Title
Developmental dynamics of RNA translation in the human brain

Journal
Nature Neuroscience, 25(10)

ISSN
1097-6256

Authors
Duffy, Erin E
Finander, Benjamin
Choi, GiHun
et al.

Publication Date
2022-10-01

DOI
10.1038/s41593-022-01164-9

Peer reviewed

HHS Public Access

Author manuscript

Nat Neurosci. Author manuscript; available in PMC 2023 May 19.

Published in final edited form as:

Nat Neurosci. 2022 October ; 25(10): 1353–1365. doi:10.1038/s41593-022-01164-9.

Developmental Dynamics of RNA Translation in the Human Brain

Erin E. Duffy[1,*], **Benjamin Finander**[1], **GiHun Choi**[1], **Ava C. Carter**[1], **Iva Pritisanac**[2,3,4,5], **Aqsa Alam**[2], **Victor Luria**[6,7,8], **Amir Karger**[9], **William Phu**[8,10,11], **Maxwell A. Sherman**[12], **Elena G. Assad**[1], **Naomi Pajarillo**[1], **Alexandra Khitun**[13], **Elizabeth E. Crouch**[14,15], **Sanika Ganesh**[1], **Jin Chen**[16], **Bonnie Berger**[12], **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. Moses**[2], **Brian T. Kalish**[1,20,21,*], **Michael E. Greenberg**[1,*]

[1.]Department of Neurobiology, Harvard Medical School; Boston, MA USA.

[2.]Department of Cell and Systems Biology, University of Toronto; Toronto, ON, Canada.

[3.]Program in Molecular Medicine, The Hospital for Sick Children; Toronto, ON, Canada.

[4.]Department of Biochemistry, University of Toronto; Toronto, ON, Canada.

[5.]Gottfried Schatz Research Center for Cell Signaling, Metabolism and Aging, Molecular Biology and Biochemistry, Medical University of Graz, 8010 Graz, Austria.

[6.]Department of Neuroscience, Yale School of Medicine; New Haven, CT USA.

[7.]Department of Systems Biology, Harvard Medical School; Boston, MA USA.

[8.]Department of Pediatrics, Division of Genetics and Genomics; Boston Children's Hospital, Boston, MA USA.

[9.]IT-Research Computing, Harvard Medical School; Boston, MA USA.

[10.]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard; Cambridge, MA USA.

[11.]Analytic and Translational Genetics Unit, Massachusetts General Hospital; Boston, MA USA.

[12.]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology; Cambridge, MA USA.

[13.]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School; Boston, MA, USA.

[*]Corresponding authors. Erin_DuffyLacy@hms.harvard.edu, Brian.Kalish@sickkids.ca, Michael_Greenberg@hms.harvard.edu.

AUTHOR CONTRIBUTIONS

EED, BTK, and MEG conceptualized the study and designed the experiments. EED and BTK performed ribosome profiling and RNA-seq. EED, BTK, and BF analyzed ribosome profiling and RNA-seq. EED and JC analyzed harringtonine-treated Ribo-seq datasets. GC prepared human embryonic stem cell-derived neurons for ribosome profiling. ACC analyzed the insertion of TEs at ORF translation start sites. IP, AA, JF-K and AMM performed physiochemical analysis. VL, AODL, A Karger, WP, and NS performed phylostratigraphy analysis. BTK prepared samples for proteomics, and BTK, BF, and A Khitun analyzed proteomics data. EED, EGA, and NP performed microprotein validation experiments in 293T cells. MS and BB performed disease heritability analysis. EEC and EJH provided prenatal brain tissue samples and technical advice for tissue processing. SG assisted with Ribo-seq quality control analyses. EED, BTK, BF, ECG and MEG drafted the manuscript, with input from all co-authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

[14.]Department of Pediatrics, University of California San Francisco; San Francisco, CA, USA.

[15.]The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research, University of California San Francisco; San Francisco, CA, USA.

[16.]Cecil H. and Ida Green Center for Reproductive Biology Sciences, UT Southwestern Medical Center; Dallas, TX USA.

[17.]Department of Pathology, University of California San Francisco; San Francisco, CA USA.

[18.]Weill Institute for Neurosciences, University of California San Francisco; San Francisco, CA USA.

[19.]Pathology Service 113B, San Francisco Veterans Affairs Healthcare System; San Francisco, CA USA.

[20.]Department of Paediatrics, Division of Neonatology, Hospital for Sick Children; Toronto, ON, Canada.

[21.]Program in Neuroscience and Mental Health, SickKids Research Institute; Toronto, ON, Canada.

Abstract

The precise regulation of gene expression is fundamental to neurodevelopment, plasticity, and cognitive function. While several studies have profiled transcription in the developing human brain, there is a gap in our understanding of accompanying translational regulation. 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- 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 tasks.[1] Proteomic diversity is generated through a series of transcriptional, post-transcriptional, and translational mechanisms that ultimately contribute to a rich and complex ‘translatome’. While 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 resolution.[2] From these data, the movement of ribosomes across codons can be determined and then used to identify

protein-coding open reading frames (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-seq analysis fails to give a complete picture of the landscape of proteins produced by a cell. Indeed, ribosomal profiling studies in yeast,[3] as well as cardiac[4] and tumor tissues,[5] have revealed the widespread active translation of previously unknown small ORFs (sORFs) encoding microproteins ≤100 amino acids. From the relatively few microproteins to be functionally studied, researchers have identified important regulators of mitochondrial metabolism, 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 find that the majority of sORFs in the brain are newly evolved in humans, where a subset of the sORFs arose via transposable element insertion at start codons. While their recent evolution might be thought to suggest that the microproteins translated from these sORFs are non-functional, >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 type.[9] Our study thus significantly 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

Translation Landscape of the Human Prenatal and Adult Brain

To characterize the human brain translational landscape at single-nucleotide resolution, we performed simultaneous RNA sequencing (RNA-seq) and ribosome profiling (Ribo-seq) from human adult dorsolateral prefrontal cortex and prenatal cortex across a range of ages (Figure 1a, Supplementary Figure 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 to 23 weeks, while adult brain donors (43 total) ranged in age from 18 to 82 years, with an average post-mortem interval of 9.9 hours (Figure 1b). Importantly, across samples, Ribo-seq data exhibited the three-nucleotide periodicity characteristic of actively translating ribosomes, a key metric for confident ORF identification (Figure 1c). Moreover, Ribo-seq reads exhibited expected fragment size distributions (Supplementary Figure 1b) and mapped primarily to annotated gene coding regions (Supplementary Figure 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 Supplementary Figure 1a-j.

High-confidence bona fide ORFs were identified based on the characteristic triplet reading frame periodicity of ribosome footprints using RibORF.[10] 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 (Figure 1d-e, Supplementary Figure 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 tissues[13] (Figure 1e, f, Supplementary Figure 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 lincRNAs or pseudogenes (Figure 1g), including the recently characterized ncRNA-encoded microproteins NoBody,[8] MOXI,[7] and Cyren.[14] 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

While 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 efficiency,[17] focusing on canonical ORFs that encode proteins of known function. Comparison of our paired transcriptome and translatome datasets revealed several distinct modes of developmental regulation (Figure 2a-b, Supplementary Figure 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 (Figure 2c, Supplementary Figure 2a-b). This coordinated translational regulation likely reflects developmental changes in mTOR signaling, as these transcripts contain 5' terminal oligopyrimidine tract (5'-TOP) motifs,[18] sequences at the 5′ ends of mRNAs that link their translation to the mTOR nutrient-sensing signaling pathway.[19] 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 (Figure 2c, Supplementary Figure 2a-b). 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 (e.g. pseudogenes, lincRNAs, antisense RNAs, and 5' and 3' untranslated regions 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 AAs) 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 (Figure 3a, Supplementary Figure 3a-b). While 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 (Figure 3b, Supplementary Figure 3c). Importantly, while the ribosome density of sORFs was on average ~10-fold lower than the translation of canonical ORFs (Supplementary Figure 3a), this low level of translation is similar to that of a number of previously reported microprotein-encoding RNAs with well-characterized functions, including RPL41,[22] SLN,[23] and NBDY,[8] 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 (Figure 3c).

