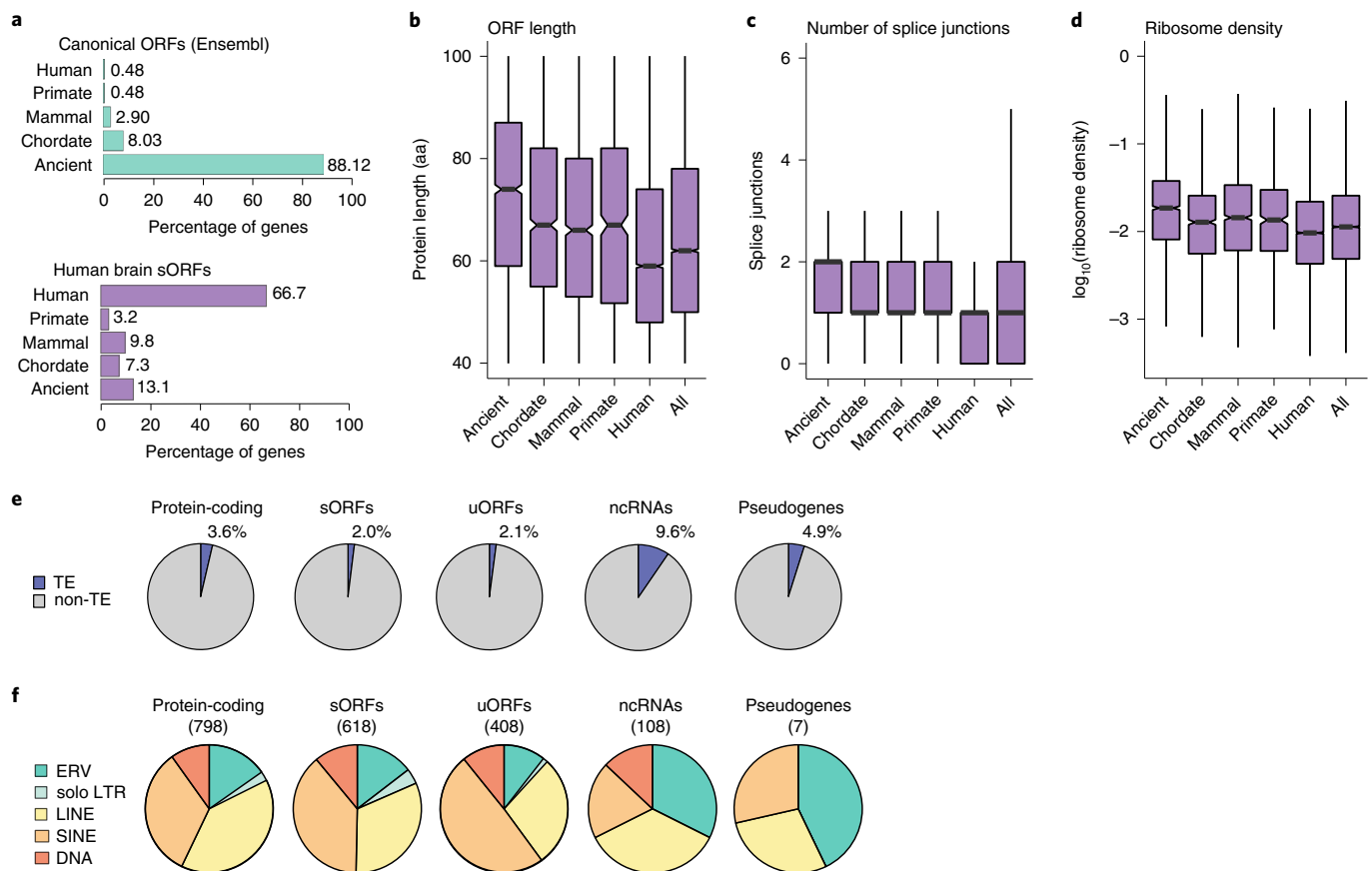


Fig. 5 | Evolutionary origins of human brain sORFs. a, Percentage of canonical ORFs (top, all ORFs in human Ensembl database, ≥ 40 aa) and sORFs (bottom, ≥ 40 aa) grouped by evolutionary age. **b**, Box and whisker plots of microprotein ORF length grouped by evolutionary age. **c**, Box and whisker plots of the number of splice junctions per microprotein ORF (40–100 aa) grouped by evolutionary age. **d**, Box and whisker plots of microprotein ORF RD grouped by evolutionary age. **b–d**, Data are shown as median \pm IQR (whiskers indicate $1.5 \times$ IQR); notches

indicate median $\pm 1.58 \times$ IQR/ \sqrt{n} ; $n = 2,488$ (ancient), 1,396 (chordate), 1,859 (mammal), 604 (primate), 12,689 (human), 19,036 (all) sORFs. **e**, Pie chart of the percentage of ORFs with a TE insertion at the start codon, grouped by ORF type or ncRNA biotype. **f**, Pie chart of the distribution of TE types, grouped by ORF type or ncRNA biotype. Numbers indicate the number of ORFs in each category. IQR, interquartile range; ncRNA, non-coding RNA.

developmentally timed regulation of *DLGAP1* translation may be required for proper neurological function. Our overall findings are, thus, consistent with a nuanced role for brain uORFs in translational regulation, with select uORFs exerting a strong negative regulatory influence on developmentally timed protein expression.

Physicochemical analysis of brain microprotein function

We also sought to gain further insight into brain microprotein function through additional primary sequence analysis. In this regard, 6,491 (17%) human brain sORFs showed significant sequence similarity ($E < 10^{-4}$) to known proteins, with 441 (~1%) matching a protein sequence encoded elsewhere in the genome. These previously characterized protein paralogs participate in a variety of processes, including cellular metabolism, transcription, translation and membrane transport (Extended Data Fig. 7a and Supplementary Table 4), raising the possibility that the newly identified sORFs encode microproteins with biological functions that are similar to the proteins translated from the corresponding canonical ORF. Indeed, we found that 31% of the sORFs with significant sequence similarity to known or predicted human proteins overlapped with an annotated protein domain, strongly suggesting that many of these sORFs encode defined folded structures or even entire structural domains (Extended Data Fig. 7b).

For sORFs lacking sequence similarity to known or predicted human proteins (69%), calculated FoldIndex scores, a rough predictive measure of intrinsic disorder⁴⁷, suggest that these protein products

do not generally adopt stable three-dimensional conformations (Fig. 7a), consistent with other sequence-based characteristics (Fig. 7b and Extended Data Fig. 7c–f). However, many disordered proteins have recently been shown to serve essential cellular functions through a variety of mechanisms, including the tuning of protein interaction specificity and affinity, as well as through the formation of biomolecular condensates^{48–50}.

To explore further the potential function of human brain microproteins, we proceeded to compare human brain sORFs with similarly sized disordered regions from known proteins on the basis of their physicochemical and bulk sequence properties using hierarchical clustering⁵¹ (Fig. 7c, Extended Data Fig. 7g and Supplementary Table 5). Strikingly, this analysis identified a strong enrichment of brain sORFs ($>5 \times$ expected, 712 total) in the resulting sequence clusters that were rich in RGG motifs and/or aromatic residues as well as, to a lesser extent, clusters rich in arginine residues ($2.8 \times$ expected). Likewise, human brain sORFs were strongly overrepresented ($>5 \times$ expected) in an aromatic amino acid-rich polypeptide cluster. By contrast, clusters of acidic, lysine-rich and polar sequences encompassing known intrinsically disordered proteins were strongly depleted for sORFs, indicating that brain sORFs display restricted sequence features, consistent with possible biological functions. Indeed, we found that 11 of the human brain microproteins that are predicted to be intrinsically disordered are important for cell survival⁹, further supporting a functional role for these microproteins in the brain.

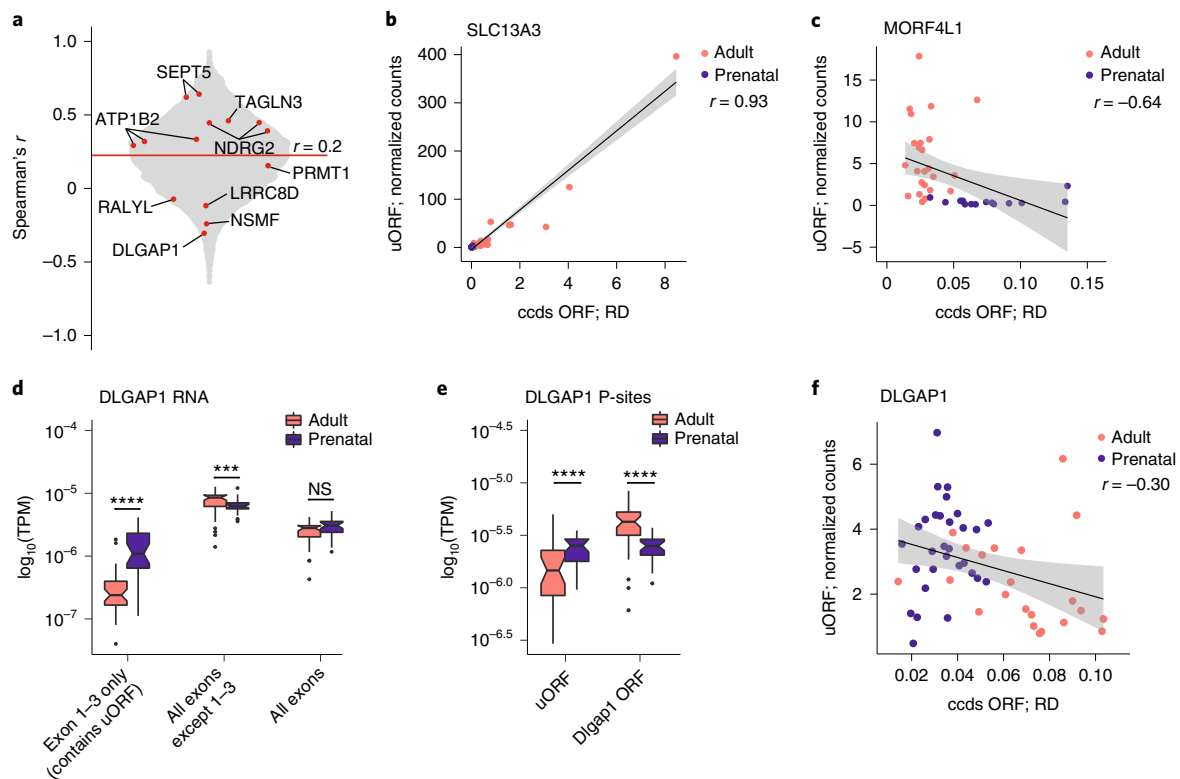


Fig. 6 | Effects of uORF expression on downstream ORF translation. a, Beeswarm dot plot showing the Spearman's r correlation between uORF translation (uORF normalized counts from ribosome profiling) and canonical ORF RD (canonical coding sequence [ccds]-normalized counts from ribosome profiling divided by normalized total RNA abundance) for individual genes across all 73 individuals. The red line represents the mean correlation across all genes. Red dots indicate developmentally regulated uORFs (described in Extended Data Fig. 6a). **b,c,** Scatter plot and Spearman's r correlation between upstream ORF translation (uORF normalized counts from ribosome profiling) and canonical ORF RD (ccds-normalized counts from ribosome profiling divided by normalized total RNA abundance) for *SLC13A3* (**b**) and *MORF4L1* (**c**) across 73 individuals. Gray shading indicates 95% confidence interval. **d,** Box and whisker plot of RNA-seq reads (TPM) from adult and prenatal samples

over *DLGAP1* exons 1–3 ($P = 7.47 \times 10^{-10}$), all exons except 1–3 ($P = 4.20 \times 10^{-4}$) and all exons (not significant). **e,** Box and whisker plot of Ribo-seq P-sites (in TPM) from adult and prenatal samples over *DLGAP1* uORF ($P = 6.15 \times 10^{-5}$) and ccds ORF ($P = 2.87 \times 10^{-8}$). **d,e,** **** $P < 0.0001$ and *** $P < 0.001$ by two-sided Kolmogorov–Smirnov test. Data are shown as median \pm IQR (whiskers indicate $1.5 \times$ IQR); notches indicate median $\pm 1.58 \times$ IQR/ \sqrt{n} ; $n = 43$ (adult) and 30 (prenatal) biologically independent tissues. **f,** Scatter plot and Spearman's r correlation between uORF translation (uORF normalized counts from ribosome profiling) and canonical ORF RD (ccds-normalized counts from ribosome profiling divided by normalized total RNA abundance) for *DLGAP1* across 73 individuals. Gray shading indicates 95% confidence interval. IQR, interquartile range; NS, not significant.

It is notable that RG and RGG motifs are important for the regulation of mRNA splicing and translation and have also been associated with RNA binding and the formation of biomolecular condensates⁵². Indeed, several previously characterized proteins in RGG-rich and R-rich clusters are known to interact with RNA in biomolecular condensates and have been implicated in splicing and mRNA binding, raising the possibility that the newly identified sORF-encoded microproteins may also interact with RNA processing complexes to control mRNA splicing, translation or DNA damage responses in the nucleus.

In vitro validation of candidate human brain microproteins

To independently verify the translation potential and subcellular localization of the newly identified sORFs, we overexpressed six selected sORFs with their endogenous 5' UTR and a FLAG-HA epitope tag in heterologous cells (Supplementary Table 6). These sORFs were all derived from annotated non-coding regions and included both evolutionarily ancient and human-specific sORFs. We confirmed expression of microproteins of expected molecular weight and found that subsequent start codon mutations prevented translation of these microproteins (Fig. 7d), providing further evidence that these sORFs can be efficiently translated to yield stable protein species. Moreover, the translated microproteins exhibited a range of subcellular localization patterns (Fig. 7e), suggesting that expression in heterologous cells provides a

useful platform for interrogating the biochemistry and cell biology of these newly identified human brain microproteins.

Discussion

RNA translation is a fundamental cellular process that is tightly regulated across human development. The fidelity of translation, as well as the stability and localization of RNA transcripts, are critical determinants of brain function, with mRNA translation regulation being a key step that can be mis-regulated in human neurodevelopment and neuropsychiatric disease^{53–56}. Importantly, studies in other human tissues such as the heart suggest that the translome is far more complex than previously appreciated^{4,5} and that the resulting proteome diversity likely contributes to a myriad of functions in these tissues. Remarkably, however, the human brain translome has remained largely uncharacterized.