While recognizing the difficulties associated with proteomic microprotein detection, we sought to independently corroborate our Ribo-seq findings at the protein level. Towards this end, we performed size-selected mass spectrometry-based proteomics for enhanced detection of protein species less than 20 kDa.[24] 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 microproteins.[25]

Collectively, these analyses identified peptides corresponding to 4,104 unique ORFs (Figure 3d, Supplementary Figure 3d-g), including 199 sORFs, 39 uORFs, and 4 noncoding 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 (Figure 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 stress,[28] 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 mass 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 significant expansion of the known brain translatome, 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 protocol[29] in which doxycycline induction of NGN2 was combined with SMAD and WNT inhibition to induce patterning toward a forebrain phenotype (Figure 4a). The resulting cultures (hereafter NGN2 neurons) have transcriptional signatures that are similar to those of well-differentiated glutamatergic neurons (Supplementary Figure 4a).

For these studies, we also exposed the differentiated NGN2 neurons to elevated levels of KCl, a treatment that is known to induce acute, synchronous membrane depolarization and to promote activity-dependent changes in RNA transcription in these neurons.[30] Thus, day 28 cultures from three independent differentiation cohorts were harvested for combined RNA-seq and ribosome profiling either prior to or 6 h following membrane depolarization with 55 mM potassium chloride (KCl). The resulting datasets passed key quality control metrics, with clear three-nucleotide periodicity observed in the Ribo-seq data (Supplementary Figure 4h) and high data correlation between separate differentiation cohorts (Supplementary Figure 4b-c). Moreover, robust induction of known activity-responsive loci was observed in all depolarized samples (Supplementary Figure 4d). Collectively, this analysis identified a total of 124,613 actively translated ORFs in NGN2 neurons (Figure 4b-c, Supplementary Figure 4g), >60% of which (78,965/124,613) were also observed in the human brain tissue samples (Supplementary Figure 4e-f). Principal component analysis (PCA) plots showed that NGN2 samples cluster more with the fetal rather than the adult translatome (Figure 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 sites.[11] The use of harringtonine is a key validation of Ribo-seq experiments not available in post-mortem 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 ORFs.[31] Compared to vehicle-treated control neurons, we observed the expected accrual of ribosome footprints

at translational initiation sites in harringtonine-treated samples (Figure 4i, Supplementary Figure 4h). We validated the start codons of 61,400 ORFs, including 8,881 sORFs, 5,258 uORFs, and 2,005 ORFs translated from ncRNAs (Figure 4J, Supplementary Figure 4i, Supplementary Table 3). Canonical ORFs showed the highest concordance (74.4%) between the harringtonine experiments and the original dataset analyzed using RibORF. Importantly, we also confirmed the start codons of a number of non-canonical ORFs using harringtonine treatment, including 43.1% of uORFs and 41.3% of noncoding ORFs. These ORFs represent the 'highest confidence' ORFs from the NGN2 dataset.

We next further characterized the novel sORFs found to be translated from previously annotated ncRNAs. In this regard, we observed active translation of novel sORFs in 128 out of 706 activity-dependent ncRNAs detected in NGN2 cultures (Figure 4e, Supplementary Figure 4j-m), many of which (101) were also detected by Ribo-seq in our human brain datasets. Notably, a number of the predicted protein products from these ncRNAs were verified biochemically using size-selected proteomics (Supplementary Table 2). Among these translated ncRNAs was *LINC00473* (Figure 4f-g), a previously characterized primate-specific and activity-dependent lincRNA[32] that has been implicated as a sex-specific driver of stress resilience when expressed ectopically in the mouse prefrontal cortex.[33] Thus, *LINC00473* 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 neurona 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 species.[34] Determination of the minimal evolutionary age for human brain sORFs revealed that, compared to most annotated protein-coding genes, a majority of sORFs are human-specific (67% of sORFs versus 12% of annotated protein-coding genes Figure 5a, Supplementary Figure 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 (Figure 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 (Supplementary Figure 5b). These features are consistent with the classic view that the more evolutionarily conserved regions of the genome are more likely to be translated,[5] 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 are 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 iPSC human cells, selecting for those sORFs that show a significant growth phenotype when knocked out (Mann-Whitney U p-value <0.05).[9] Of the 124 sORFs that satisfy these criteria, a striking 101 are human-specific (Chi-squared 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 (KS test, $p<2.2*10^{-16}$, Supplementary Figure 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 formation.[36] 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% vs. 4%, respectively, $p < 2.2*10^{-16}$ by two proportions z-test; Figure 5e, Supplementary Figure 5e-f). 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 (Figure 5f); however, the functional significance 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 transcripts,[40] 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 (Figure 6a-c, Supplementary Figure 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 protein[45] (Figure 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) (Figure 3d-e, Supplementary Figure 3d), and was associated with a reduction in translation of the canonical DLGAP1 protein (Figure 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 gene,[46] raising the possibility that this 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 (Supplementary Figure 7a, 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 (Supplementary Figure 7b).

For sORFs lacking sequence similarity to known or predicted human proteins (69%), calculated FoldIndex scores, a rough predictive measure of intrinsic disorder,[47] suggest that these protein products do not generally adopt stable three-dimensional conformations (Figure 7a), consistent with other sequence-based characteristics (Figure 7b, Supplementary Figure 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 clustering[51] (Figure 7c, Supplementary Figure 7i, Supplementary Table 4). Strikingly, this analysis identified a strong enrichment of brain sORFs (>5x expected, 712 total) in the resulting sequence clusters that were rich in arginine-glycine-glycine (RGG) motifs and/or aromatic residues, as well as, to a lesser extent, clusters rich in arginine residues (2.8x expected). Likewise, human brain sORFs were strongly overrepresented (>5x 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 survival,[9] further supporting a functional role for these microproteins in the brain.

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 condensates.[52] Indeed, several previously characterized proteins in RGG- 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 over-expressed six selected sORFs with their endogenous 5' UTR and a FLAG-HA epitope tag in heterologous cells (Supplementary Table 5). 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 (Figure 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 (Figure 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 translatome 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 translatome has remained largely uncharacterized.

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 open reading frames in the human brain, including thousands of novel microproteins. We identified in the brain a subset of ncRNAs, uORFs, and other 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 ribosome density 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 traits.[57] 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 the majority of sORFs are human-specific renders them interesting candidates in the study of uniquely human features of the brain. While 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 viability.[9] 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 post-mortem brain tissue, and post-mortem 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. While 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 post-mortem tissue predominantly reflect physiologically relevant translation events.

In conclusion, our study provides the first large-scale resource for the investigation of translation regulation in the human brain. Importantly, 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- and brain-specific microproteins.

METHODS

Human brain samples

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. Adult brain samples with a post-mortem interval <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 manuscript. 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).

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 15x in 400 μL ice cold lysis buffer: 20 mM Tris pH 7.4, 150 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT, 100 μg/mL cycloheximide (Sigma). The lysate was further sheared using a 26-gauge syringe. The lysate was clarified by centrifugation at 20,000 x g for 10 minutes at 4°C. Supernatant (10 μL) was removed, added to 300 μL 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 RNaseI per μg RNA) at room temperature for 45 minutes with gentle agitation.

After RNase digestion, 10 μL SuperasIN (Thermo) 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 11x34 mm centrifuge tubes (Beckman Coulter) and underlaid with 0.9 mL sucrose cushion. Samples were centrifuged in a TLS-55 rotor at 51,000 rpm for 2 hours at 4°C. The supernatant was discarded and the pellet was resuspended in 300 μL TRIzol. Ribosome-protected fragments were purified from TRIzol using the Zymo Direct-zol kit. RNA was precipitated by adding 38.5 μL RNase-free water, 1.5 μL Glycoblue, 10 μL 3 M sodium acetate pH 5.5, and 150 μL isopropanol to 100 μL eluted RNA. The mixture was incubated overnight at −20°C. Samples were centrifuged for 30 minutes at 20,000 x g at 4°C. The supernatant was discarded, and the RNA pellet was

resuspended in 5 μL 10 mM Tris pH 8. Five μL 2X denaturing sample loading buffer (980 μL formamide, 20 μL 500 mM EDTA, 300 μg bromophenol blue) was added to each sample, then the sample was denatured at 80°C for 90 seconds. Ribosome-protected fragments, along with control oligos,[58] were run on a 15% polyacrylamide gel at 200 V with 12 μL NEB miRNA marker. The gel was stained with SYBR gold in 1X TBE. Gel fragments between 17 and 34 nucleotides were excised and placed in a microfuge tube with 400 μL gel extraction buffer. Samples were frozen on dry ice for 30 minutes, then thawed overnight with gentle agitation.