In this study, we applied ribosome profiling and proteomics to the prenatal and adult human cortex, as well as to hESC-derived neuronal cultures, providing the first large-scale resource of translation events in the developing human brain and demonstrating that translation is an important mode of regulation for shaping the brain proteome. Collectively, our data reveal widespread translation of non-canonical ORFs in the human brain, including thousands of novel microproteins. We identified in the brain a subset of ncRNAs, uORFs and other

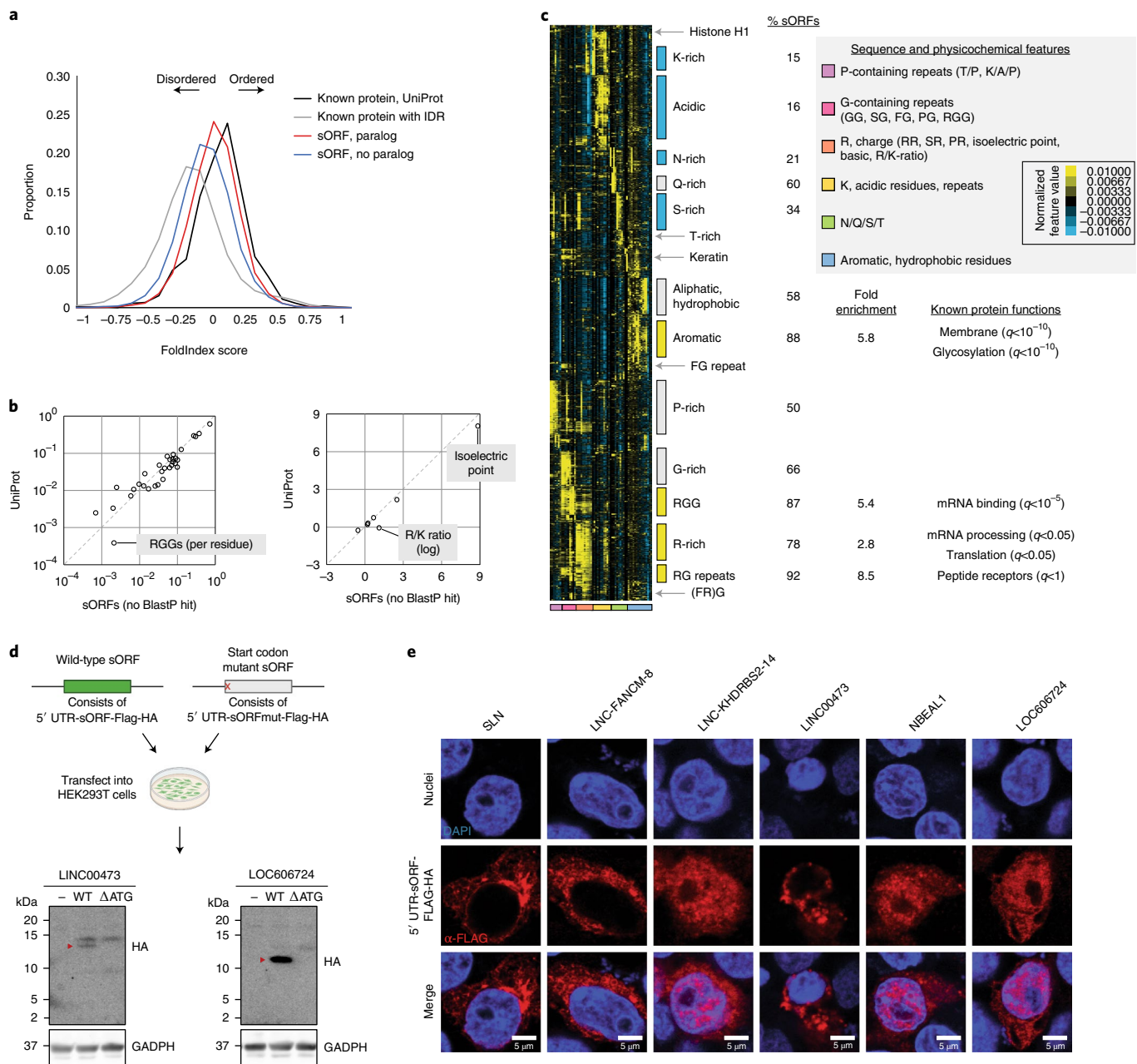


Fig. 7 | Microprotein functional characterization. **a**, FoldIndex score distribution of proteins annotated in UniProt (black), annotated proteins with IDRs (gray) and sORFs with and without a BlastP hit (red and blue, respectively). **b**, Scatter plot of average enrichment per residue of sequence and physicochemical properties in sORFs with no BlastP homology versus annotated proteins (UniProt). RGG repeats were the most highly enriched of the tested sequence and physicochemical properties in sORFs. **c**, Heat map and hierarchical clustering of z-scores for physicochemical parameters associated with the known disordered proteome (IDRs 21–100 aa in length) as well as sORFs with predicted IDRs that do not have a paralog and do not overlap annotated coding ORFs. For the purposes of this analysis, sORFs that overlapped with an annotated

ccds ORF (for example, internal, external and readthrough) were excluded from this analysis. Boxes to the right of the heat map indicate clusters of IDRs with similar properties. Blue indicates clusters depleted for sORFs; yellow indicates clusters significantly enriched for sORFs. **d**, Western blot of FLAG-HA-tagged unmodified and ATG-mutated *LINC00473* and *LOC606724* lincRNAs, which includes the endogenous 5' UTR of each transcript. Experiment was repeated twice with similar results. Unprocessed blots are provided in the source data. **e**, Immunofluorescence of FLAG-HA-tagged sORFs (*SLN*, *LNC-FANCM-8*, *LNC-KHDRBS2-14*, *LINC00473*, *NBEAL1* and *LOC606724*) containing the endogenous 5' UTR expressed in HEK293T cells. Experiment was repeated twice with similar results. WT, wild-type.

annotated non-coding transcripts that encode translated proteins, some of which were directly confirmed by mass spectrometry. We identified translational control as a widespread mode of canonical gene regulation across development while also acknowledging that effects of non-translational variables, such as protein stability and post-translational modifications, can influence resulting protein levels. Furthermore, the developmentally regulated changes in RD that we

identified could be due to changes in translational initiation or elongation, and future experiments will be required to disentangle these key modes of regulation. Further investigation of these pervasive forms of translational regulation promises new insights into the gene expression mechanisms that control various aspects of human neurodevelopment.

In addition to studying the developmental regulation of RNA translation in human brain tissue, we profiled the activity-dependent

translatome in hESC-derived neuronal cultures and found that many activity-dependent lincRNAs that were thought to be non-coding are actually translated in this context. What is unclear for individual ‘non-coding’ transcripts is whether they function in the brain solely as protein-coding RNAs or whether the RNA and the encoded protein possess independent functions. Recent studies using expression quantitative trait locus analysis suggest that hundreds of lincRNAs have associations with human diseases and that rare variants in lincRNAs impact complex human traits⁵⁷. These findings underscore the importance of discerning the protein-coding potential of brain-expressed lincRNAs in future studies.

The identification of sORFs is a necessary and critical first step in understanding their role in the human brain, and much work remains to understand the function of individual human brain sORFs. Our analysis of microprotein amino acid sequence identified a marked absence of structured domains and likely enrichment for proteins with RNA-binding functions. Given these findings, it is tempting to speculate that these disordered microproteins may impact RNA metabolism by enhancing or inhibiting the formation of biomolecular condensates. Moreover, the fact that most sORFs are human-specific renders them interesting candidates in the study of uniquely human features of the brain. Although these findings may suggest that some newly evolved microproteins are non-functional, we isolated many human-specific microproteins that appear to play a key role in cell growth and viability⁹. It will be of great interest in the future to understand how sORFs may expand and become fixed in the genome through continued evolution.

It is important to consider several caveats of the current study. First, although we were able to use existing single-cell RNA-seq data from human adult brain tissue to estimate the distribution of cell types in our adult and prenatal brain tissue samples, our ribosome profiling was restricted to bulk tissue measurements. In addition, ribosome profiling was largely performed from postmortem brain tissue, and postmortem interval-dependent decreases in translation initiation and/or ribosome/RNA binding likely contributed to some loss of ORF resolution that may result in an overestimation of truncated ORF annotations. However, due to our stringent filtering, our findings in the present study likely represent an overall underestimate of translated ORFs. Although we carefully curated our final list of ORFs, all ORF identification is limited by the sliding scale of confidence in computational ORF-calling algorithms, and only a small fraction of predicted ORFs were subsequently validated through biochemistry. Similarly, alternative splicing events can masquerade as alternate translation events, which could lead to false-positive identification of novel ORFs. Despite these limitations, our finding that the translatome of hESC-derived neurons largely mimics translation in prenatal cortex tissue suggests that our measurements in postmortem tissue predominantly reflect physiologically relevant translation events.

In conclusion, our study provides, to our knowledge, the first large-scale resource for the investigation of translation regulation in the human brain. Notably, our results identify previously unannotated microproteins as candidates for future functional characterization, opening new opportunities for the investigation of translational regulation in the nervous system and for the elucidation of the function of many new human-specific and brain-specific microproteins.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41593-022-01164-9>.

References

- Kim, M.-S. et al. A draft map of the human proteome. *Nature* **509**, 575–581 (2014).
- 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**, 218–223 (2009).
- Guydosh, N. R. & Green, R. Dom34 rescues ribosomes in 3' untranslated regions. *Cell* **156**, 950–962 (2014).
- van Heesch, S. et al. The translational landscape of the human heart. *Cell* **178**, 242–260.e29 (2019).
- Prensner, J. R. et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. *Nat. Biotechnol.* **39**, 697–704 (2021).
- Makarewich, C. A. et al. The DWORF micropeptide enhances contractility and prevents heart failure in a mouse model of dilated cardiomyopathy. *eLife* **7**, e38319 (2018).
- Makarewich, C. A. et al. MOXI is a mitochondrial micropeptide that enhances fatty acid β -oxidation. *Cell Rep.* **23**, 3701–3709 (2018).
- D’Lima, N. G. et al. A human microprotein that interacts with the mRNA decapping complex. *Nat. Chem. Biol.* **13**, 174–180 (2017).
- Chen, J. et al. Pervasive functional translation of noncanonical human open reading frames. *Science* **367**, 1140–1146 (2020).
- Ji, Z. RibORF: identifying genome-wide translated open reading frames using ribosome profiling. *Curr. Protoc. Mol. Biol.* **124**, e67 (2018).
- 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).
- Ji, Z., Song, R., Regev, A. & Struhl, K. Many lincRNAs, 5' UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).
- Kalish, B. T. et al. Maternal immune activation in mice disrupts proteostasis in the fetal brain. *Nat. Neurosci.* **24**, 204–213 (2021).
- Slavoff, S. A., Heo, J., Budnik, B. A., Hanakahi, L. A. & Saghatelian, A. A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. *J. Biol. Chem.* **289**, 10950–10957 (2014).
- Miller, J. A. et al. Transcriptional landscape of the prenatal human brain. *Nature* **508**, 199–206 (2014).
- Kang, H. J. et al. Spatio-temporal transcriptome of the human brain. *Nature* **478**, 483–489 (2011).
- Chothani, S. et al. deltaTE: detection of translationally regulated genes by integrative analysis of Ribo-seq and RNA-seq data. *Curr. Protoc. Mol. Biol.* **129**, e108 (2019).
- Levy, S., Avni, D., Hariharan, N., Perry, R. P. & Meyuhas, O. Oligopyrimidine tract at the 5' end of mammalian ribosomal protein mRNAs is required for their translational control. *Proc. Natl Acad. Sci. USA* **88**, 3319–3323 (1991).
- Jeong, H. et al. Evolution of DNA methylation in the human brain. *Nat. Commun.* **12**, 2021 (2021).
- Huh, G. S. et al. Functional requirement for class I MHC in CNS development and plasticity. *Science* **290**, 2155–2159 (2000).
- Stevens, B. et al. The classical complement cascade mediates CNS synapse elimination. *Cell* **131**, 1164–1178 (2007).
- Klaudiny, J., von der Kammer, H. & Scheit, K. H. Characterization by cDNA cloning of the mRNA of a highly basic human protein homologous to the yeast ribosomal protein YL41. *Biochem. Biophys. Res. Commun.* **187**, 901–906 (1992).
- Odermatt, A. et al. Characterization of the gene encoding human sarcolipin (SLN), a proteolipid associated with SERCA1: absence of structural mutations in five patients with Brody disease. *Genomics* **45**, 541–553 (1997).
- Khitun, A. & Slavoff, S. A. Proteomic detection and validation of translated small open reading frames. *Curr. Protoc. Chem. Biol.* **11**, e77 (2019).