After overnight gel extraction, 400 μL eluate was transferred to a new microfuge tube. The RNA was precipitated by adding 1.5 μL glycoblue and 500 μL isopropanol. After overnight incubation at −20°C, the sample was centrifuged at 20,000 x g for 30 minutes at 4°C. The supernatant was discarded, and precipitated RNA was resuspended in 4 μL 10mM Tris pH 8. Samples were then dephosphorylated using T4 PNK (4 μL RNA in 10 mM Tris pH 8, 0.5 μL T4 PNK enzyme, 0.5 μL T4 PNK buffer, and 0.5 μL Superasin) at 37°C for 1 hour. Samples were then subjected to SPRI clean up: 50 μL of sample in RNase-free water was added to 90 μL RNAclean beads and 270 μL isopropanol. After washing with 85% ethanol, beads were resuspended in 7 μL 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 1x50 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 (ESCRO) Committee. The transgenic H9 NGN2 hESC line was a generous gift from Alban Ordureau and J. Wade Harper. H9 NGN2 was generated by inserting dox-inducible NGN2 cassette into the *AAVS1* 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 (Supplementary Figure 6A).[29] On the day prior to cell harvest, neurons were silenced with TTX and APV, which antagonize sodium channels and NMDA receptors, respectively. 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, Life Technologies) until ready for differentiation. A published protocol that combines developmental patterning and NGN2 induction was adapted to differentiate H9 NGN2 into neurons.[29] At day 0, cells were treated with Accutase (StemPro Accutase, Life Technologies) and plated in single cells at 50,000 cells/cm^2 in mTeSR Plus media supplemented with 10 μM Y-27632 (STEMCELL Technologies) on tissue culture plates coated with 336.67 μg/mL 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, 1X MEM non-essential amino acids (MEM NEAA), 1X penicillin/streptomycin (pen/strep) and 1X 2-mercaptoethanol (all Gibco)) supplemented with 100 nM LDN193189, 2 μM XAV939 (STEMCELL Technologies), 10 μM SB431542 hydrate and 2 μg/mL doxycycline hyclate (Sigma). Day 2 media was 50%

KSR media/50% NIM media supplemented with LDN/XAV/SB and 2 μg/mL doxycycline. NIM media consisted of DMEM/F-12 medium, 1X GlutaMAX, 1X MEM NEAA, 1X pen/strep, 0.16% D-glucose (Sigma) and 1X N2 supplement-B (STEMCELL Technologies). Day 3 media was NIM media supplemented with 2 μg/mL doxycycline. At day 4, cells were treated with Accutase and plated in single cells at 40,000 cells/cm^2 in NB media (Neurobasal medium (without glutamine), 1X GlutaMAX, 1X MEM NEAA, 1X pen/strep and 1X N2 supplement-B) supplemented with 1X B27 without Vitamin A, 2-2.4 μg/mL mouse laminin (Gibco), 1 μM ascorbic acid, 2 μM dibutyryl cyclic-AMP (Sigma), 20 ng/mL brain-derived neurotrophic factor, 10 ng/mL glial-derived neurotrophic factor (rhBDNF and rhGDNF, Peprotech), 10 μM Y-27632 and 2 μg/mL doxycycline on the tissue culture plates coated with 336.67 μg/mL 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 doxycycline on the third day of the week (days 8, 15 and 22). Cells were silenced on day 27 with TTX and APV, which antagonize sodium channels and NMDA receptors, respectively. Cells were collected at day 28 in 1X PBS supplemented with 1X cycloheximide after being stimulated with 55 mM KCl for 0 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 μg/mL harringtonine or an equal volume of DMSO was added to cell culture media and incubated at 37°C for 2 min 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 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 1x75 nt reads. Samples were always processed in large batches of a maximum of 24 samples to minimize sample processing biases.

Analysis of RNA sequencing 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) annotation.[60] 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 *H. 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 transcriptome[60] 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 amino acids and translation probability >0.7. Only samples in which frame 0 periodicity was >50% for at least two read lengths and an overall 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 zero 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","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 only detected in 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. In order to be designated a sORF, an ORF had to be 100 amino acids 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 samples[62] 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 rerun using SCDC.

Differential expression and GO 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 2 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 deltaTE[17] in R 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. 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 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, upstream open reading frame sequences identified by RibORF were joined to produce one singular

sample-specific upstream open reading frame region. For each of the upstream open reading frame regions, raw counts were generated by quantifying total P-sites across each region, from which TPMs were calculated. These TPMs were compared to canonical open reading frame translational efficiency values to characterize the relationship between upstream open reading frame utilization and canonical open reading frame translational dynamics at the level of individual genes.

Protein Sequence Analysis by LC-MS/MS

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 mm^3 pieces. Gel pieces were then subjected to a modified in-gel trypsin digestion procedure. Gel pieces were washed and dehydrated with acetonitrile for 10 min, 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 modified sequencing-grade trypsin (Promega, Madison, WI) at 4°C. After 45 min, 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 hr). 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, 0.1% formic acid). A nano-scale reverse-phase HPLC capillary column was created by packing 2.6 μm C18 spherical silica beads into a fused silica capillary (100 μm inner diameter x ~30 cm length) with a flame-drawn tip. After equilibrating the column, each sample was loaded via a Famos auto sampler (LC Packings, San Francisco CA) onto the column. A gradient was formed, and peptides were eluted with increasing concentrations of solvent B (97.5% acetonitrile, 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, Waltham, MA). 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 2-level FDR control was used: peptide level and protein group level, both with a 1% FDR threshold. A posterior error probability calculation is performed based on a target-decoy search. Common mass spec contaminants were filtered out. Only peptides with a score >50 were considered for subsequent analysis.

Physiochemical 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.gencodegenes.org/human/release_29.html; downloaded on Aug 2, 2019). GO terms enriched in known proteins with significant sequence similarity to sORFs were determined using GORilla[61] and plotted using GraphPadPrism V8 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 (Supplementary Figure 4D). FoldIndex score [47] is defined as 2.785*hydropathy - abs(net charge) - 1.151. Physicochemical and sequence properties of sORFs and 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 and 100 amino acids. 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 3 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 v1.1.6r4 (ref. [66]) and enrichment analysis was performed by selecting the IDRs from each cluster and using the 6910 IDRs in the entire cluster analysis as the background set. These lists were entered into the GOrilla webserver.[64] 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 amino acids were analyzed using TimeTree[67] 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) NCBI database with a Blastp e-value threshold of 10e-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 Hominidae (91 – 6.6 Mya)), and Humans (phylostrata 30-31, including Homo 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 lab of Dr. Didier Trono.[68] First, we created a list of all start codons in different categories (protein-coding, ncRNA, uORF, sORF, 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, dsDNA 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 was 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) using Lipofectamine 3000 reagent (Invitrogen). After 24 h, cells were harvested and resuspended in RIPA buffer (Sigma) supplemented with protease inhibitor cocktail (Roche). Protein concentration was measured by Bradford assay, and 20 μg protein lysate was denatured at 95°C for 5 min and then separated on a 10-20% Tris-tricine gel (Invitrogen) at 125 V for 90 min. Proteins were transferred to a nitrocellulose membrane (Bio-Rad) at 115 V for 90 min, and the membrane was blocked with 5% non-fat dry milk in TBST for 1 h. Membranes were incubated with anti-HA (C29F4) (1:1000, CST) or anti-GAPDH (1:1000, Sigma Aldrich) primary antibody in 5% non-fat dry milk in TBST overnight at 4°C. Membranes were washed 4x in TBST at room temperature 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 h and transfected as described above. Cells were fixed with 4% paraformaldehyde for 30 min at room temperature and washed three times with ice-cold PBS. The cells were permeabilized and blocked for 1 h at room temperature using 5% donkey serum in PBST (1X PBS, 0.1% Triton X-100). Coverslips were incubated with anti-FLAG mouse primary antibody (1:1000) (Sigma-Aldrich) overnight at 4°C. Coverslips were washed 3x in PBST at room temperature then incubated with fluorescently-labeled secondary antibody (1:2000, Alexa Fluor 488

anti-mouse) for 1 h at room temperature. Coverslips were washed 3x 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 63x objective and analyzed using ImageJ Software (version 2.1.0).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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. EED was supported by the Damon Runyon Cancer Research Foundation (DRG-2397-20). ACC was supported by the Hanna H. Gray Fellowship through the Howard Hughes Medical Institute. VL was supported by a Boston Children's Hospital Career Development Award (to AODL), and VL and NS by a National Human Genomic Research Institute (NHGRI) R01HG010898. EJH was supported by NIH P01 NS083513. AODL and WP were supported by a Manton Center Endowed Scholar Award and the NHGRI U01HG008900. MAS was supported by a National Institute of Mental Health (NIMH) F31MH124393. AMM and JDF-K acknowledge funding from the Canadian Institutes of Health Research (CIHR PJT-148532). BTK was supported by a National Institute of Neurological Diseases and Stroke (NINDS) K08 NS112338. MEG was supported by funding from a 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 lab 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 Dr. Sarah 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 lab of Dr. Didier Trono for sharing the human transposable element annotation and to Dr. Wade 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.

DATA AVAILABILITY

Data reprocessed from Wang *et al.*[70] was accessed from ArrayExpress with accession number 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. NGN2 RNA-seq and Ribo-seq data have been submitted to the Gene Expression Omnibus (GEO) under accession number GSE180240. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD035950. Our web-based searchable database is available from http://greenberg.hms.harvard.edu/project/human-brain-orf-database/

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 is previously published and cited in the Methods section.

REFERENCES

1. Kim M-S et al. A draft map of the human proteome. Nature 509, 575–581 (2014). [PubMed: 24870542]

2. Ingolia NT, Ghaemmaghami S, Newman JRS & Weissman JS Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. Science 324, 218–223 (2009). [PubMed: 19213877]

3. Guydosh NR & Green R Dom34 rescues ribosomes in 3' untranslated regions. Cell 156, 950–962 (2014). [PubMed: 24581494]

4. van Heesch S et al. The Translational Landscape of the Human Heart. Cell 178, 242–260.e29 (2019). [PubMed: 31155234]

5. Prensner JR et al. Noncanonical open reading frames encode functional proteins essential for cancer cell survival. Nat Biotechnol (2021) doi:10.1038/s41587-020-00806-2.

6. Makarewich CA et al. The DWORF micropeptide enhances contractility and prevents heart failure in a mouse model of dilated cardiomyopathy. Elife 7, (2018).

7. Makarewich CA et al. MOXI Is a Mitochondrial Micropeptide That Enhances Fatty Acid β-Oxidation. Cell Rep 23, 3701–3709 (2018). [PubMed: 29949755]

8. D'Lima NG et al. A human microprotein that interacts with the mRNA decapping complex. Nat Chem Biol 13, 174–180 (2017). [PubMed: 27918561]

9. Chen J et al. Pervasive functional translation of noncanonical human open reading frames. Science 367, 1140–1146 (2020). [PubMed: 32139545]

10. Ji Z. RibORF: Identifying Genome-Wide Translated Open Reading Frames Using Ribosome Profiling. Curr Protoc Mol Biol 124, e67 (2018). [PubMed: 30178897]

11. Ingolia NT, Lareau LF & Weissman JS Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147, 789–802 (2011). [PubMed: 22056041]

12. Ji Z, Song R, Regev A & Struhl K Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. Elife 4, e08890 (2015). [PubMed: 26687005]

13. Kalish BT et al. Maternal immune activation in mice disrupts proteostasis in the fetal brain. Nat Neurosci (2020) doi:10.1038/s41593-020-00762-9.

14. Slavoff SA, Heo J, Budnik BA, Hanakahi LA & Saghatelian A A human short open reading frame (sORF)-encoded polypeptide that stimulates DNA end joining. J Biol Chem 289, 10950–10957 (2014). [PubMed: 24610814]

15. Miller JA et al. Transcriptional landscape of the prenatal human brain. Nature 508, 199–206 (2014). [PubMed: 24695229]

16. Kang HJ et al. Spatio-temporal transcriptome of the human brain. Nature 478, 483–489 (2011). [PubMed: 22031440]

17. 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). [PubMed: 31763789]

18. Levy S, Avni D, Hariharan N, Perry RP & Meyuhas O Oligopyrimidine tract at the 5' end of mammalian ribosomal protein mRNAs is required for their translational control. Proc Natl Acad Sci U S A 88, 3319–3323 (1991). [PubMed: 2014251]

19. Jeong H et al. Evolution of DNA methylation in the human brain. Nat Commun 12, 2021 (2021). [PubMed: 33795684]

20. Huh GS et al. Functional requirement for class I MHC in CNS development and plasticity. Science 290, 2155–2159 (2000). [PubMed: 11118151]

21. Stevens B et al. The classical complement cascade mediates CNS synapse elimination. Cell 131, 1164–1178 (2007). [PubMed: 18083105]

22. Klaudiny J, von der Kammer H & Scheit KH 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). [PubMed: 1326959]

23. 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). [PubMed: 9367679]

24. Khitun A & Slavoff SA Proteomic Detection and Validation of Translated Small Open Reading Frames. Curr Protoc Chem Biol 11, e77 (2019). [PubMed: 31750990]

25. Johnson ECB 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). [PubMed: 32284590]
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). [PubMed: 22924626]
27. Lander SS et al. Glutamate Dehydrogenase-Deficient Mice Display Schizophrenia-Like Behavioral Abnormalities and CA1-Specific Hippocampal Dysfunction. Schizophr Bull 45, 127–137 (2019). [PubMed: 29471549]
28. Sinvani H et al. Translational tolerance of mitochondrial genes to metabolic energy stress involves TISU and eIF1-eIF4GI cooperation in start codon selection. Cell Metab 21, 479–492 (2015). [PubMed: 25738462]
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). [PubMed: 29791859]
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). [PubMed: 35649373]
31. Zhang P et al. Genome-wide identification and differential analysis of translational initiation. Nat Commun 8, 1749 (2017). [PubMed: 29170441]
32. Ataman B et al. Evolution of Osteocrin as an activity-regulated factor in the primate brain. Nature 539, 242–247 (2016). [PubMed: 27830782]
33. Issler O et al. Sex-Specific Role for the Long Non-coding RNA LINC00473 in Depression. Neuron 106, 912–926.e5 (2020). [PubMed: 32304628]
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). [PubMed: 18029048]
35. Calviello L et al. Detecting actively translated open reading frames in ribosome profiling data. Nat. Methods 13, 165–170 (2016). [PubMed: 26657557]
36. Playfoot CJ et al. Transposable elements and their KZFP controllers are drivers of transcriptional innovation in the developing human brain. Genome Res. 31, 1531–1545 (2021). [PubMed: 34400477]
37. Carlevaro-Fita J et al. Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. Genome Res 29, 208–222 (2019). [PubMed: 30587508]
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). [PubMed: 23637635]
39. Johnson R & Guigó R The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. RNA 20, 959–976 (2014). [PubMed: 24850885]
40. Uhlén M et al. Proteomics. Tissue-based map of the human proteome. Science 347, 1260419 (2015). [PubMed: 25613900]
41. Brar GA et al. High-resolution view of the yeast meiotic program revealed by ribosome profiling. Science 335, 552–557 (2012). [PubMed: 22194413]
42. Chew G-L, Pauli A & Schier AF Conservation of uORF repressiveness and sequence features in mouse, human and zebrafish. Nat Commun 7, 11663 (2016). [PubMed: 27216465]
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). [PubMed: 33826921]
44. Aspden JL et al. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq. Elife 3, e03528 (2014). [PubMed: 25144939]
45. Rasmussen AH, Rasmussen HB & Silahtaroglu A The DLGAP family: neuronal expression, function and role in brain disorders. Mol. Brain 10, 43 (2017). [PubMed: 28870203]
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). [PubMed: 27271353]