25. Johnson, E. C. B. et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. *Nat. Med* **26**, 769–780 (2020).
26. Frigerio, F. et al. Deletion of glutamate dehydrogenase 1 (*Glud1*) in the central nervous system affects glutamate handling without altering synaptic transmission. *J. Neurochem* **123**, 342–348 (2012).
27. Lander, S. S. et al. Glutamate dehydrogenase-deficient mice display schizophrenia-like behavioral abnormalities and CA1-specific hippocampal dysfunction. *Schizophr. Bull.* **45**, 127–137 (2019).
28. Sinvani, H. et al. Translational tolerance of mitochondrial genes to metabolic energy stress involves TISU and eIF1-eIF4G1 cooperation in start codon selection. *Cell Metab.* **21**, 479–492 (2015).
29. Nehme, R. et al. Combining NGN2 programming with developmental patterning generates human excitatory neurons with NMDAR-mediated synaptic transmission. *Cell Rep.* **23**, 2509–2523 (2018).
30. Sanchez-Priego, C. et al. Mapping cis-regulatory elements in human neurons links psychiatric disease heritability and activity-regulated transcriptional programs. *Cell Rep.* **39**, 110877 (2022).
31. Zhang, P. et al. Genome-wide identification and differential analysis of translational initiation. *Nat. Commun.* **8**, 1749 (2017).
32. Ataman, B. et al. Evolution of Osteocrin as an activity-regulated factor in the primate brain. *Nature* **539**, 242–247 (2016).
33. Issler, O. et al. Sex-specific role for the long non-coding RNA LINC00473 in depression. *Neuron* **106**, 912–926 (2020).
34. Domazet-Loso, T., Brajković, J. & Tautz, D. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends Genet.* **23**, 533–539 (2007).
35. Calviello, L. et al. Detecting actively translated open reading frames in ribosome profiling data. *Nat. Methods* **13**, 165–170 (2016).
36. Playfoot, C. J. et al. Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. *Genome Res.* **31**, 1531–1545 (2021).
37. Carlevaro-Fita, J. et al. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome Res.* **29**, 208–222 (2019).
38. Kapusta, A. et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.* **9**, e1003470 (2013).
39. Johnson, R. & Guigó, R. The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *RNA* **20**, 959–976 (2014).
40. Uhlén, M. et al. Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
41. Brar, G. A. et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. *Science* **335**, 552–557 (2012).
42. 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).
43. Tresenrider, A. et al. Integrated genomic analysis reveals key features of long undecoded transcript isoform-based gene repression. *Mol. Cell* **81**, 2231–2245.e11 (2021).
44. Aspden, J. L. et al. Extensive translation of small open reading frames revealed by Poly-Ribo-Seq. *eLife* **3**, e03528 (2014).
45. Rasmussen, A. H., Rasmussen, H. B. & Silaharoglu, A. The DLGAP family: neuronal expression, function and role in brain disorders. *Mol. Brain* **10**, 43 (2017).
46. Xing, J. et al. Resequencing and association analysis of six PSD-95-related genes as possible susceptibility genes for schizophrenia and autism spectrum disorders. *Sci. Rep.* **6**, 27491 (2016).
47. Prilusky, J. et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435–3438 (2005).
48. Tsang, B. et al. Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation. *Proc. Natl Acad. Sci. USA* **116**, 4218–4227 (2019).
49. Gueroussov, S. et al. Regulatory expansion in mammals of multivalent hnRNP assemblies that globally control alternative splicing. *Cell* **170**, 324–339 (2017).
50. Hnisz, D., Shrinivas, K., Young, R. A., Chakraborty, A. K. & Sharp, P. A. A phase separation model for transcriptional control. *Cell* **169**, 13–23 (2017).
51. Zarin, T. et al. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. *eLife* **8**, e46883 (2019).
52. Chong, P. A., Vernon, R. M. & Forman-Kay, J. D. RGG/RG motif regions in RNA binding and phase separation. *J. Mol. Biol.* **430**, 4650–4665 (2018).
53. Jishi, A., Qi, X. & Miranda, H. C. Implications of mRNA translation dysregulation for neurological disorders. *Semin. Cell Dev. Biol.* **114**, 11–19 (2021).
54. Chen, Y.-C., Chang, Y.-W. & Huang, Y.-S. Dysregulated translation in neurodevelopmental disorders: an overview of autism-risk genes involved in translation. *Dev. Neurobiol.* **79**, 60–74 (2019).
55. Kelleher, R. J. & Bear, M. F. The autistic neuron: troubled translation? *Cell* **135**, 401–406 (2008).
56. Kapur, M., Monaghan, C. E. & Ackerman, S. L. Regulation of mRNA translation in neurons—a matter of life and death. *Neuron* **96**, 616–637 (2017).
57. de Goede, O. M. et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. *Cell* **184**, 2633–2648.e19 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

¹Department of Neurobiology, Harvard Medical School, Boston, MA, USA. ²Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada. ³Program in Molecular Medicine, The Hospital for Sick Children, Toronto, ON, Canada. ⁴Department of Biochemistry, University of Toronto, Toronto, ON, Canada. ⁵Gottfried Schatz Research Center for Cell Signaling, Metabolism and Aging, Molecular Biology and Biochemistry, Medical University of Graz, Graz, Austria. ⁶Department of Neuroscience, Yale School of Medicine, New Haven, CT, USA. ⁷Department of Systems Biology, Harvard Medical School, Boston, MA, USA. ⁸Department of Pediatrics, Division of Genetics and Genomics, Boston Children's Hospital, Boston, MA, USA. ⁹IT-Research Computing, Harvard Medical School, Boston, MA, USA. ¹⁰Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. ¹¹Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. ¹²Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. ¹³Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. ¹⁴Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA. ¹⁵The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California, San Francisco, San Francisco, CA, USA. ¹⁶Cecil H. and Ida Green Center for Reproductive Biology Sciences, UT Southwestern Medical Center, Dallas, TX, USA. ¹⁷Department of Pathology, University of California, San Francisco, San Francisco, CA, USA. ¹⁸Weill Institute for Neurosciences, University of California, San Francisco, San Francisco, CA, USA. ¹⁹Pathology Service 113B, San Francisco Veterans Affairs Healthcare System, San Francisco, CA, USA. ²⁰Department of Paediatrics, Division of Neonatology, Hospital for Sick Children, Toronto, ON, Canada. ²¹Program in Neuroscience and Mental Health, SickKids Research Institute, Toronto, ON, Canada. ✉e-mail: erin_duffylacy@hms.harvard.edu; brian.kalish@sickkids.ca; michael_greenberg@hms.harvard.edu

Methods

Human brain samples

All human tissue research was approved by the Harvard Medical School institutional review board (IRB). De-identified adult brain tissue samples were obtained from the National Institute of Health (NIH) NeuroBioBank. NIH NeuroBioBank sample collection has been approved by the following IRBs: University of Miami IRB, University of Maryland IRB, Maryland Department of Health and Mental Hygiene IRB, Partners Human Research Committee, Department of Veterans Affairs—Los Angeles, Bronx VA Medical Center IRB and University of Pittsburgh IRB. Adult brain samples with a postmortem interval of <15 hours were included in the final cohort. Prenatal brain samples were obtained from the Human Developmental Biology Resource (HDBR) and the University of California, San Francisco (UCSF) Pediatric Neuropathology Laboratory. De-identified tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations. All cases were determined by chromosomal analysis, physical examination and/or pathological analysis to be control tissues, which indicates that they were absent of neurological disease. Cases with any abnormalities in these parameters were not used for this study. Patients were consented specifically for research purposes and were not compensated for their tissue donation. Tissue collection protocols were approved by the Human Gamete, Embryo and Stem Cell Research Committee (IRB) at UCSF and the Newcastle & North Tyneside 1 Research Ethics Committee (for HDBR).

Ribosome profiling of human brain tissue

Ribosome profiling was performed using a protocol modified from McGlincy et al.⁵⁸. Frozen brain tissue (~80 mg) was thawed on ice. Each sample was dounced 15× in 400 µl of ice-cold lysis buffer: 20 mM Tris pH 7.4, 150 mM NaCl, 5 mM MgCl₂, 1 mM DTT and 100 µg ml⁻¹ of cycloheximide (Sigma-Aldrich). The lysate was further sheared using a 26-gauge syringe. The lysate was clarified by centrifugation at 20,000g for 10 minutes at 4 °C. Supernatant (10 µl) was removed, added to 300 µl of TRIzol and frozen at -80 °C for future RNA-seq library preparation. RNA concentration in the remaining supernatant was quantified using RNA Qubit. Lysate (30 µg) was subjected to RNase I digestion (0.5 U of RNase I per µg of RNA) at room temperature for 45 minutes with gentle agitation.

After RNase digestion, 10 µl of Superasin (Thermo Fisher Scientific) was added to each sample, and the samples were transferred to ice. To isolate ribosome protected fragments, the RNase-digested lysate was transferred to Ultra-Clear 11 × 34 mm centrifuge tubes (Beckman Coulter) and underlaid with 0.9 ml of sucrose cushion. Samples were centrifuged in a TLS-55 rotor at 51,000 r.p.m. for 2 hours at 4 °C. The supernatant was discarded, and the pellet was resuspended in 300 µl of TRIzol. Ribosome-protected fragments were purified from TRIzol using the Zymo Direct-zol kit. RNA was precipitated by adding 38.5 µl of RNase-free water, 1.5 µl of GlycoBlue, 10 µl of 3 M sodium acetate pH 5.5 and 150 µl of isopropanol to 100 µl of eluted RNA. The mixture was incubated overnight at -20 °C. Samples were centrifuged for 30 minutes at 20,000g at 4 °C. The supernatant was discarded, and the RNA pellet was resuspended in 5 µl of 10 mM Tris pH 8. Next, 5 µl of 2× denaturing sample loading buffer (980 µl of formamide, 20 µl of 500 mM EDTA and 300 µg of bromophenol blue) was added to each sample, and then the sample was denatured at 80 °C for 90 seconds. Ribosome-protected fragments, along with control oligos⁵⁸, were run on a 15% polyacrylamide gel at 200 V with 12 µl of NEB miRNA marker. The gel was stained with SYBR Gold in 1× TBE. Gel fragments between 17 nt and 34 nt were excised and placed in a microfuge tube with 400 µl of gel extraction buffer. Samples were frozen on dry ice for 30 minutes and then thawed overnight with gentle agitation.

After overnight gel extraction, 400 µl of eluate was transferred to a new microfuge tube. The RNA was precipitated by adding 1.5 µl of GlycoBlue and 500 µl of isopropanol. After overnight incubation

at -20 °C, the sample was centrifuged at 20,000g for 30 minutes at 4 °C. The supernatant was discarded, and precipitated RNA was resuspended in 4 µl of 10 mM Tris pH 8. Samples were then dephosphorylated using T4 PNK (4 µl of RNA in 10 mM Tris pH 8, 0.5 µl of T4 PNK enzyme, 0.5 µl of T4 PNK buffer and 0.5 µl of Superasin) at 37 °C for 1 hour. Samples were then subjected to SPRI cleanup: 50 µl of sample in RNase-free water was added to 90 µl of RNAClean beads and 270 µl of isopropanol. After washing with 85% ethanol, beads were resuspended in 7 µl of RNase-free water. The supernatant was collected, and we proceeded with next-generation sequencing library preparation using the Clontech smRNA library prep kit according to the manufacturer's instructions. Libraries were sequenced on an Illumina NovaSeq S2 with single-end 1× 50-nt reads. Samples were always processed in large batches of a maximum of 24 samples.

Human neuron differentiation

The use of hESCs was approved by the Harvard Medical School Embryonic Stem Cell Research Oversight Committee. The transgenic H9 NGN2 hESC line was a generous gift from Alban Ordureau and J. Wade Harper. H9 NGN2 was generated by stably inserting a doxycycline-inducible NGN2 cassette into the *AAVSI* locus of H9 cells (WA09, WiCell)⁵⁹. We collected human neurons from three independent differentiation cohorts, and each replicate exhibited characteristic gene expression patterns reported previously (Extended Data Fig. 4a)²⁹. H9 NGN2 cells were cultured in mTeSR Plus media (STEMCELL Technologies) on tissue culture plates coated with hESC-qualified Matrigel (Corning). They were passaged using Dispase (1 mg ml⁻¹, Life Technologies) until ready for differentiation. A published protocol that combines developmental patterning and NGN2 induction was adapted to differentiate H9 NGN2 into neurons²⁹. At day 0, cells were treated with Accutase (StemPro Accutase, Life Technologies) and plated in single cells at 50,000 cells per cm² in mTeSR Plus media supplemented with 10 µM Y-27632 (STEMCELL Technologies) on tissue culture plates coated with 336.67 µg ml⁻¹ of Growth Factor Reduced Matrigel (Corning). On day 1, the medium was replaced with KSR media (Knockout DMEM medium, 15% knockout serum replacement (KOSR), 2 mM L-glutamine, 1× MEM non-essential amino acids (MEM NEAA), 1× penicillin–streptomycin (pen/strep) and 1× 2-mercaptoethanol (all Gibco)) supplemented with 100 nM LDN193189, 2 µM XAV939 (STEMCELL Technologies), 10 µM SB431542 hydrate and 2 µg ml⁻¹ of doxycycline hyclate (Sigma-Aldrich). Day 2 media was 50% KSR media/50% NIM media supplemented with LDN/XAV/SB and 2 µg ml⁻¹ of doxycycline. NIM media consisted of DMEM/F-12 medium, 1× GlutaMAX, 1× MEM NEAA, 1× pen/strep, 0.16% D-glucose (Sigma-Aldrich) and 1× N2 supplement-B (STEMCELL Technologies). Day 3 media was NIM media supplemented with 2 µg ml⁻¹ of doxycycline. At day 4, cells were treated with Accutase and plated in single cells at 40,000 cells per cm² in NB media (Neurobasal medium (without glutamine), 1× GlutaMAX, 1× MEM NEAA, 1× pen/strep and 1× N2 supplement-B) supplemented with 1× B27 without vitamin A, 2–2.4 µg ml⁻¹ of mouse laminin (Gibco), 1 µM ascorbic acid, 2 µM dibutyl cyclic-AMP (Sigma-Aldrich), 20 ng ml⁻¹ of brain-derived neurotrophic factor, 10 ng ml⁻¹ of glial-derived neurotrophic factor (rhBDNF and rhGDNF, PeproTech), 10 µM Y-27632 and 2 µg ml⁻¹ of doxycycline on the tissue culture plates coated with 336.67 µg ml⁻¹ of Growth Factor Reduced Matrigel. The media at day 4 without Y-27632 and doxycycline is referred to as complete NB (cNB) media. On day 5, the media was replaced with cNB media. Thereafter, half of the media was replaced weekly with cNB_{2x}, where concentrations of all the supplements to the NB media (except Y-27632) were doubled. In between each media change, media was directly supplemented with 2 µg ml⁻¹ of doxycycline on the third day of the week (days 8, 15 and 22). Cells were silenced on day 27 with TTX (Abcam) and APV (Tocris), which antagonize sodium channels and NMDA receptors, respectively. Cells were collected at day 28 in 1× PBS supplemented with 1× cycloheximide after being stimulated with 55 mM KCl for 0 hours or 6 hours.

Ribosome profiling of human neurons

Ribosome profiling of human NGN2-induced neurons was performed as described above for human brain tissue, except that RNase I digestion time was 15 minutes. For harringtonine treatment, $2 \mu\text{g ml}^{-1}$ of harringtonine or an equal volume of DMSO was added to cell culture media and incubated at 37°C for 2 minutes before proceeding with Ribo-seq cell lysis and sample preparation as described above.

RNA-seq library preparation

RNA-seq libraries were prepared from 10 ng of total RNA using the SMARTer Stranded Total RNA-seq Pico Input Mammalian V2 kit (Takara Bio) according to the manufacturer's instructions. Samples were multiplexed with Illumina TruSeq HT barcodes and sequenced on a NextSeq 2000 with single-end 1×75 -nt reads. Samples were always processed in large batches of a maximum of 24 samples to minimize sample processing biases.

Analysis of RNA-seq data

In an effort to capture the most complete picture of translation, including the potential translation of brain-specific ncRNAs, RNA-seq and Ribo-seq reads were mapped to the lncRNA knowledge base (lncRNAKB) annotation⁶⁰. This annotation includes experimental evidence of ncRNA expression across 31 solid human normal tissues, including the brain, providing a comprehensive resource of transcripts and transcript isoforms in the human brain. Sequencing reads were aligned using Hisat2 (version 2.1.0) to the *Homo sapiens* genome (GRCh30) and transcriptome (lncRNAKB). Alignments and analysis were performed on the Orchestra2 high performance computing cluster through Harvard Medical School. Aligned BAM files were sorted using Picard Tools (version 2.8.0) and filtered for reads that uniquely aligned to remove multi-mapped reads using samtools (version 1.9); stranded bedGraphs were generated using STAR; and reads were quantified over annotated exons using HTSeq-count (version 0.9.1).

ORF calling and filtering with RibORF

Sequencing adapters were removed using Cutadapt (version 1.14); trimmed FASTQ files were aligned to hg38 ribosomal RNA sequences using Bowtie2 (version 2.3.4.3); and unaligned reads were mapped to the hg38 genome and lncRNAKB transcriptome⁶⁰ using STAR (version 2.7.3a) with standard settings and the following modified parameters: `-clip5pNbases 3`, `-seedSearchStartLmax 15`, `-outSJfilterOverhangMin 30 8 8 8`, `-outFilterScoreMin 0`, `-outFilterScoreMinOverLread 0.66`, `-outFilterMatchNmin 0`, `-outFilterMatchNminOverLread 0.66`, `-outSAMtype BAM Unsorted`. Aligned BAM files were filtered for only uniquely mapped reads and sorted using Picard Tools (version 2.8.0), and stranded bedGraphs were generated using STAR. The RibORF pipeline was run on each sample individually using standard parameters. Due to template switching during library preparation, reads contained three untemplated bases at the 3' end that were not included in the alignment but added to the length of each read. Therefore, reads 30–33 nt in length (corresponding to RNA fragments 27–30 nt) were analyzed for three-nucleotide periodicity within known protein-coding ORFs (RefSeq). For each sample, we selected only the read lengths for which at least 50% of the reads matched the primary ORF of known protein-coding genes in a meta-gene analysis. Samples with fewer than two read lengths passing filtering were removed from further analysis. Read lengths were offset-corrected, and RibORF was used to predict ORFs with a minimum length of 8 aa and translation probability >0.7 . Only samples in which frame 0 periodicity was $>50\%$ for at least two read lengths and an overall area under the curve (AUC) >0.9 were included in the final analysis.

After running the RibORF pipeline on each brain sample individually, information from RibORF output files was used to generate GTF and BED files for all ORFs identified in each sample. ORFs with lengths of 0 and ORFs annotated as non-coding despite being detected in

protein-coding genes were eliminated. Using bedtools version 2.27.1 and the GRCh38 primary assembly human genome file, DNA sequences were associated with each exon of each remaining ORF. ORFs that did not end in stop codon sequences ('TGA', 'TAA' and 'TAG') were eliminated. Using the R library micropan version 2.1, DNA sequences for each complete ORF were translated into protein sequences. Of note, ORFs with start codons 'GTG' or 'TTG' are translated with a methionine as the initial amino acid despite these sequences not typically encoding methionine in other positions in a protein-coding DNA sequence, per existing literature on non-canonical start codon usage in translation. Finally, all remaining ORF information was collapsed into one table, and duplicate ORFs, defined as ORFs in the same genomic position with identical protein sequences, were eliminated. When eliminating duplicates, ORFs identified of the most common ORF type identified by RibORF were conserved, according to the following order of priority, from highest to lowest: canonical, truncation, extension, overlap, uORF, internal, external, polycistronic, readthrough and non-coding ORFs. ORFs annotated as type 'seqerror' were eliminated. After combining ORF outputs from all samples, ORFs that were detected in only one sample were eliminated. After the removal of singleton ORFs, duplicate ORFs were once again eliminated according to the same priority scheme, leaving only one entry for each ORF that was detected in at least two samples in the dataset. Importantly, the large number of unique protein sequences in the unannotated protein dataset does not reflect a large number of unique genes; rather, these represent alternative coding regions and/or isoforms of genes. The lncRNAKB annotation was used to assign a specific ORF type to each ORF. To be designated a sORF, an ORF had to be 100 aa or less in length and not fully overlap in-frame with a canonical protein-coding ORF.

Single-cell deconvolution from bulk RNA-seq data

The SCDC R package (version 0.0.0.9000)⁶¹ was used to approximate the distribution of cell types in our human postmortem tissue samples, using single-cell RNA-seq data from phenotypically normal human dorsolateral prefrontal cortex samples⁶² as a cell type reference. Raw counts data from control samples in the reference dataset were normalized, and the distribution of cell types present in each of the human postmortem tissue samples in this study were determined using SCDC. Cell types not represented in any sample were removed from the single-cell reference dataset, and then the cell type distribution was re-run using SCDC.

Differential expression and Gene Ontology enrichment analysis

RNA-seq gene expression was quantified as described above, and lowly expressed genes were filtered for counts per million >1 in at least two samples using edgeR (version 3.26.8). Ribo-seq expression was quantified by counting the number of P-sites over a given ORF. To identify differences in transcription and translation between adult and prenatal human brain, two-way differential expression analysis was performed using deltaTE¹⁷ in R version 4.0.1. Read normalization and size factor estimation were performed on RNA-seq and Ribo-seq data simultaneously; samples were corrected for batch effects; and ORF types were subsetted for display purposes. Gene Ontology (GO) enrichment analysis was performed using gProfiler2 in R (version 0.2.0), with a custom background of expressed genes based on expression-filtered RNA-seq genes and false discovery rate (FDR) <0.05 .

ORF validation in NGN2 neurons with harringtonine-treated Ribo-Seq

Paired harringtonine-treated and vehicle-treated Ribo-Seq BAM files were analyzed using RiboTISH (<https://github.com/zhpn1024/ribotish>)³¹. ORFs were identified using a negative binomial model to fit background from harringtonine-treated samples, followed by testing significance of translation initiation sites. ORFs were filtered for FDR q -value <0.05 , and ORFs detected by RibORF were considered

validated if the start codon of the ORF called by ORF-RATER and RibORF perfectly matched.

Paired harringtonine-treated and vehicle-treated Ribo-seq BAM files were analyzed using ORF-RATER (<https://github.com/alexfields/ORF-RATER>)⁶³. ORFs were filtered for an orfrating >0, and ORFs detected by RibORF were considered validated if the start codon of the ORF called by ORF-RATER and RibORF perfectly matched.

uORF/canonical ORF correlation analysis

For each mRNA transcript detected in each individual human brain sample, uORF sequences identified by RibORF were joined to produce one singular sample-specific uORF region. For each of the uORF regions, raw counts were generated by quantifying total P-sites across each region, from which transcripts per million reads (TPMs) were calculated. These TPMs were compared to canonical ORF translational efficiency values to characterize the relationship between uORF utilization and canonical ORF translational dynamics at the level of individual genes.

Protein sequence analysis by liquid chromatography with tandem mass spectrometry

Size-selected proteomics of the human adult and prenatal brain, as well as hESC-derived neurons, was performed at the Taplin Biological Mass Spectrometry Facility at Harvard Medical School. Excised gel bands were cut into approximately 1-mm³ pieces. Gel pieces were then subjected to a modified in-gel trypsin digestion procedure. Gel pieces were washed and dehydrated with acetonitrile for 10 minutes, followed by removal of acetonitrile. Pieces were then completely dried in a speed-vac. Rehydration of the gel pieces was with 50 mM ammonium bicarbonate solution containing 12.5 ng μl^{-1} of modified sequencing-grade trypsin (Promega) at 4 °C. After 45 minutes, the excess trypsin solution was removed and replaced with 50 mM ammonium bicarbonate solution to just cover the gel pieces. Samples were then placed in a 37 °C room overnight. Peptides were later extracted by removing the ammonium bicarbonate solution, followed by one wash with a solution containing 50% acetonitrile and 1% formic acid. The extracts were then dried in a speed-vac (-1 hour). The samples were stored at 4 °C until analysis.

On the day of analysis, samples were reconstituted in 5–10 μl of HPLC solvent A (2.5% acetonitrile and 0.1% formic acid). A nano-scale reverse-phase HPLC capillary column was created by packing 2.6 μm of C18 spherical silica beads into a fused silica capillary (100- μm inner diameter \times -30-cm length) with a flame-drawn tip. After equilibrating the column, each sample was loaded via a Famos auto sampler (LC Packings) onto the column. A gradient was formed, and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile and 0.1% formic acid).

As peptides eluted, they were subjected to electrospray ionization and then entered into an LTQ Orbitrap Velos Pro ion-trap mass spectrometer (Thermo Fisher Scientific). Peptides were detected, isolated and fragmented to produce a tandem mass spectrum of specific fragment ions for each peptide.

Mass spectrometry analysis

Thermo Fisher raw files were loaded into MaxQuant version 1.6.17.0 for the peptide search. Each file corresponded to one brain sample and was labeled as its own experiment in the search. Default parameters, including specific trypsin digestion, methionine oxidation and protein N-terminal acetyl variable modifications and carbamidomethyl-fixed modifications, were used. We uploaded a custom protein FASTA file for our search using the protein sequence identified in our RibORF post-processing. For adult brain mass spectrometry, we used a protein FASTA file containing only sequences from adult samples that passed our quality control metrics and the same for prenatal brain mass spectrometry. The size of each search database was as follows: adult brain – 53,326 ORFs; prenatal brain – 98,410 ORFs; NGN2 neurons – 84,450

ORFs. In each case, ‘truncation’ type ORFs were excluded because of their redundancy to canonical protein sequences. The protein search in MaxQuant was run using an Amazon Web Services client to optimize speed and efficiency. A default two-level FDR control was used: peptide level and protein group level, both with a 1% FDR threshold. A posterior error probability calculation was performed based on a target-decoy search. Common mass spectrometry contaminants were filtered out. Only peptides with a score >50 were considered for subsequent analysis.

Physicochemical analysis

sORFs were searched using Blastp (version 2.6.0, -evalue 0.0001, -word_size 4) against a database of all protein translations from GENCODE v29 (https://www.genecodegenes.org/human/release_29.html; downloaded on 2 August 2019). GO terms enriched in known proteins with significant sequence similarity to sORFs were determined using GORilla⁶⁴ and plotted using GraphPad Prism version 8 software. Locations of significant hits were then compared to Pfam annotations (from Ensembl, accessed through Ensembl API), and any sORF with at least one residue overlap was considered overlapping. Although we allowed any degree of overlap, we found that many of the sORFs had near-complete overlap with Pfam domains (Extended Data Fig. 7b). FoldIndex score⁴⁷ is defined as $2.785 \times \text{hydropathy} - \text{abs}(\text{net charge}) - 1.151$. Physicochemical and sequence properties of sORFs and intrinsically disordered regions (IDRs) were computed using custom Python codes (<https://github.com/IPritisanac/idr.mol.feats>). All analyses of sORFs, IDRs and reference proteins were performed on protein sequences between 21 aa and 100 aa. In total, 27,110 sORFs, 19,652 IDRs and 908 UniProt human reference proteins met this criterion. Normalization, filtering and clustering of sequence properties was performed using Cluster3.0 (ref. ⁶⁵) with the following parameters: median centering of columns; normalization of columns; retaining sequences with at least three observations with absolute value >0.01 and weighting columns using default options; and clustering using average linkage hierarchical clustering. This process left 16,905 sequences in the cluster analysis, of which 6,910 were IDRs and 10,095 (59%) were sORFs. Clusters were visualized and selected manually using Treeview version 1.1.6r4 (ref. ⁶⁶), and enrichment analysis was performed by selecting the IDRs from each cluster and using the 6,910 IDRs in the entire cluster analysis as the background set. These lists were entered into the GORilla webserver⁶⁴. The enrichment or depletion of sORFs in each cluster was computed by comparing the ratio of sORFs in each cluster to the expected ratio of 10,095/6,910.

Phylostratigraphy analysis

All ORFs with an amino acid length of ≥ 40 aa were analyzed using TimeTree⁶⁷ to identify the minimal evolutionary age for every protein-coding gene. The evaluation is based on sequence similarity scored with Blastp and identifying the most distant sequence in which a sufficiently similar sequence appears. Each protein sequence was used to query the non-redundant (nr) National Center of Biotechnology Information (NCBI) database with a Blastp E-value threshold of 10^{-3} and a maximum number of 200,000 hits. We identified the phylostratum in which each ORF appeared. Each phylostratum corresponds to an evolutionary node in the lineage of the species, as listed in the NCBI Taxonomy database. For clarity, we aggregated results into the following evolutionary eras: Ancient (phylostrata 1–7, from cellular organisms through Deuterostomia (290–747 millions of years ago (Mya))), Chordates (phylostrata 8–17, from Chordata through Amniota (747–320 Mya)), Mammals (phylostrata 18–22, from Mammalia through Euarchontoglires (320–91 Mya)), Primates (phylostrata 23–29, from Primates through Hominae (91–6.6 Mya)) and Humans (phylostrata 30–31, including *H. sapiens*, 6.6 Mya to present).

Transposable element insertion at start codons

To identify ORFs whose start codons derive from transposable elements, we intersected our ORF start codons with a TE annotation kindly

provided by the laboratory of Didier Trono⁶⁸. First, we created a list of all start codons in different categories (protein-coding, non-coding RNA, uORF, sORF and pseudogene) by collapsing all ORFs that share a start position. We extended this start codon position to a 10-bp window and intersected this with a BED file of all TEs in the human genome using bedtools. For any ORF start codon that overlapped a TE, we used a table of TE subfamily ages from Dfam to estimate the oldest possible lineage in which that TE may exist in the human genome^{36,69}.

Microprotein overexpression and western blot

To test the translatability of sORFs, double-strand DNA sequences containing the sORF endogenous pseudo 5' UTR (defined as the upstream DNA sequence from the sORF start codon), sORF protein sequence and a FLAG-HA tag in-frame with the sORF protein sequence were synthesized by GenScript and cloned into an FUGW overexpression vector (Addgene, 14883). A negative control construct in which the start codon was mutated to an ATT was generated using a QuikChange II Site-directed mutagenesis kit (Agilent, 200521). The wild-type and mutant plasmids were verified by Sanger sequencing. The designed sequences used in this study are listed in Supplementary Table 6.

Both wild-type and mutant plasmids were transfected into lentiX-293T cells (Takara Bio) using Lipofectamine 3000 reagent (Invitrogen). After 24 hours, cells were harvested and resuspended in RIPA buffer (Sigma-Aldrich) supplemented with protease inhibitor cocktail (Roche). Protein concentration was measured by Bradford assay, and 20 µg of protein lysate was denatured at 95 °C for 5 minutes and then separated on a 10–20% Tris-tricine gel (Invitrogen) at 125 V for 90 minutes. Proteins were transferred to a nitrocellulose membrane (Bio-Rad) at 115 V for 90 minutes, and the membrane was blocked with 5% non-fat dry milk in TBST for 1 hour. Membranes were incubated with anti-HA (C29F4) (1:1,000, CST) or anti-GAPDH (1:1,000, Sigma-Aldrich) primary antibody in 5% non-fat dry milk in TBST overnight at 4 °C. Membranes were washed 4× in TBST at room temperature and then incubated with secondary antibodies conjugated to IRDye 800 (1:10,000) and imaged with LiCOR Odyssey.