47. Prilusky J et al. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21, 3435–3438 (2005). [PubMed: 15955783]
48. Tsang B et al. Phosphoregulated FMRP phase separation models activity-dependent translation through bidirectional control of mRNA granule formation. Proc Natl Acad Sci U S A 116, 4218–4227 (2019). [PubMed: 30765518]
49. Gueroussov S et al. Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. Cell 170, 324–339.e23 (2017). [PubMed: 28709000]
50. Hnisz D, Shrinivas K, Young RA, Chakraborty AK & Sharp PA A Phase Separation Model for Transcriptional Control. Cell 169, 13–23 (2017). [PubMed: 28340338]
51. Zarin T et al. Proteome-wide signatures of function in highly diverged intrinsically disordered regions. Elife 8, (2019).
52. Chong PA, Vernon RM & Forman-Kay JD RGG/RG Motif Regions in RNA Binding and Phase Separation. J Mol Biol 430, 4650–4665 (2018). [PubMed: 29913160]
53. Jishi A, Qi X & Miranda HC Implications of mRNA translation dysregulation for neurological disorders. Semin. Cell Dev. Biol 114, 11–19 (2021). [PubMed: 34024497]
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). [PubMed: 30430754]
55. Kelleher RJ & Bear MF The autistic neuron: troubled translation? Cell 135, 401–406 (2008). [PubMed: 18984149]
56. Kapur M, Monaghan CE & Ackerman SL Regulation of mRNA translation in neurons-a matter of life and death. Neuron 96, 616–637 (2017). [PubMed: 29096076]
57. de Goede OM et al. Population-scale tissue transcriptomics maps long non-coding RNAs to complex disease. Cell 184, 2633–2648.e19 (2021). [PubMed: 33864768]
58. McGlincy NJ & Ingolia NT Transcriptome-wide measurement of translation by ribosome profiling. Methods 126, 112–129 (2017). [PubMed: 28579404]
59. 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). [PubMed: 32142685]
60. Seifuddin F et al. lncRNAKB, a knowledgebase of tissue-specific functional annotation and trait association of long noncoding RNA. Sci. Data 7, 326 (2020). [PubMed: 33020484]
61. Dong M et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform 22, 416–427 (2021). [PubMed: 31925417]
62. 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). [PubMed: 32341540]
63. Fields AP et al. A regression-based analysis of ribosome-profiling data reveals a conserved complexity to mammalian translation. Mol. Cell 60, 816–827 (2015). [PubMed: 26638175]
64. Eden E, Navon R, Steinfeld I, Lipson D & Yakhini Z GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10, 48 (2009). [PubMed: 19192299]
65. de Hoon MJL, Imoto S, Nolan J & Miyano S Open source clustering software. Bioinformatics 20, 1453–1454 (2004). [PubMed: 14871861]
66. Saldanha AJ Java Treeview—extensible visualization of microarray data. Bioinformatics 20, 3246–3248 (2004). [PubMed: 15180930]
67. Kumar S, Stecher G, Suleski M & Hedges SB TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol 34, 1812–1819 (2017). [PubMed: 28387841]
68. Turelli P et al. Primate-restricted KRAB zinc finger proteins and target retrotransposons control gene expression in human neurons. Sci. Adv 6, eaba3200 (2020). [PubMed: 32923624]
69. Hubley R et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44, D81–D89 (2016). [PubMed: 26612867]
70. Wang Z-Y et al. Transcriptome and translatome co-evolution in mammals. Nature 588, 642–647 (2020). [PubMed: 33177713]

71. Boulting GL 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). [PubMed: 33542524]

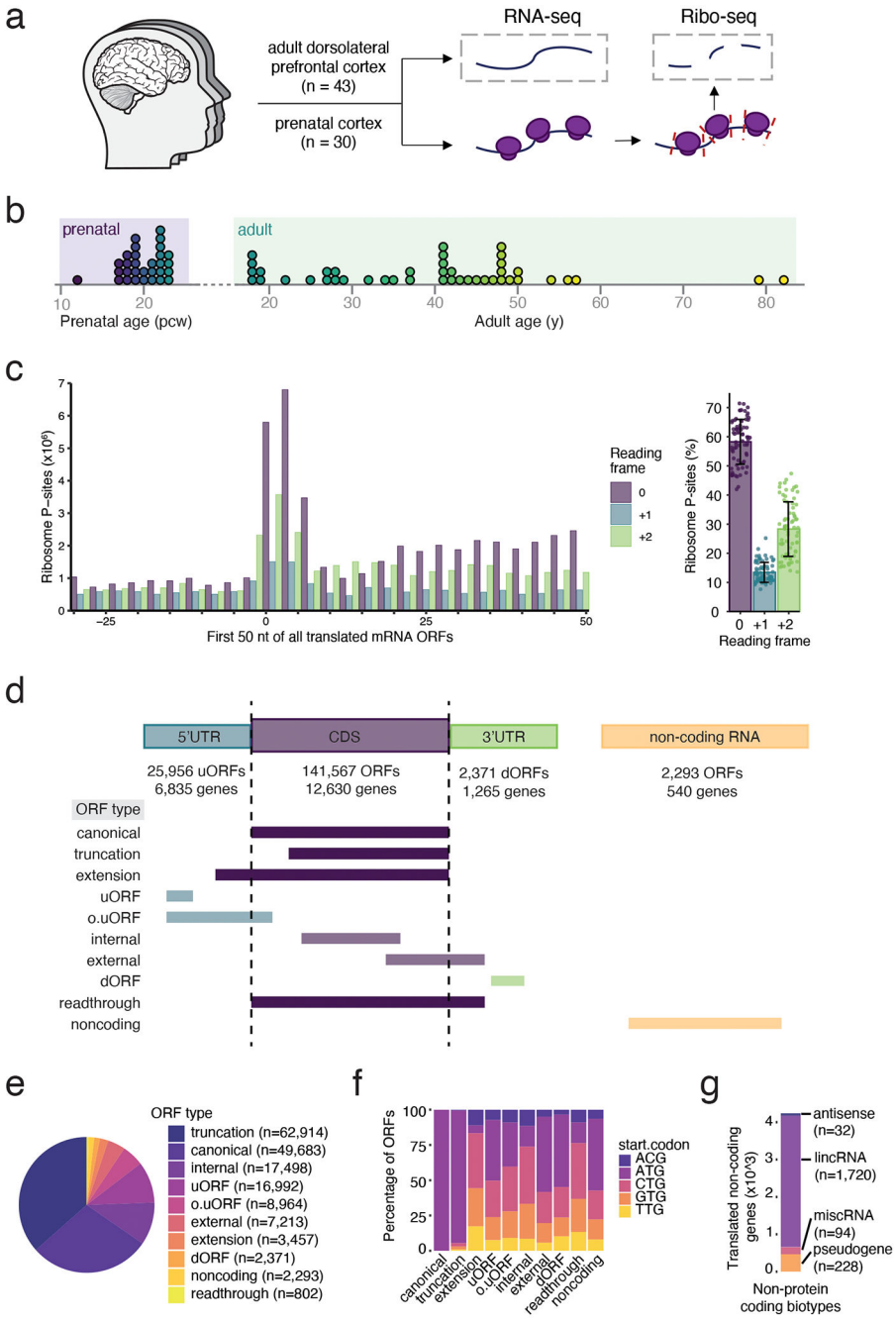

Figure 1: Ribosome profiling captures active translation in the human adult and prenatal brain.

(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 ± SD, n=73 biologically independent tissues. (d) Schematic overview of ORF types detected by RibORF. (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 non-coding RNAs 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.