Immunofluorescence

HEK293T cells were grown on glass slides for 24 hours and transfected as described above. Cells were fixed with 4% paraformaldehyde for 30 minutes at room temperature and washed three times with ice-cold PBS. The cells were permeabilized and blocked for 1 hour at room temperature using 5% donkey serum in PBST (1× PBS and 0.1% Triton X-100). Coverslips were incubated with anti-FLAG mouse primary antibody (1:1,000) (Sigma-Aldrich) overnight at 4 °C. Coverslips were washed 3× in PBST at room temperature and then incubated with fluorescently labeled secondary antibody (1:2,000, Alexa Fluor 488 anti-mouse) for 1 hour at room temperature. Coverslips were washed 3× in PBST at room temperature and mounted onto Superfrost glass slides using DAPI Fluoromount-G (Thermo Fisher Scientific). Images were visualized using a Leica SP8 confocal microscope using a ×63 objective and analyzed using ImageJ software (version 2.1.0).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Data reprocessed from Wang et al.⁷⁰ were accessed from ArrayExpress with accession number [E-MTAB-7247](https://www.ebi.ac.uk/ena/arrayexpress/experiments/E-MTAB-7247). Human brain primary tissue RNA-seq and Ribo-seq data have been submitted to the database of Genotypes and Phenotypes (dbGaP) under accession number [phs002489](https://www.ncbi.nlm.nih.gov/gap/studies/phs002489). NGN2 RNA-seq and Ribo-seq data have been submitted to the Gene Expression Omnibus under accession number [GSE180240](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180240). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with

the dataset identifier [PXD035950](https://www.ebi.ac.uk/ena/arrayexpress/experiments/PXD035950). Our web-based searchable database is available from <https://greenberg.hms.harvard.edu/project/human-brain-orf-database/>. Source data are provided with this paper.

Code availability

Custom Python code for physicochemical analysis is available at <https://github.com/IPritisanac/idr.mol.feats>. All other code used in this study was previously published and is cited in the Methods section.

References

- McGlincy, N. J. & Ingolia, N. T. Transcriptome-wide measurement of translation by ribosome profiling. *Methods* **126**, 112–129 (2017).
- Ordureau, A. et al. Global landscape and dynamics of Parkin and USP30-dependent ubiquitylomes in iNeurons during mitophagic signaling. *Mol. Cell* **77**, 1124–1142.e10 (2020).
- Seifuddin, F. et al. lncRNAKB, a knowledgebase of tissue-specific functional annotation and trait association of long noncoding RNA. *Sci. Data* **7**, 326 (2020).
- Dong, M. et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).
- Nagy, C. et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. *Nat. Neurosci.* **23**, 771–781 (2020).
- Fields, A. P. et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. *Mol. Cell* **60**, 816–827 (2015).
- Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOrrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).
- de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
- Saldanha, A. J. Java Treeview—extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
- Kumar, S., Stecher, G., Suleski, M. & Heddes, S. B. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
- Turelli, P. et al. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. *Sci. Adv.* **6**, eaba3200 (2020).
- Huble, R. et al. The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44**, D81–D89 (2016).
- Wang, Z.-Y. et al. Transcriptome and translome co-evolution in mammals. *Nature* **588**, 642–647 (2020).
- Boulting, G. L. et al. Activity-dependent regulome of human GABAergic neurons reveals new patterns of gene regulation and neurological disease heritability. *Nat. Neurosci.* **24**, 437–448 (2021).

Acknowledgements

This research was supported by the Allen Discovery Center program, a Paul G. Allen Frontiers Group advised program of the Paul G. Allen Family Foundation. E.E.D. was supported by the Damon Runyon Cancer Research Foundation (DRG-2397-20). A.C.C. was supported by the Hanna H. Gray Fellowship through the Howard Hughes Medical Institute. V.L. was supported by a Boston Children's Hospital Career Development Award (to A.O.D.L.), and V.L. and N.S. were supported by National Human Genomic Research Institute (NHGRI) R01HG010898 and NIH 1R01HG010898-01A1. E.J.H. was supported by NIH P01 NS083513. A.O.D.L. and W.P. were supported by a Manton Center Endowed Scholar Award and NHGRI U01HG008900. M.A.S. was supported by National Institute of Mental Health (NIMH) F31MH124393. A.M.M. and J.D.F.-K. acknowledge funding from the Canadian Institutes of Health Research (CIHR PJT-148532). B.T.K.

was supported by National Institute of Neurological Diseases and Stroke (NINDS) K08 NS112338. M.E.G. was supported by funding from NINDS R01 NS115965. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank members of the Greenberg laboratory for helpful discussions on the manuscript. We thank the Taplin Mass Spectrometry Facility at Harvard Medical School for their technical expertise and analysis of proteomics samples and S. Slavoff for advice on size-selection proteomics. We thank the Broad Institute Genomics Program for next-generation sequencing of ribosome profiling libraries. We are grateful to the NIH NeuroBioBank and the Human Developmental Biology Resource for providing human adult and prenatal brain tissue, respectively. We are grateful to the laboratory of D. Trono for sharing the human transposable element annotation and to W. Harper for reagents and technical advice related to iPSC-derived human neurons. We thank the Neurobiology Department and the Neurobiology Imaging Facility for consultation and instrument availability that supported this work. This facility is supported, in part, by the Neural Imaging Center as part of an NINDS P30 Core Center grant (NS072030). Figures 3d, 4a and 7d were made with BioRender.

Author contributions

E.E.D., B.T.K. and M.E.G. conceptualized the study and designed the experiments. E.E.D. and B.T.K. performed ribosome profiling and RNA-seq. E.E.D., B.T.K. and B.F. analyzed ribosome profiling and RNA-seq. E.E.D. and J.C. analyzed harringtonine-treated Ribo-seq datasets. G.C. prepared human embryonic stem-cell-derived neurons for ribosome profiling. A.C.C. analyzed the insertion of TEs at ORF translation start sites. I.P., A.A., J.F.-K. and A.M.M. performed

physiochemical analysis. V.L., A.O.D.L, A. Karger, W.P. and N.S. performed phylostratigraphy analysis. B.T.K. prepared samples for proteomics, and B.T.K., B.F. and A. Khitun analyzed proteomics data. E.E.D., E.G.A. and N.P. performed microprotein validation experiments in 293T cells. M.S. and B.B. performed disease heritability analysis. E.E.C. and E.J.H. provided prenatal brain tissue samples and technical advice for tissue processing. S.G. assisted with Ribo-seq quality control analyses. E.E.D., B.T.K., B.F., E.C.G. and M.E.G. drafted the manuscript, with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

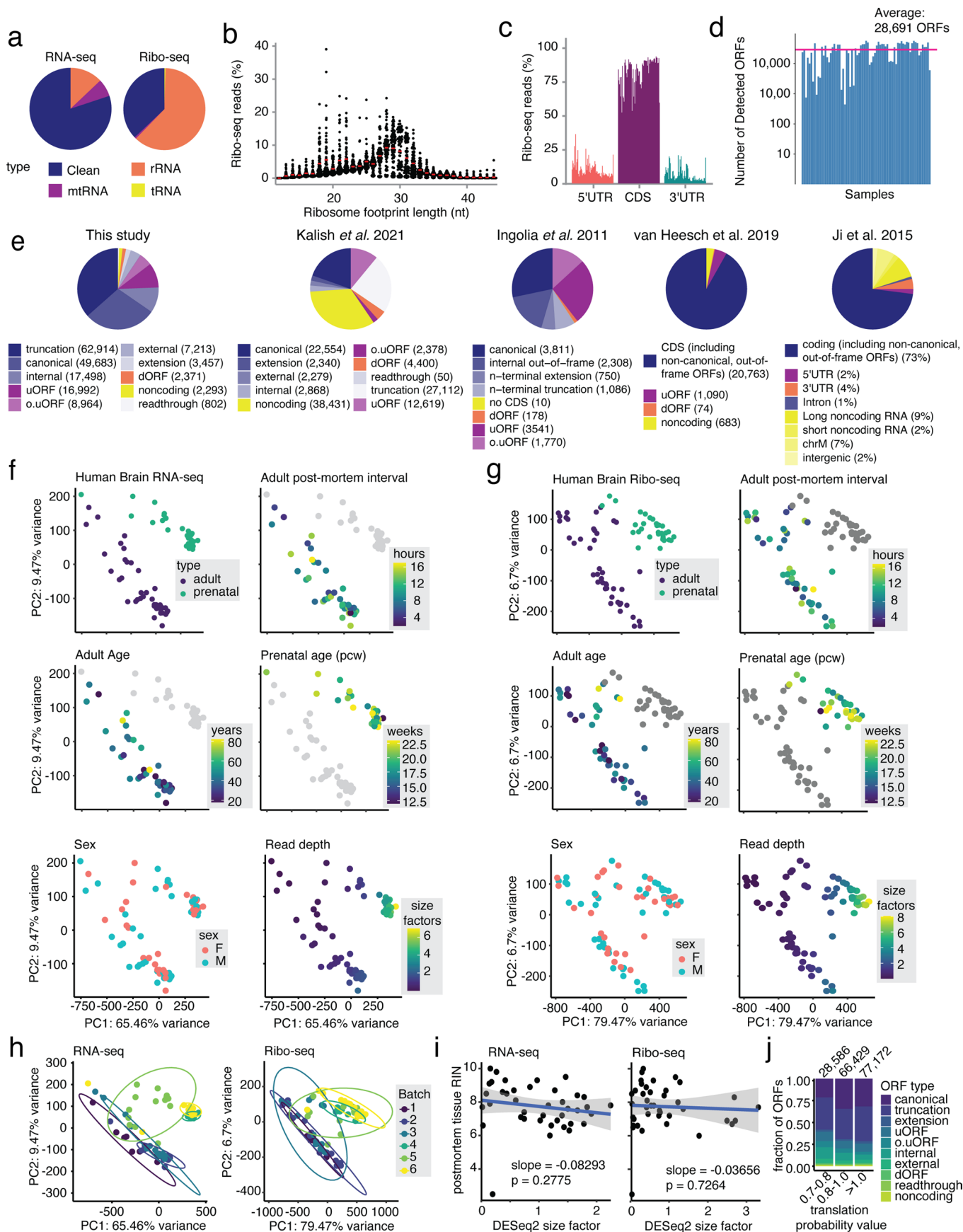
Extended data is available for this paper at <https://doi.org/10.1038/s41593-022-01164-9>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41593-022-01164-9>.

Correspondence and requests for materials should be addressed to Erin E. Duffy, Brian T. Kalish or Michael E. Greenberg.

Peer review information *Nature Neuroscience* thanks Linnaea Ostroff and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

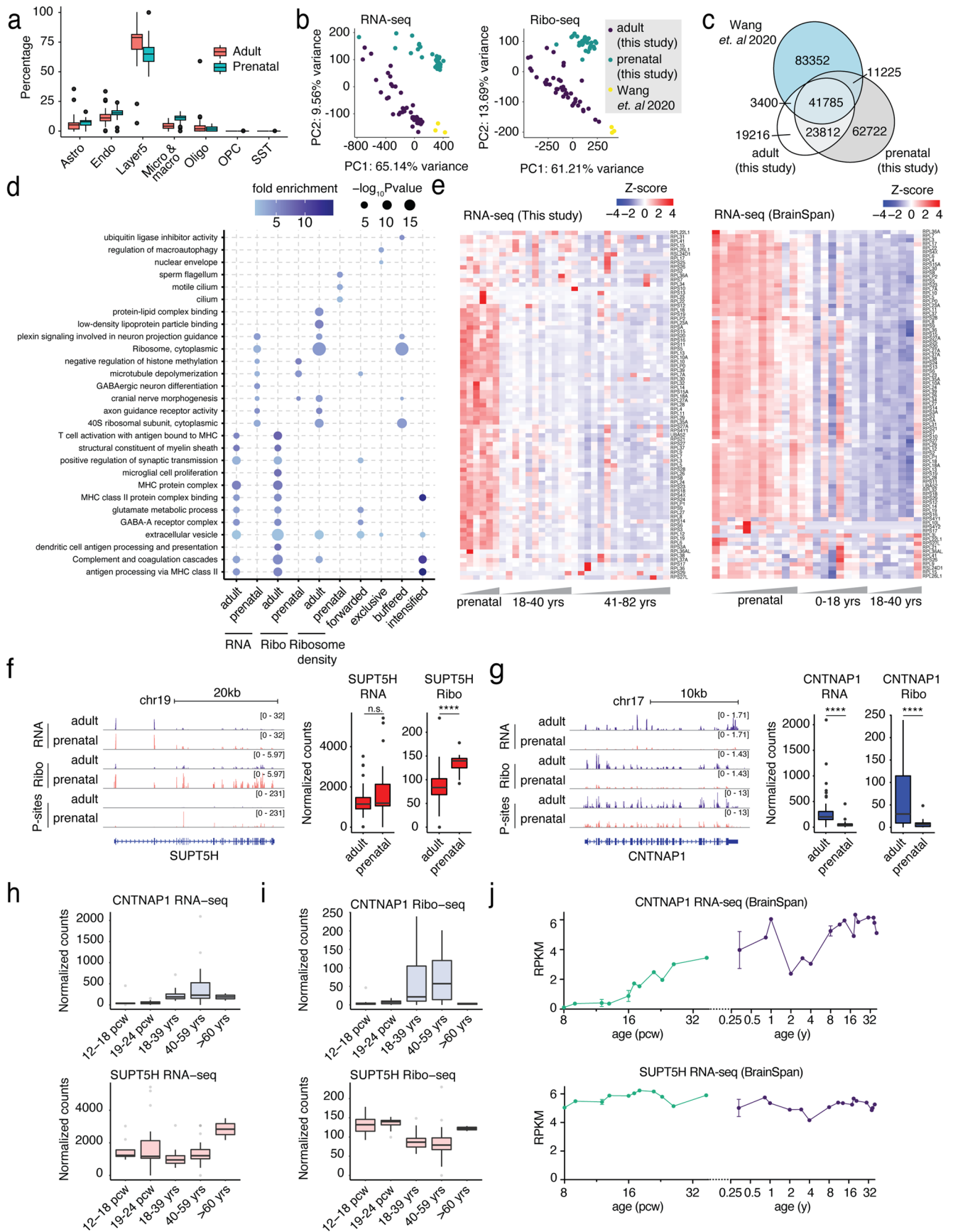
Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Ribosome profiling captures active translation in the human adult and prenatal brain, Related to Fig. 1. (a) Pie chart displaying the fraction of raw sequence reads derived from tRNA, ribosomal RNA (rRNA), mitochondrial RNA (mtRNA), and remaining aligned reads (clean) from human adult and prenatal brain RNA-seq and Ribo-seq. (b) Beeswarm plot of sequenced ribosome footprint lengths across all 73 adult and prenatal brain samples. Red line indicates the average percentage of Ribo-seq reads assigned to a given read length across all samples. (c) Bar plot of the percentage of reads mapping to the coding sequence (CDS) and untranslated regions (5' and 3' UTR) of annotated protein-coding genes (Refseq hg38). Each bar represents an individual sample. (d) Bar plot of the number of ORFs identified by RibORF in each sample after filtering (see Methods). Notably, we identify many more ORFs in the prenatal brain compared to the adult brain, which is likely at least in part a result of the longer post-mortem interval in adult compared to prenatal samples. (e) Pie chart of ORF types detected in our study as well as four previous studies^{4,11-13}. Because each study defines ORF types differently, ORFs are in hues denoting similar ORF types. Blue = CDS ORFs including non-canonical, out-of-frame ORFs; pink = uORFs and overlapping uORFs; orange = dORFs; yellow = ORFs translated from previously annotated non-coding RNAs. (f-g) PCA analysis of all genes in

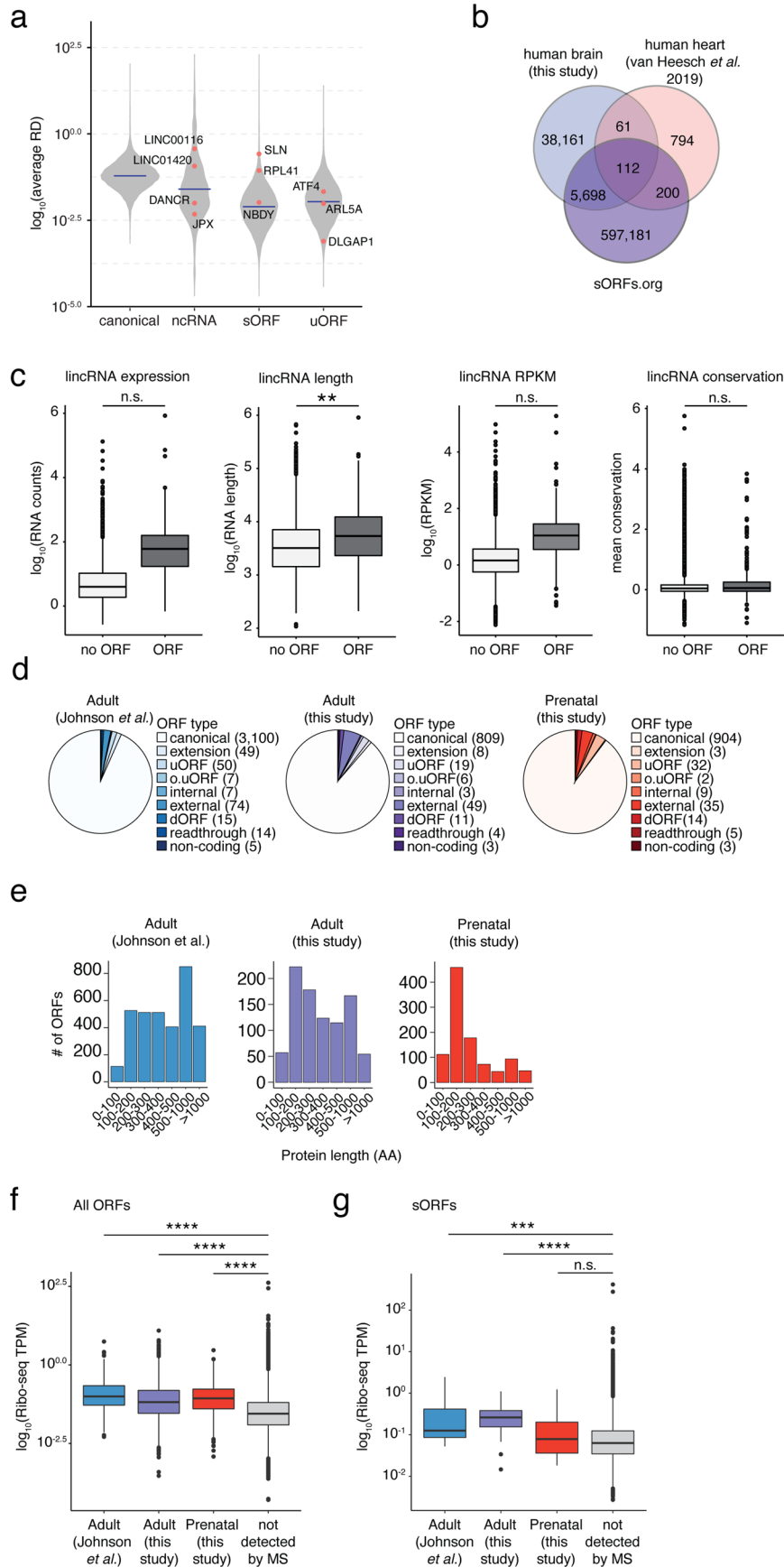
the human brain (f) RNA-seq and (g) Ribo-seq, colored by sample type (adult vs prenatal), post-mortem interval for adult samples, adult age, prenatal age (pcw), sex, and read depth (based on DESeq2 scale factors of estimated library size). The validity of combining samples into two groups in subsequent analyses was confirmed by the finding that these two groups were well separated by PCA analysis for both the transcriptome and translome. (h) PCA analysis of batches of samples processed for RNA-seq (left) and Ribo-seq (right) to test for batch effect. We note a mild batch effect in some cases. To address the remaining batch effects in our differential expression analysis, we employed DTEG, an algorithm that uses DESeq2 to normalize across samples and includes batch correction. (i) Linear regression between postmortem tissue RNA integrity number (RIN) and read depth for RNA-seq and Ribo-seq. There is no significant correlation between the two variables (two-sided significance test for linear regression, $p > 0.05$ for RNA-seq and Ribo-seq). Gray shading = 95% CI. (j) Stacked bar plot of ORF types distributed by translation probability value, as calculated by RibORF. While RibORF uses a translation probability cutoff of 0.7 to determine significantly translated ORFs, we detect novel ORFs with a large range of translation probabilities.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Transcriptional and translational regulation across human brain development, Related to Fig. 2. While the transcriptome has not been previously characterized in the developing human brain, our measurements of the transcriptome are consistent with published gene expression data from the BrainSpan Atlas of the Developing Human Brain¹⁵. **(a)** Bulk gene expression deconvolution of adult and prenatal brain RNA-seq using multiple single-cell RNA sequencing references (SCDC)⁶¹. Differences in cell type composition are not statistically significant by two-way ANOVA (F-statistic = 0, P-value = 1), strongly suggesting that observed differences in transcription and translation between prenatal and adult brain samples are not driven by differences in cell type composition. Data are shown as median \pm IQR (whiskers = 1.5*IQR), notches indicate median \pm $-1.58*IQR/\sqrt{n}$, n = 43 (adult) and 30 (prenatal) biologically independent tissues. **(b)** We sought to benchmark the Ribo-seq data in human postmortem brain tissue in this study to a recently published dataset of Ribo-seq in the human cerebellum⁷⁰. PCA analysis of human prenatal and adult Ribo-seq (this study), as well as Ribo-seq of human adult cerebellum from Wang et al⁷⁰. **(c)** Diagram of overlapping ORFs identified from human prenatal and adult Ribo-seq (this study), as well as Ribo-seq of human adult cerebellum from Wang et al⁷⁰. Overall, we found a similar level of overlap between the Wang data and our human brain tissue samples as we found with our human NGN2 neuron samples. We do not find this limited overlap surprising, however, given that the Wang data is from adult cerebellum whereas we sample dorsolateral prefrontal cortex across a variety of ages. **(d)** Dot plot of the top enriched GO terms in each regulatory category defined in Fig. 2B. **(e)** Heatmap of RNA-seq expression (row-normalized) for all ribosomal genes in Fig. 2D across all human adult and prenatal samples in this study (left) and in

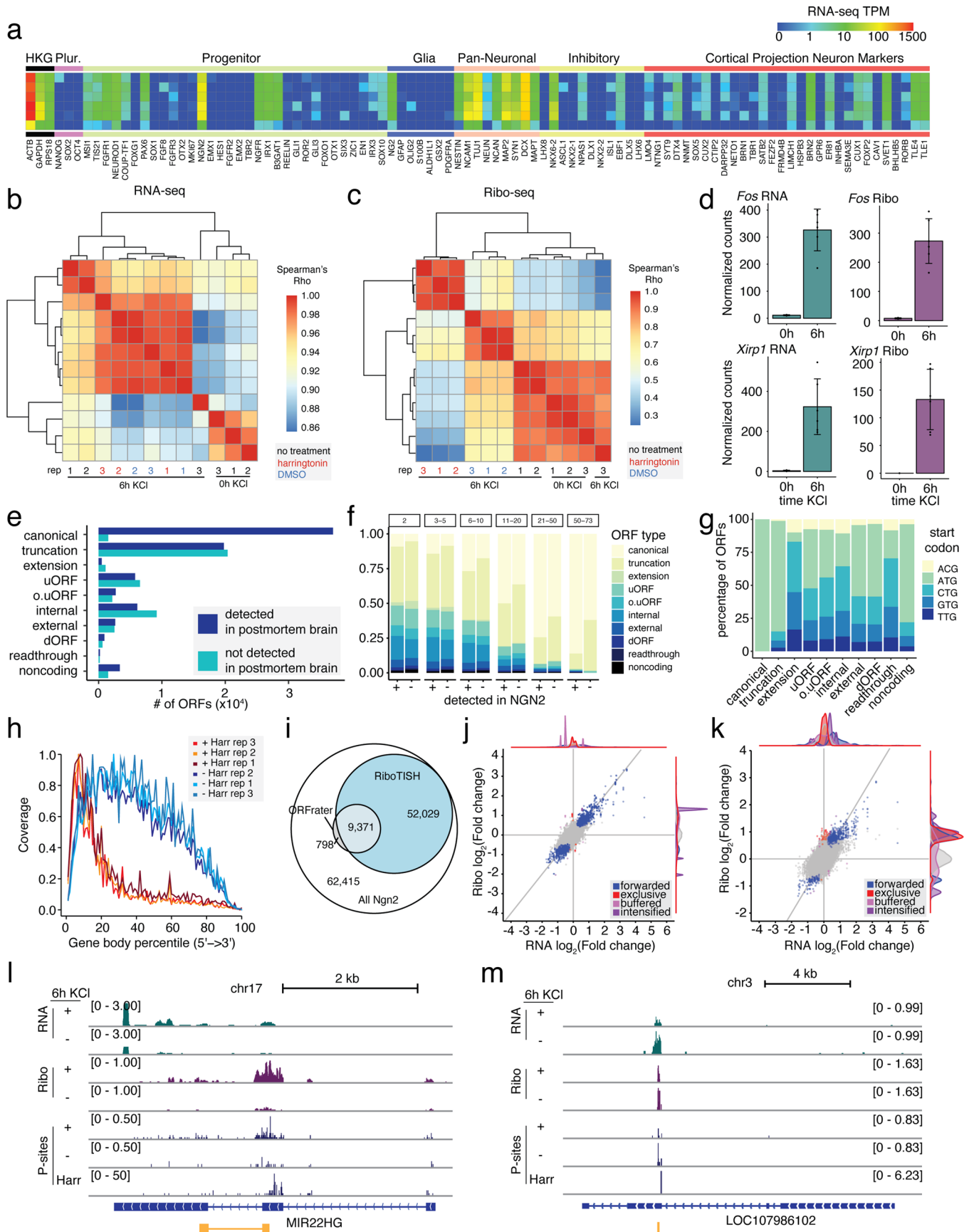
the dorsolateral prefrontal cortex from the BrainSpan Atlas of the Developing Human Brain¹⁵. **(f)** Genomic locus of *SUPTSH*, a transcript regulated only at the level of translation, and **(g)** *CNTNAPI*, a transcript regulated only at the level of transcription. Tracks represent merged and depth-normalized reads across all adult vs. prenatal samples for RNA-seq, Ribo-seq, as well as P-site positions. Box and whisker plots indicate DESeq2-normalized RNA-seq reads and Ribo-seq P-sites in adult vs. prenatal samples. **(f)** **** = Ribo-seq $p_{adj} = 2.26*10^{-4}$ by DESeq2. **(g)** **** = RNA-seq $p_{adj} = 3.74*10^{-12}$ by DESeq2; **** = Ribo-seq $p_{adj} = 1.35*10^{-6}$ by DESeq2. **(f-g)** Data are shown as median \pm IQR (whiskers = 1.5*IQR), n = 43 (adult) and 30 (prenatal) biologically independent tissues. **(h)** Box and whisker plots of DESeq2-normalized RNA-seq reads across human brain samples divided into five age categories (prenatal = 12-18 pcw and 19-24 pcw, adult = 20-39 yrs, 40-59 yrs, >60 yrs) for *CNTNAPI* (top), a transcriptionally regulated gene, and *SUPTSH* (bottom), a translationally regulated gene. **(i)** Box and whisker plots of DESeq2-normalized Ribo-seq reads across human brain samples divided into five age categories (prenatal = 12-18 pcw (n = 8 biologically independent tissues) and 19-24 pcw (n = 22 biologically independent tissues), adult = 18-39 yrs (n = 18 biologically independent tissues), 40-59 yrs (n = 23 biologically independent tissues), >60 yrs (n = 2 biologically independent tissues)) for *CNTNAPI* (left) and *SUPTSH* (right). **(h-i)** Data are shown as median \pm IQR (whiskers = 1.5*IQR). **(j)** Line plot of RPKM values from the BrainSpan Atlas of the Developing Human Brain for *CNTNAPI* (top) and *SUPTSH* (bottom) across development in the dorsolateral prefrontal cortex. Data are shown as mean \pm SD, n = 1 (8, 9, 17, 19, 21, 24, 26 & 37 pcw; 0.83, 1, 2, 33, 4, 11, 13, 18, 19, 21, 30, 36, 37, 40 yrs), 2 (0.33, 8 yrs), 3 (12, 13, 16 pcw) biologically independent tissues.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Microprotein expression and validation across brain development, Related to Fig. 3. (a) Violin plot of average ribosome density (RD) by ORF type. Previously described ORFs are shown in red. Average ribosome density is shown in blue. (b) Venn diagram of sORFs detected in human brain (this study), human heart (van Heesch et al.⁴), or the sORF.org database. In total, 6,071 translated sORFs identified in the human brain perfectly matched the amino acid sequence of a previously reported entry in the sORFs.org database or identified in the human heart, a degree of overlap consistent with prior studies⁴. (c) Out of 8,590 lincRNA genes expressed across all brain samples, 415 lincRNA genes encode at least one translated ORF. We examined possible differences between ORF-encoding lincRNAs and non-translated lincRNAs. Box and whisker plots of annotated lincRNA features (expression, length, RPKM, and conservation) comparing RNAs that contain at least one ORF in the human brain to lincRNAs that do not contain any ORFs, data are shown as median \pm IQR (whiskers = 1.5*IQR), $n = 8,175$ (no ORF detected) and 415 (ORF detected) lincRNAs. ** $p = 0.009079$ by two-sided Welch two-sample T-test, n.s. = not significant. (d) Number and type of ORFs identified by size-selection proteomics