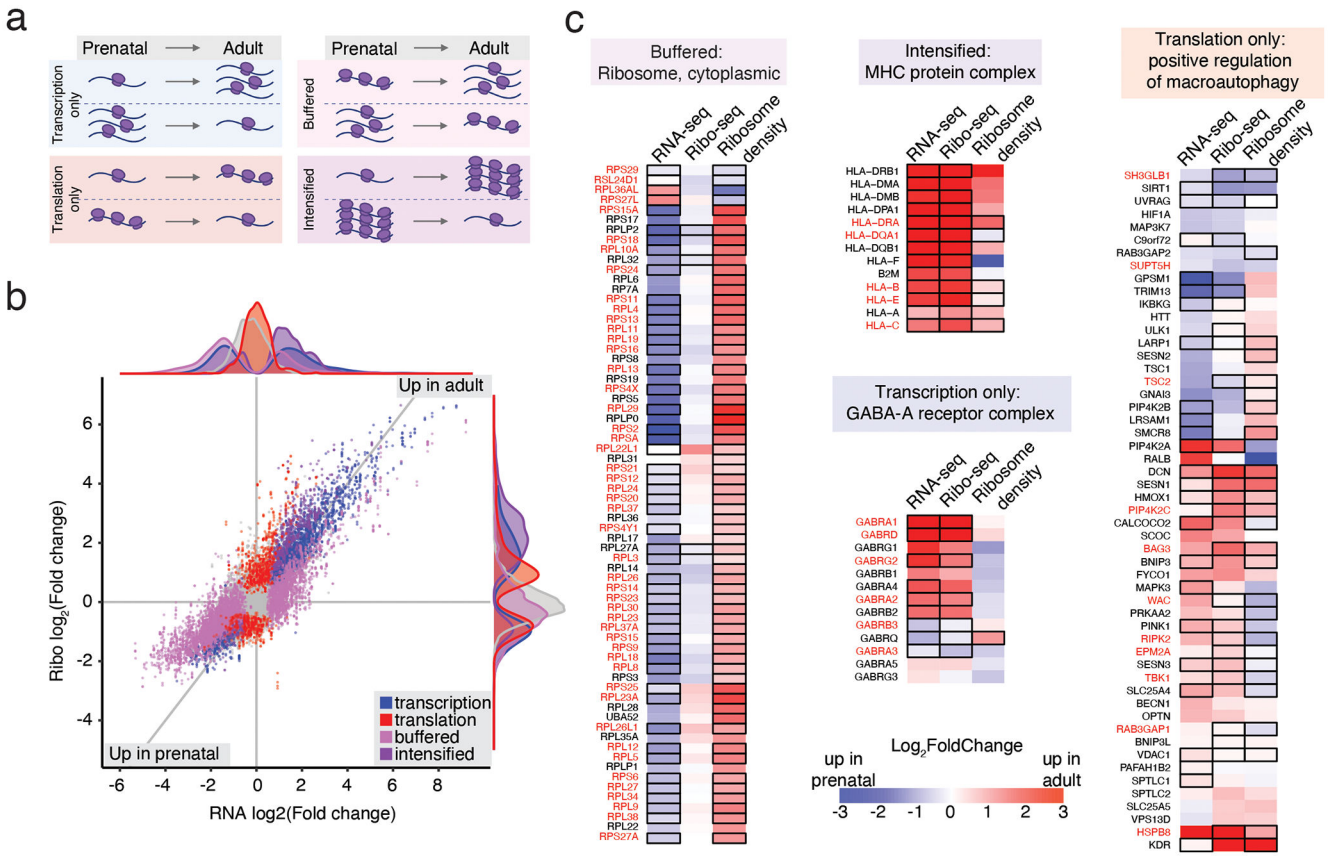

Figure 2: Transcriptional and translational regulation across human brain development.

(a) Classification of genes based on RNA-seq, Ribo-seq, and ribosome density measurements. (b) Scatterplot 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 ribosome density), translationally regulated genes (red; change in ribosome density with no change in transcription), buffered genes (light purple; change in ribosome density that counterbalances the change in mRNA transcription), and intensified genes (dark purple; change in ribosome density that amplifies the change in mRNA) are highlighted. (c) Heatmap of genes associated with the top GO term in each regulatory category identified in A. Black outlines indicate DESeq2 $p_{adj} < 0.05$, gene names in red indicate inclusion in a given regulatory category.

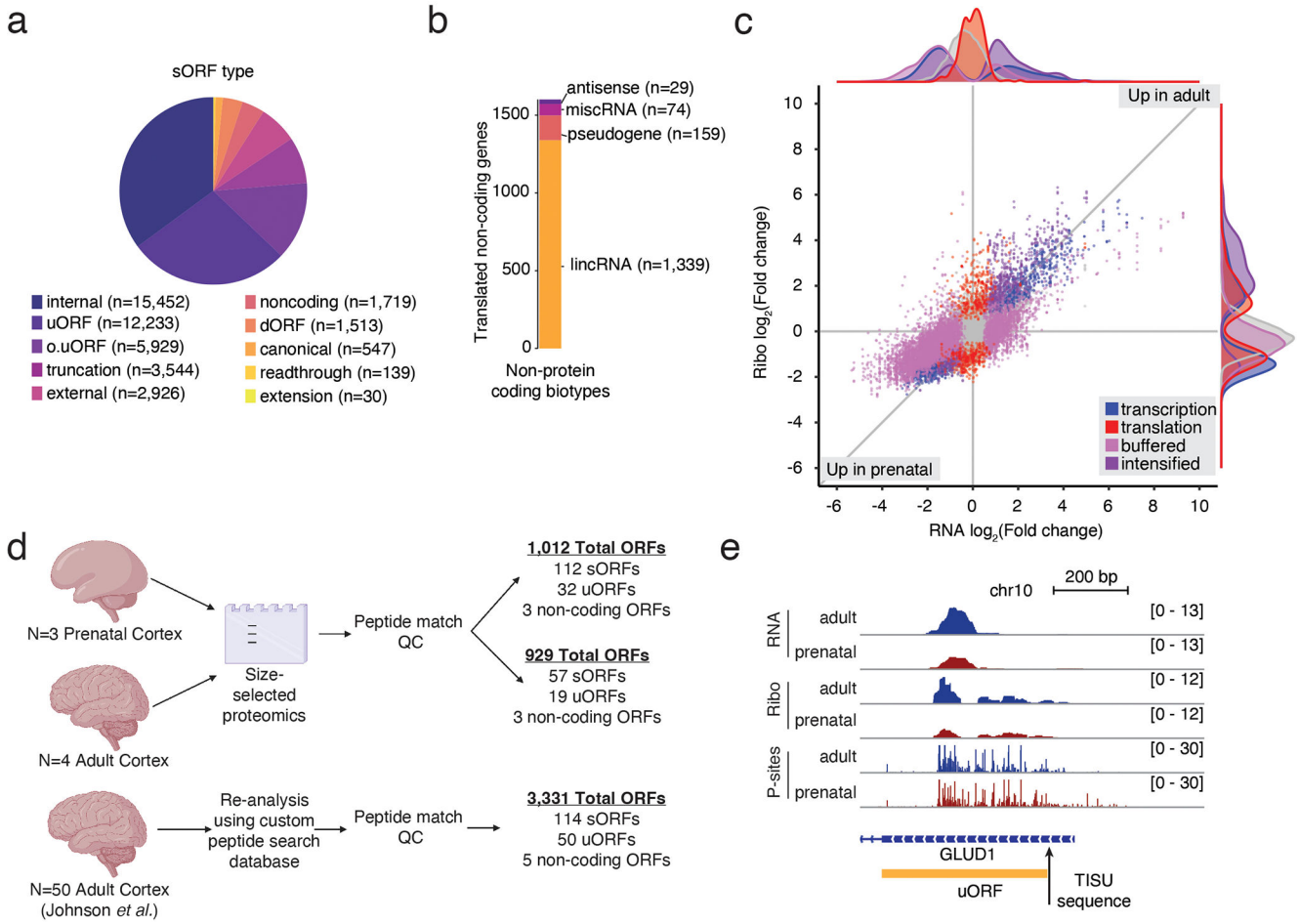

Figure 3: Microprotein expression and validation across brain development.

(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 non-coding RNAs containing at least one sORF, separated by transcript biotype. (c) Scatterplot 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 vs. 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.