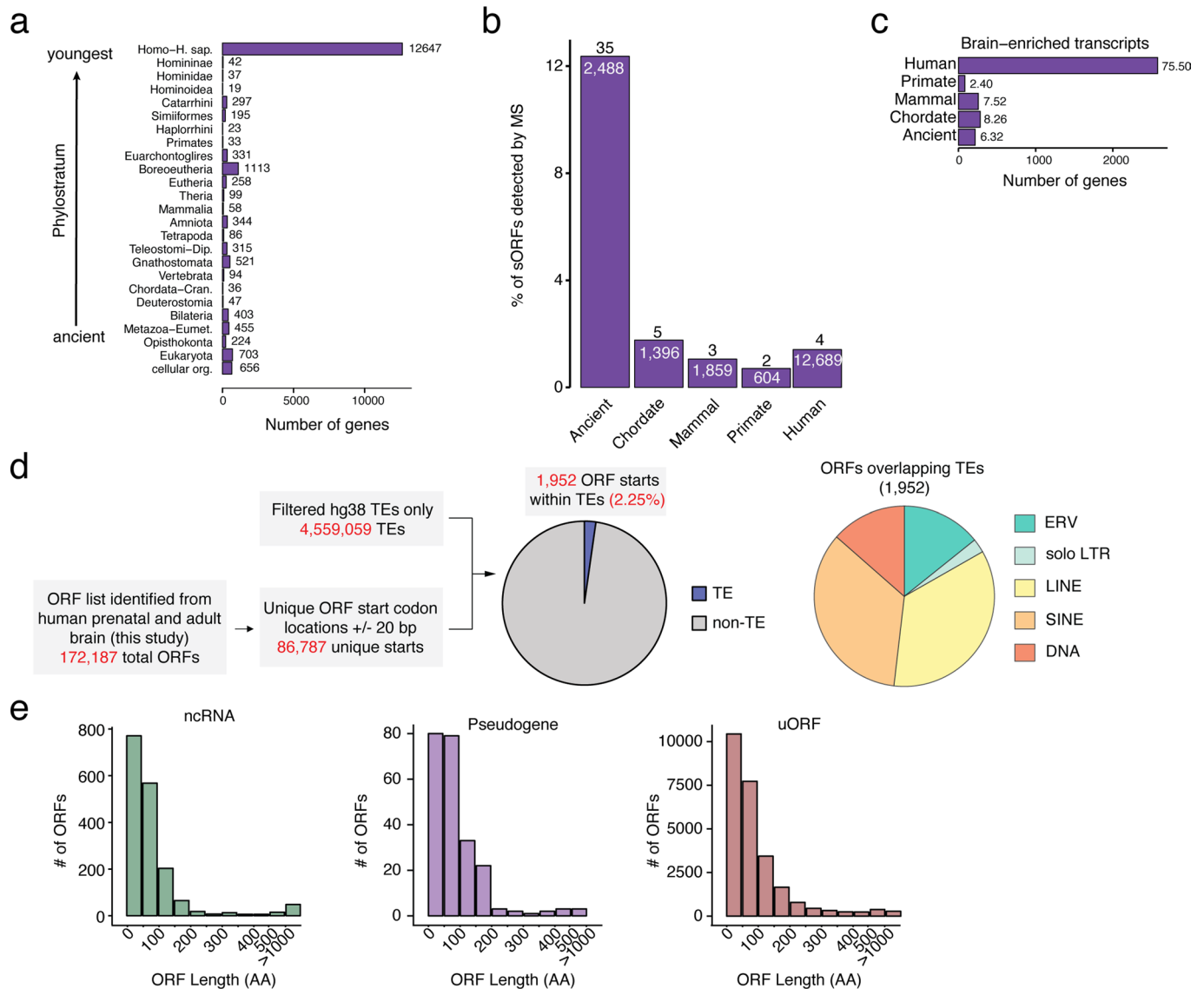
in the adult and prenatal brain, or by Johnson et al.²⁵. (e) Histograms of number of proteins identified by size-selection proteomics in the adult and prenatal brain, or by Johnson et al.²⁵, binned by protein length. (f) Box and whisker plots of Ribo-seq TPM for all ORFs detected by MS and ORFs not detected, data are shown as median \pm IQR (whiskers = 1.5*IQR), $n = 352$ (adult, Johnson et al.), 3331 (adult, this study), 419 (prenatal, this study), 168,085 (not detected by MS). (g) Box and whisker plots of Ribo-seq TPM for sORFs detected by MS and ORFs not detected. (f-g) data are shown as median \pm IQR (whiskers = 1.5*IQR), $n = 14$ (adult, Johnson et al.), 17 (adult, this study), 31 (prenatal, this study), 16838,125085 (not detected by MS). *** $p = 2.92 \times 10^{-5}$, **** $p < 2.2 \times 10^{-16}$, by two-sided Kolmogorov-Smirnov test. While only a fraction of the sORFs identified by ribosome profiling were detected by our mass spectrometry analysis, this is not surprising given that such shotgun proteomic approaches have low sensitivity for the detection of individual proteins, particularly if the proteins are transient or low in abundance. Consistent with this finding, the sORF-encoded proteins that we were able to detect by proteomics exhibited a higher average ribosome density compared to all sORFs detected by ribosome profiling.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Activity-dependent translation in hESC-derived neurons, Related to Fig. 4. (a) Heatmap of RNA-seq TPM from hESC-derived neuronal cultures for marker genes associated with neuronal and non-neuronal cell types. Rows indicate individual samples and biological replicates. The pattern of gene expression observed in these cultures largely mimics the findings in Nehme et al.²⁹. (b) Heatmap of pairwise Spearman's r correlation between the top 2,000 expressed genes in RNA-seq samples from NGN2-derived neurons, including cells that were treated with cycloheximide alone (no treatment), harringtonine, or a vehicle control (DMSO). (c) Heatmap of pairwise Spearman's r correlation between the top 2,000 expressed genes in Ribo-seq samples from NGN2-derived neurons, including cells that were treated with cycloheximide alone (no treatment), harringtonine, or a vehicle control (DMSO). (d) Bar chart of normalized RNA-seq and Ribo-seq reads mapping to *FOS*, a classic activity-induced gene in neurons, and *XIRP1*, an activity-induced gene that was previously reported in human GABAergic neuron cultures⁷¹. Data are shown as mean \pm SD, $n = 3$ independent cell differentiations. (e) Bar plot of ORF types detected by Ribo-seq in NGN2 neurons, separated based on whether or not the exact ORF was also detected in postmortem adult or prenatal brain tissue. Most canonical ORFs detected in NGN2 neurons are also detected in human postmortem brain tissue. Likewise, we detect substantial, albeit somewhat lower, overlap for a range of non-canonical ORFs across both sample types as well. We are not surprised at the extent of the overlap between these datasets, as NGN2 neuronal cultures largely mimic embryonic cortical excitatory neurons; while cortical excitatory neurons represent a major constituent cell type in the dorsolateral prefrontal cortex, many other cell types exist in this tissue, and a number of the candidate sORFs/noncanonical ORFs (for example PLN, SLN, DWORF) and ncRNAs exhibit cell type specificity in the RNAs from which they are translated. Therefore, it is expected that we would not detect all of the same ORFs between NGN2 neurons and human postmortem tissue. (f) Importantly, the nature of the overlapping ORF population does not simply reflect the likelihood of ORF detection in brain samples. In this regard, we categorized ORFs detected in adult and/or prenatal postmortem tissue based on the number of samples in which the identical ORF was detected, and binned ORFs in each category by whether they were also detected in NGN2 neurons. Stacked bar chart of ORF types detected by Ribo-seq in human postmortem tissue, separated based on the number of tissue samples

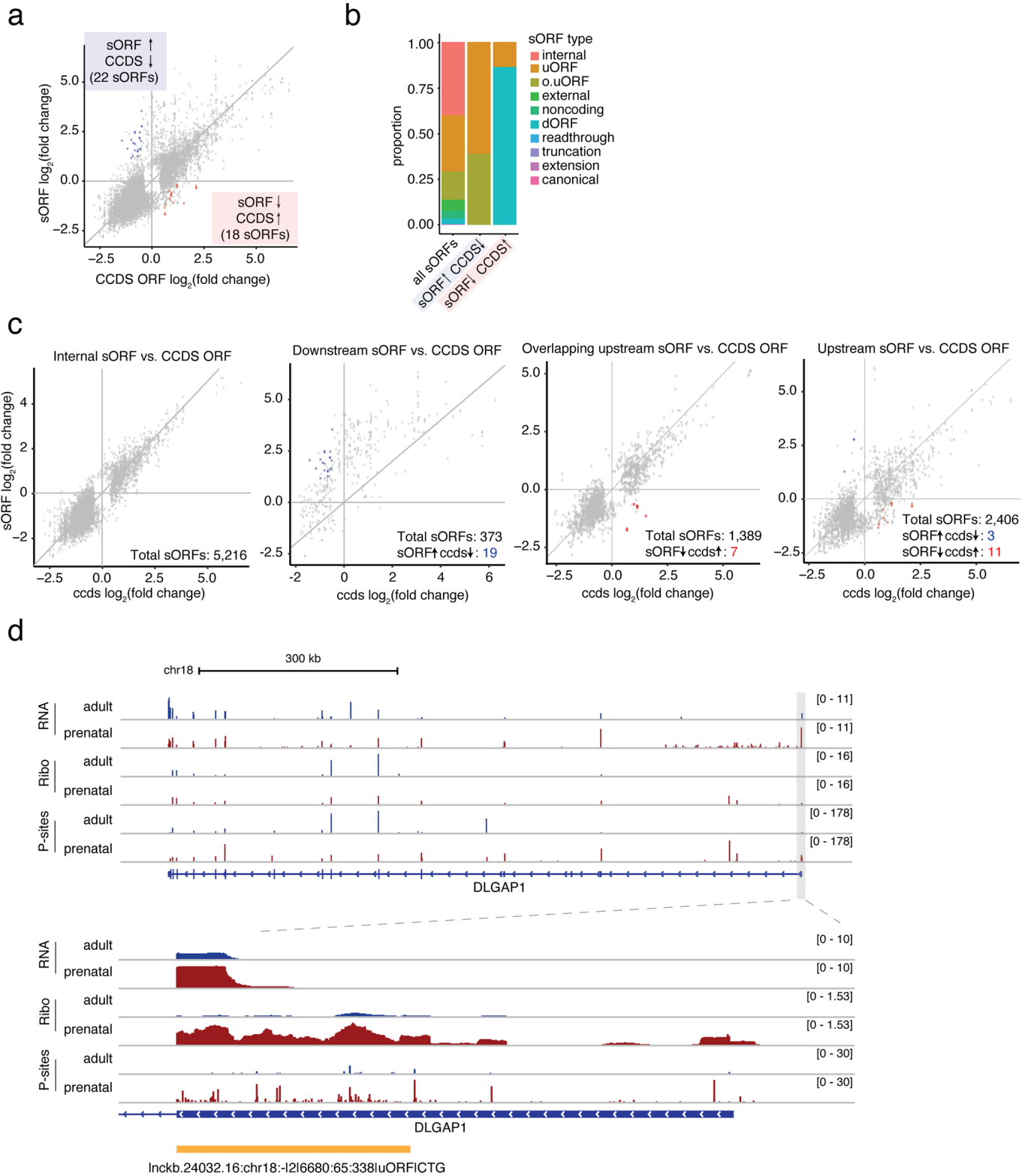
in which a given ORF was detected (2, 3-5, 6-10, 11-20, 21-50, 50-73) and whether the ORF was also detected in NGN2 neurons. This analysis did not find a notable difference in the proportion of canonical or various non-canonical ORFs based on whether they were detected in NGN2 neurons. (g) Stacked bar plot of start codon usage by ORF type in NGN2 neurons. (h) Metagene plot of Ribo-seq P-sites across the gene body of annotated Refseq ORFs in 3 independent replicates of harringtonine-treated (+ Harr) NGN2 neurons or vehicle-treated controls (- Harr). The enrichment for P-sites near start codons in + Harr samples is a hallmark of translational stalling induced by harringtonine treatment. (i) Number of ORFs in which the start codon was confirmed in harringtonine-treated NGN2-treated neurons. Two independent computational pipelines were used, ORF-RATER⁶⁰ and RiboTISH³¹. While ORF-RATER validates fewer ORFs compared to RiboTISH, this result is expected as ORF-RATER utilizes a different, more stringent filtering process, consistent with our finding that nearly all ORFs validated by ORF-RATER are also validated by RiboTISH. A full list of ORFs validated by ORF-RATER and RiboTISH can be found in Supplementary Table 3. (j&k) Activity-dependent changes in ORF translation were largely driven by transcriptional changes rather than a shift in ribosome density for both canonical ORFs and sORFs. This finding is consistent with observations that activity-dependent translation events coupled to transcription are transient and likely return to basal levels within six hours of membrane depolarization. (l) Scatterplot of fold-changes between stimulated and unstimulated neurons for all canonical ORFs in Ribo-seq data and the corresponding gene in RNA-seq data. Transcriptionally regulated genes (blue), translationally regulated genes (red), buffered genes (light purple), and intensified genes (dark purple) are highlighted. (m) Scatterplot of fold-changes between stimulated and unstimulated neurons for all sORFs in Ribo-seq data and the corresponding gene in RNA-seq data. Transcriptionally regulated genes (blue), translationally regulated genes (red), buffered genes (light purple), and intensified genes (dark purple) are highlighted. (n&o) Genomic loci of two activity-dependent ncRNAs with evidence of translation, *MIR22HG* (n) and *LOC107986102* (o). Tracks represent merged and depth-normalized reads across 3 biological replicates of membrane-depolarized (6 h KCl) and unstimulated neurons for RNA-seq, Ribo-seq, as well as P-site positions for Ribo-seq and harringtonine-treated Ribo-seq. sORFs identified by RibORF are shown in gold.