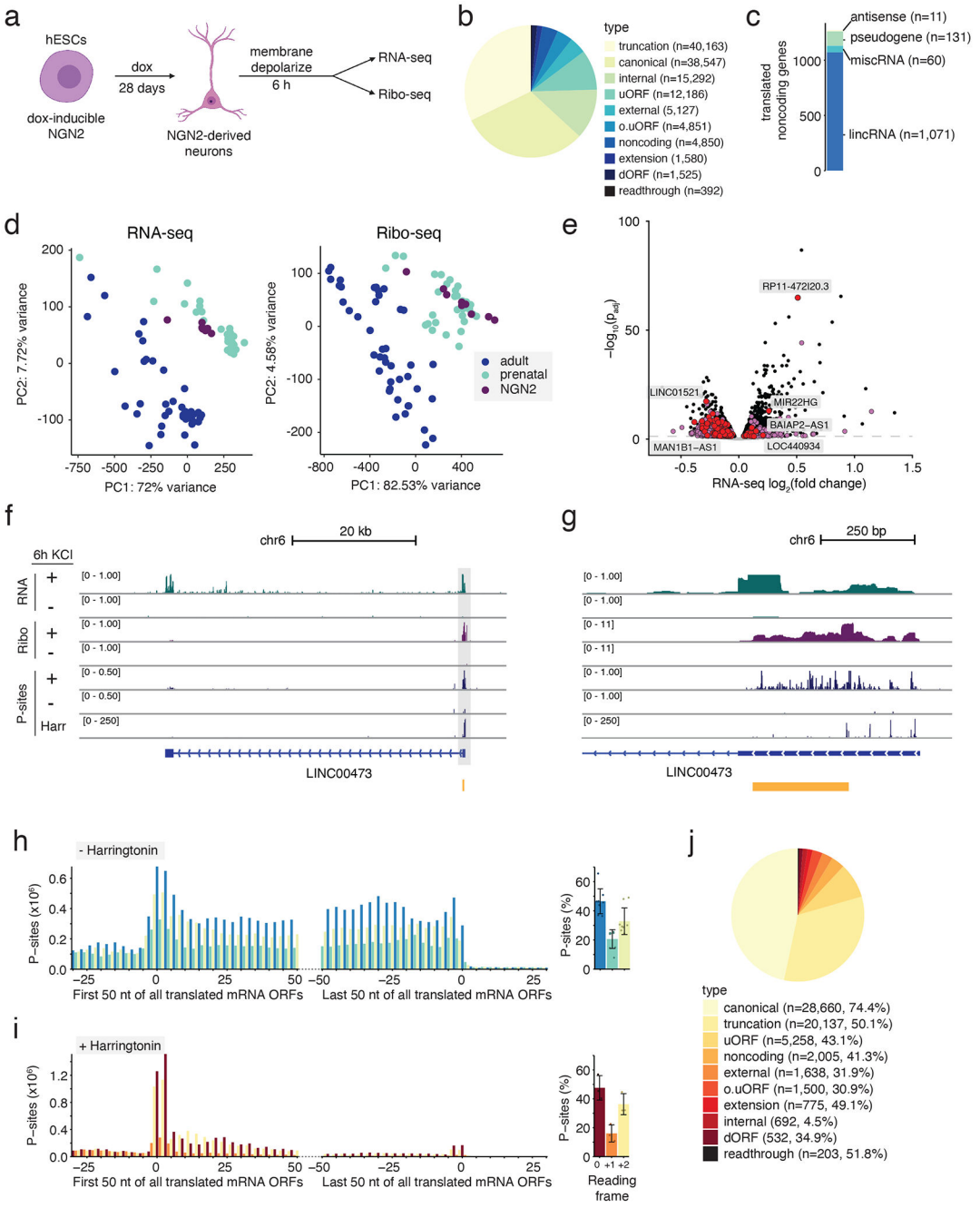

Figure 4: Activity-dependent translation in hESC-derived neurons.

(a) Schematic of Ribo-seq and RNA-seq from NGN2-derived hESCs following 6 h membrane depolarization. (b) Breakdown of translated ORFs of each type identified in NGN2-derived neurons. (c) Stacked bar plot of numbers and percentages of translated non-coding RNAs 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 $-\log_{10}(p_{adj})$ versus $\log_2$(fold-change) in RNA-seq expression between membrane-depolarized and unstimulated NGN2 neurons. Black indicates DEseq2 $p_{adj} < 0.05$, purple indicates activity-dependent non-coding RNAs

with no evidence of translation in human brain or NGN2 neurons, red indicates activity-dependent non-coding RNAs with evidence of translation in human brain and/or NGN2 neurons. (f) Genomic locus of *LINC00473* in NGN2 neurons. 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. The sORF identified by RibORF is shown in gold. (g) High resolution depiction of genomic locus of the ORF encoded by *LINC00473* in NGN2 neurons. 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. 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 last 50 nt (right) of annotated ORFs are shown. Data are shown as mean ± SD, 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 last 50 nt (right) of annotated ORFs are shown. Data are shown as mean ± SD, 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.

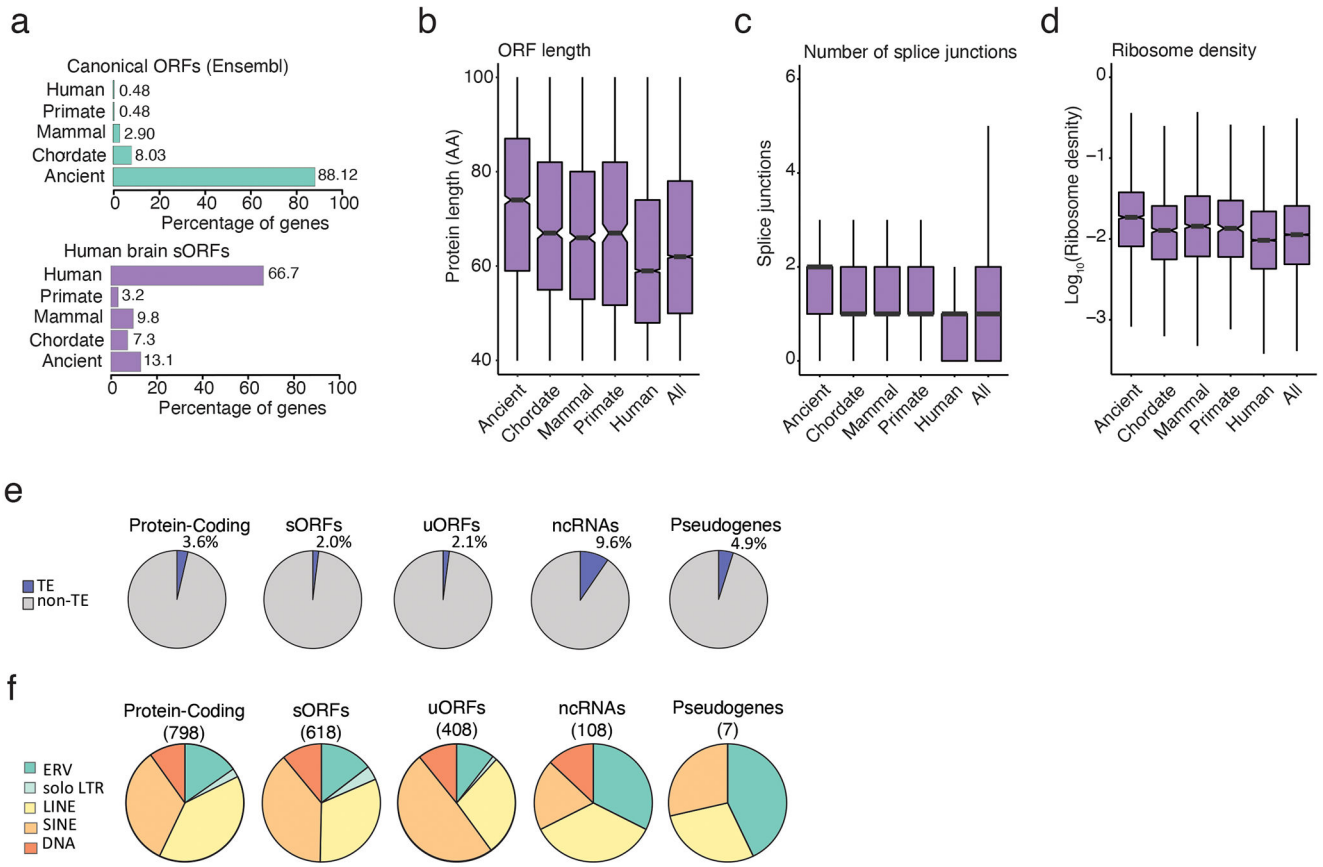

Figure 5: Evolutionary origins of human brain sORFs.

(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 ribosome density grouped by evolutionary age. (b-d) Data are shown as median ± IQR (whiskers = 1.5*IQR), notches indicate median +/− 1.58*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 non-coding RNA biotype. (f) Pie chart of the distribution of TE types, grouped by ORF type or non-coding RNA biotype. Numbers indicate the number of ORFs in each category. IQR, interquartile range; ncRNA, non-coding RNA.