Extended Data Fig. 5 | Evolutionary origins of human brain sORFs,

Related to Fig. 5. (a) Bar plot of the number of sORFs (40-110 AA) grouped by evolutionary age. (b) Bar plot of the number of sORFs (40-100 AA) detected by mass spectrometry (See Fig. 3d) grouped by evolutionary age. We detected many evolutionarily conserved ORFs that were previously detected in other species. For example, uORFs translated from the 5'UTRs of *FTH1* and *CCN1* show conservation within the mammalian lineage by phylostratigraphy and show

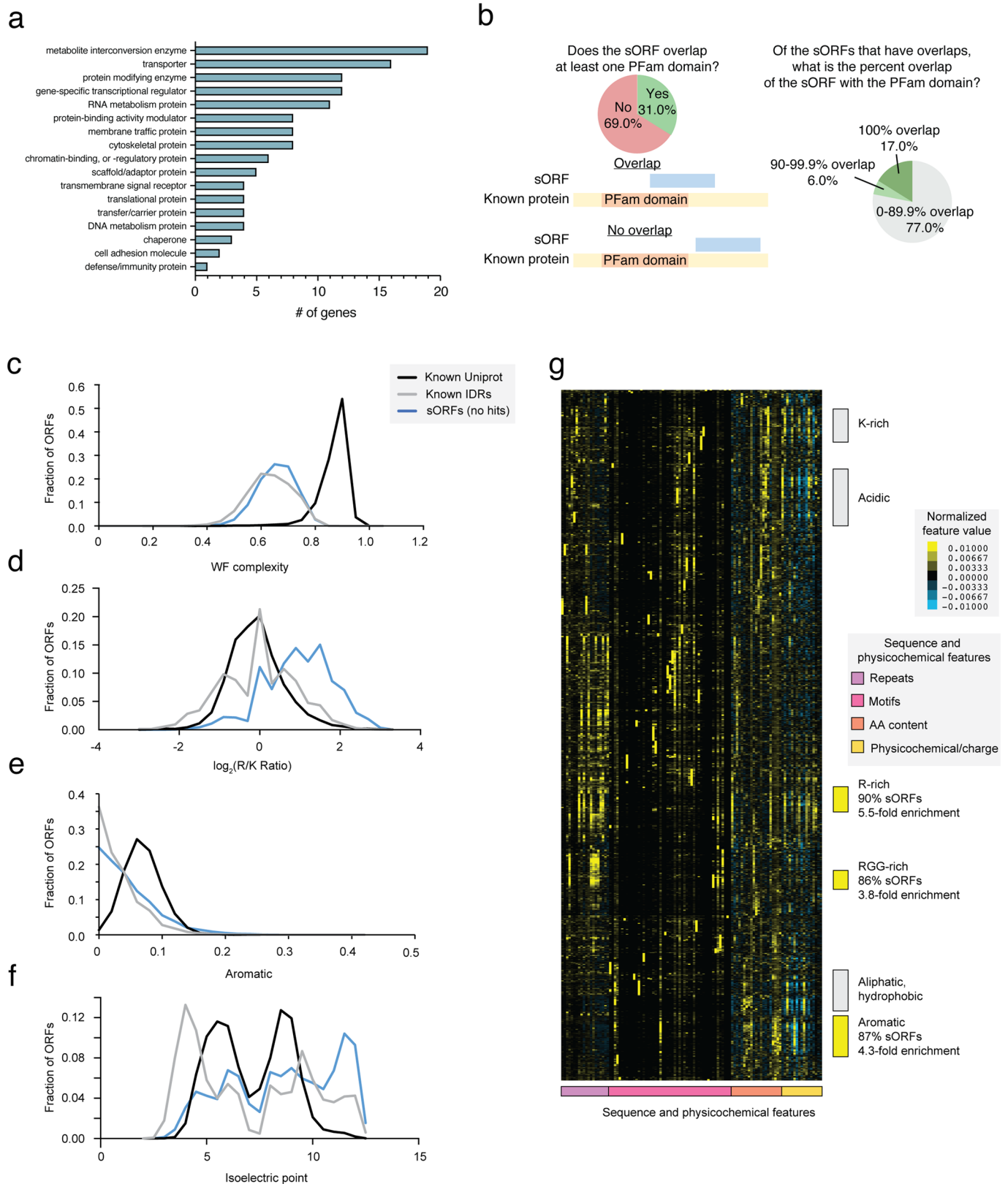
experimental evidence of translation in rodents^{4,36}. (c) Number and percentage of sORFs ≥ 40 AA that are translated from brain-enriched transcripts, grouped by evolutionary age. (d) Criteria for filtering TE insertion events at start codons (left) and pie chart of TE type for all ORFs in our dataset with a TE insertion at the start codon. (e) Histogram of ORF length for all ORFs encoded within ncRNAs, pseudogenes, and uORFs. Most ORFs encoded by ncRNAs, pseudogenes, and uORFs are sORFs.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Effects of uORF expression on downstream ORF translation, Related to Fig. 6. (a) Scatterplot of fold-changes in translation between adult and prenatal brain for sORFs and canonical ORFs expressed from the same transcript. Positive values indicate enrichment in the adult brain, whereas negative values indicate enrichment in the prenatal brain. Red points indicate genes where sORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in prenatal brain while canonical ORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in adult brain. Blue points indicate genes where sORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in adult brain while canonical ORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in prenatal brain. While most sORFs exhibited concordant translation with their associated canonical ORFs across development, we identified 50 sORFs that were discordant with nearby canonical ORF translation, and these discordant sORFs were strongly enriched for uORFs translated from 5'UTRs of annotated protein-coding genes. (b) Stacked bar plot of numbers and percentages of sORFs

detected in human brain (all sORFs), or sORFs exhibiting oppositely regulated expression across development compared to a canonical ORF translated from the same gene, separated by sORF type. (c) Scatterplot of fold-changes in translation between adult and prenatal brain for sORFs and canonical ORFs expressed from the same transcript, separated by type of ORF. Positive values indicate enrichment in the adult brain, whereas negative values indicate enrichment in the prenatal brain. Red points indicate genes where sORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in prenatal brain whereas canonical ORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in adult brain. Blue points indicate genes where sORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in adult brain whereas canonical ORF translation is significantly ($\text{DESeq2 } p_{\text{adj}} < 0.05$) enriched in prenatal brain. (d) Genomic locus of *DLGAP1*. Tracks represent merged and depth-normalized reads across all adult vs. prenatal samples for RNA-seq, Ribo-seq, as well as P-site positions. The sORF identified by RibORF is shown in gold.



Extended Data Fig. 7 | Microprotein functional characterization, Related to Fig. 7. (a) Protein functions of known genes that share homology with sORFs. **(b)** Characterization of sORFs that overlap with a domain in the PFam database. **(c)** Line plot of sequence complexity in the known human proteome (Known Uniprot), the disordered proteome (Known IDRs), and sORFs detected in human postmortem tissue that do not share homology with known protein domains. **(d)** Line plot of the \log_2 ratio of arginine to lysine in the categories of ORFs described in c. **(e)** Line plot of the proportion of aromatic amino acids (phenylalanine,

tyrosine, and tryptophan) in the categories of ORFs described in c. **(f)** Line plot of the isoelectric point for the categories of ORFs described in c. **(g)** Heatmap and hierarchical clustering of z-scores for 109 sequence and physicochemical features associated with the known disordered proteome as well as all sORFs that do not contain a BlastP hit and do not overlap with annotated ORFs. Boxes to the right of the heatmap indicate clusters of IDRs with similar properties. Yellow = clusters significantly enriched for sORFs.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used
Data analysis	<p>GraphPadPrism V8 (GraphPad Software) ImageJ Software version 2.1.0 R version 4.0.1 RibORF version 1.0 Cutadapt version 1.14 Bowtie2 version 2.3.4.3 Hisat2 version 2.1.0 STAR version 2.7.3a Picard Tools version 2.8.0 Bedtools version 2.27.1 Samtools version 1.9 HTSeq version 0.9.1 DTEG.R commit 60b61ca (https://github.com/SGDDNB/translational_regulation) edgeR version 3.26.8 gProfiler2 version 0.2.0 Blastp version 2.6.0 treeview version 1.1.6r4 MaxQuant version 1.6.17.0 ORF-RATER commit 5ae4cd6 (https://github.com/alexfields/ORF-RATER) Ribo-TISH version 0.2.6 (https://github.com/zhpn1024/ribotish) GORilla (no version info available) (http://cbl-gorilla.cs.technion.ac.il/)</p>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Human brain primary tissue RNA-seq and Ribo-seq data have been submitted to the database of Genotypes and Phenotypes (dbGaP) under accession number phs002489. NGN2 RNA-seq and Ribo-seq data have been submitted to the Gene Expression Omnibus (GEO) under accession number GSE180240.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size. Sample sizes were determined based on similar, previously conducted studies (main text ref. 4) and also magnitude and consistency of measurable differences.
Data exclusions	For ribosome profiling, samples were excluded if they exhibited poor three-nucleotide periodicity. RNA-seq samples were excluded if Ribo-seq from the same tissue was excluded. Data exclusion criteria were pre-established prior to complete data analysis.
Replication	RNA-seq and Ribo-seq experiments were reproduced 3 times with independent differentiation cohorts. Mass spectrometry experiments from adult (n=4) and prenatal (n=3) tissue were reproduced at least 3 times. Western blots and immunostaining were reproduced twice. Replication of studies statistically supported our conclusion in all experiments.
Randomization	For human postmortem tissue experiments, sample selection was dictated by tissue availability; all tissue was collected from individuals absent of neurological disease (described in more detail below in Human research participants - population characteristics (below))
Blinding	No blinding was performed as the identity of the samples was central to the analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Antibodies for immunostaining and imaging:
Mouse monoclonal anti-FLAG (M2) Sigma Aldrich Cat#F1804 (1:1000)

Donkey anti-mouse 488 secondary antibody (Thermo Fisher Scientific, A31571, 1:2000)

Antibodies for immunoblotting:

Rabbit monoclonal anti-HA Cell Signaling Technology Cat#3724S (1:1000)

Rabbit polyclonal anti-GAPDH Sigma Aldrich Cat#G9545 (1:1000)

Dylight 800nm conjugated secondary antibodies (Rockland Immunochemical 75934-472) (1:5,000)

Validation

Antibodies for immunostaining and imaging:

Mouse monoclonal anti-FLAG (M2) Sigma Aldrich Cat#F1804 (1:1000): validated by the manufacturer for the application and species. See manufacturer's website for validation statement (https://www.sigmaaldrich.com/specification-sheets/469/360/F1804-5MG_____SIGMA____.pdf)

Donkey anti-mouse 488 secondary antibody (Thermo Fisher Scientific, R37114, 1:2000): validated by the manufacturer for the application and species. See manufacturer's website for validation statement (https://www.thermofisher.com/order/genome-database/dataSheetPdf?producttype=antibody&productsubtype=antibody_secondary&productId=R37114&version=243)

Antibodies for immunoblotting:

Rabbit monoclonal anti-HA (C29F4) Cell Signaling Technology Cat#3724S (1:1000): validated by the manufacturer for the application and species. See manufacturer's website for validation statement (<https://www.cellsignal.com/products/primary-antibodies/ha-tag-c29f4-rabbit-mab/3724>)

Rabbit polyclonal anti-GAPDH Sigma Aldrich Cat#G9545 (1:1000): validated by the manufacturer for the application and species. See manufacturer's website for validation statement (<https://www.sigmaaldrich.com/deepweb/assets/sigmaaldrich/product/documents/778/720/g9545dat.pdf>)

Dylight 800nm conjugated secondary antibodies (Rockland Immunochemical 611-145-122) (1:5000): validated by the manufacturer for the application and species. See manufacturer's website for validation statement (<https://www.rockland.com/categories/secondary-antibodies/rabbit-igg-hl-antibody-dylight-800-conjugated-pre-adsorbed-611-145-122/GetProductDataSheet/?code=611-145-122>)

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

The H9 NGN2 hESC line was obtained as a gift from the laboratory of Kevin Eggan and validated in main text ref. 54. LentiX 293T cells were purchased from Takara (Cat. num. 632180).

Authentication

We did not independently authenticate the NGN2 cell line as it was a gift from the Eggan lab. RNA-seq expression patterns of NGN2 neurons largely matched those published in main text ref. 54. LentiX-293T cells were validated by morphology.

Mycoplasma contamination

Manufacturer validation: Each production lot is tested for sterility and mycoplasma. In-study validation: Cultures were tested with LookOut Mycoplasma PCR Detection Kit, Sigma MP0035, and always tested negative (no PCR product for mycoplasma DNA).

Commonly misidentified lines (See [ICLAC](#) register)

No commonly misidentified cell lines were used in the study.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

The gestational age of prenatal cortex samples (30 total) ranged from 12 to 23 weeks, while adult brain donors (43 total) ranged in age from 18 to 82 years, balancing male and female samples in both cohorts (prenatal: 14F, 16M; adult: 18F, 25M). Adult brain samples with a post-mortem interval <15 hours were included in the final cohort. All cases were determined by chromosomal analysis, physical examination, and/or pathological analysis to be control tissues, which indicates that they were absent of neurological disease. Cases with any abnormalities in these parameters were not used for this manuscript.

Recruitment

De-identified tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations.

Ethics oversight

All human tissue research was approved by the Harvard Medical School Institutional Review Board. De-identified adult brain tissue samples were obtained from the National Institute of Health (NIH) NeuroBioBank. NIH NeuroBioBank sample collection has been approved by the following Institutional Review Boards: University of Miami Institutional Review Board, The University of Maryland Institutional Review Board, The Maryland Department of Health and Mental Hygiene IRB, Partners Human Research Committee, Department of Veterans Affairs - Los Angeles, Bronx VA Medical Center Institutional Review Board, and University of Pittsburgh Institutional Review Board. Prenatal brain samples were obtained from the Human Developmental Biology Resource (HDBR) and the University of California San Francisco (UCSF) Pediatric Neuropathology Laboratory. De-identified tissue samples were collected with previous patient consent in strict observance of the legal and institutional ethical regulations. Patients were consented specifically for research purposes and were not compensated for their tissue donation. Tissue collection protocols were approved by the Human Gamete, Embryo, and Stem Cell Research Committee (institutional review board) at the University of California, San Francisco, and the Newcastle & North Tyneside 1 Research Ethics Committee (for HDBR).

Note that full information on the approval of the study protocol must also be provided in the manuscript.