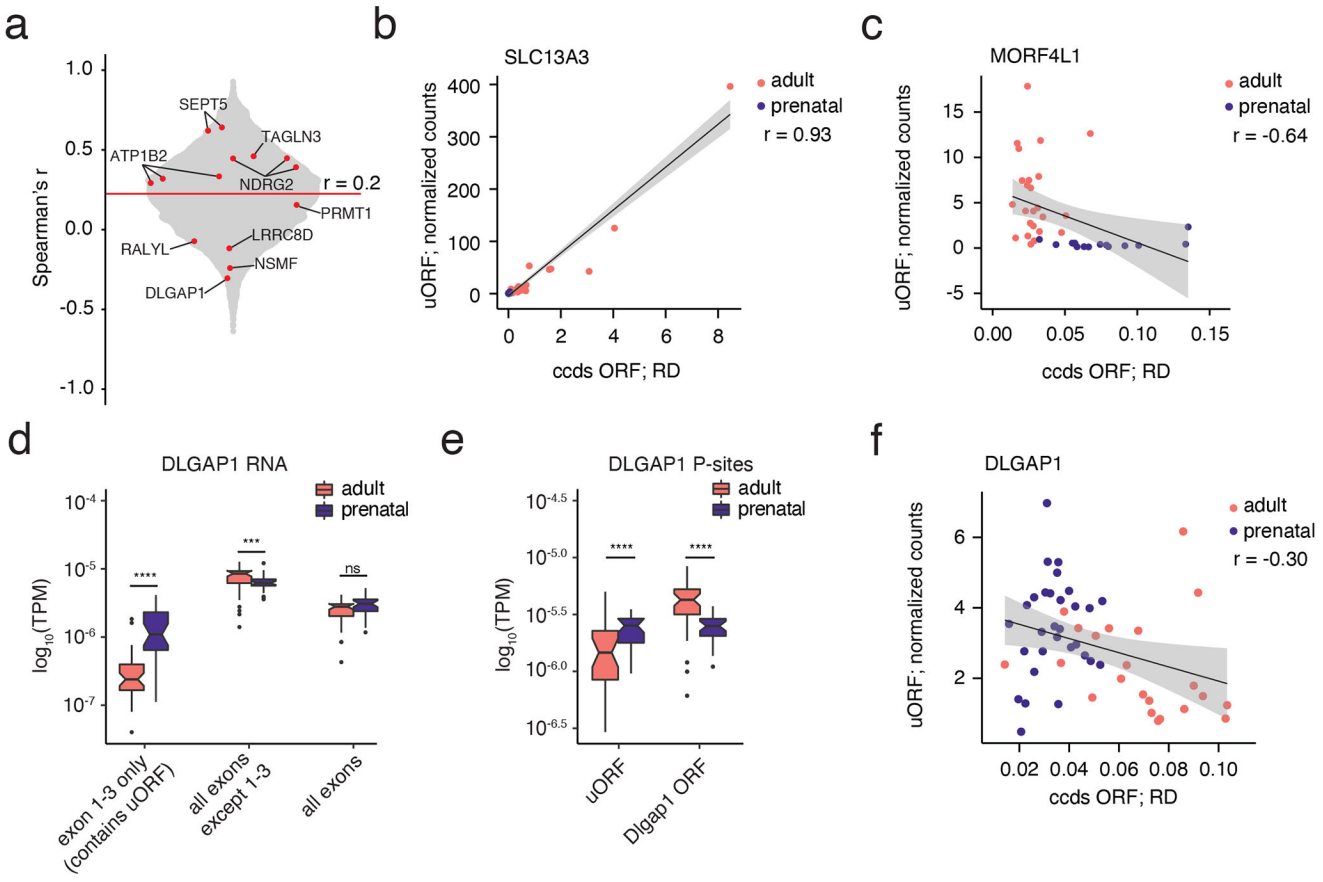

Figure 6: Effects of uORF expression on downstream ORF translation.

(a) Beeswarm dot plot showing the Spearman's r correlation between uORF translation (uORF normalized counts from ribosome profiling) and canonical ORF ribosome density (ccds normalized counts from ribosome profiling divided by normalized total RNA abundance) for individual genes across all 73 individuals. Red line represents the mean correlation across all genes. Red dots indicate developmentally regulated uORFs (described in Supplementary Figure 6a). (b-c) Scatterplot and Spearman's r correlation between upstream ORF translation (uORF normalized counts from ribosome profiling) and canonical ORF ribosome density (ccds normalized counts from ribosome profiling divided by normalized total RNA abundance) for *MAP2K1* (b) and *PIK3C2B* (c) across 73 individuals. Gray shading = 95% CI. (d) Box and whisker plot of RNA-seq reads (transcripts per million reads, TPM) from adult and prenatal samples over *DLGAP1* exons 1-3 ($p = 7.47*10^{-10}$), all exons except 1-3 ($p = 4.20*10^{-4}$), 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 ($p = 6.15*10^{-5}$) and ccds ORF ($2.87*10^{-8}$). (d-e) **** $p < 0.0001$, *** $p < 0.001$ by two-sided Kolmogorov–Smirnov test. Data are shown as median ± IQR (whiskers = 1.5*IQR), notches indicate median +/− 1.58*IQR/sqrt(n), n = 43 (adult) and 30 (prenatal) biologically independent tissues. (f) Scatterplot and Spearman's r correlation between upstream ORF translation (uORF normalized counts from ribosome profiling) and canonical ORF ribosome density (ccds normalized counts from ribosome profiling divided by normalized total RNA

abundance) for *DLGAP1* across 73 individuals. Gray shading = 95% CI. IQR, interquartile range; NS, not significant.

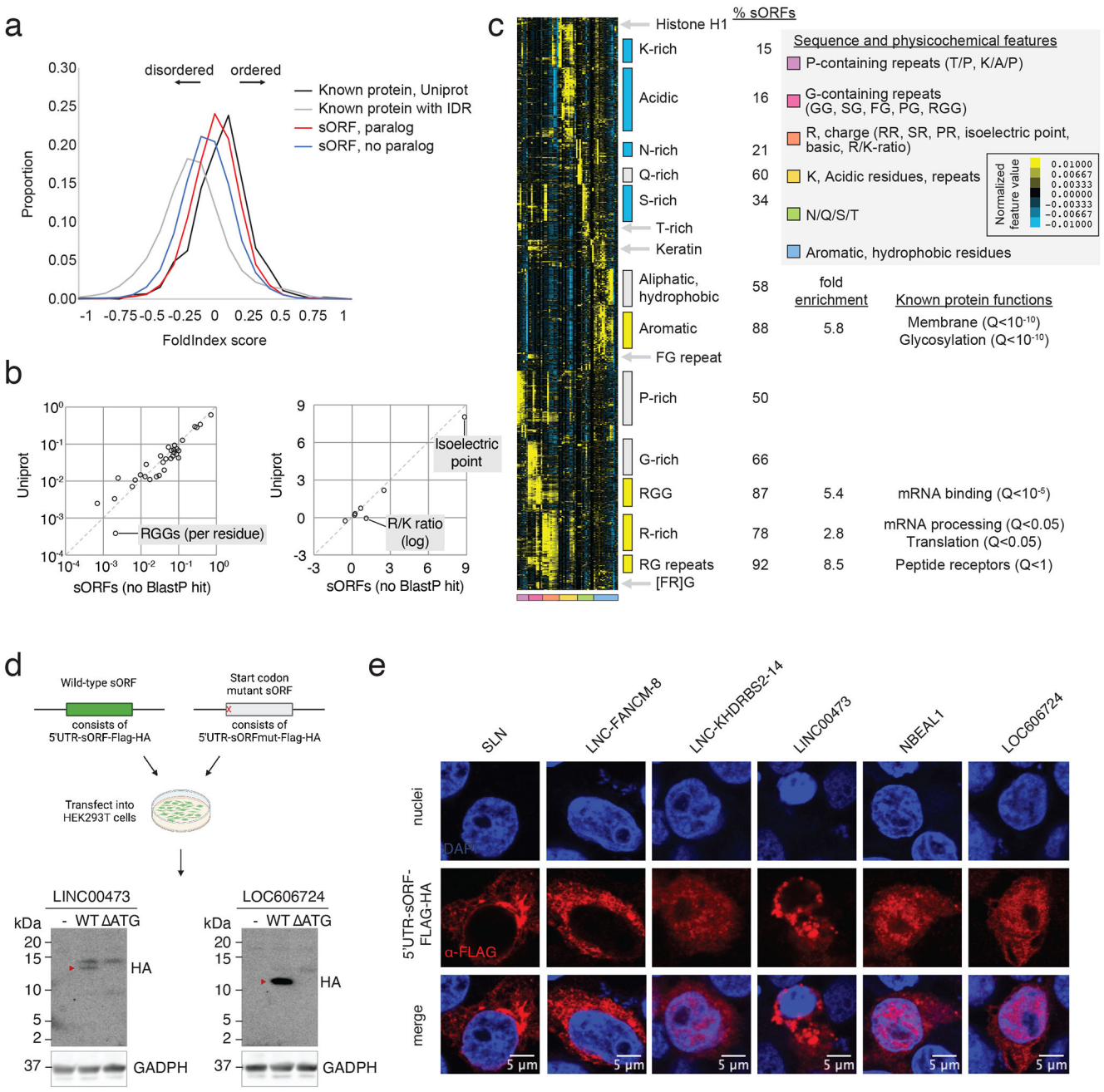

Figure 7: Microprotein functional characterization.

(a) FoldIndex score distribution of proteins annotated in Uniprot (black), annotated proteins with intrinsically disordered regions (gray), and sORFs with and without a BlastP hit (red and blue, respectively). (b) Scatterplot 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) Heatmap 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 (e.g. internal, external, readthrough), were excluded from this analysis. Boxes to the right of the heatmap indicate clusters of IDRs with similar properties. Blue = clusters depleted for sORFs, yellow = clusters significantly enriched for sORFs. (d) Western blot of FLAG-HA-tagged unmodified and ATT-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.