PNAS

RESEARCH ARTICLE | BIOCHEMISTRY

Cryptic intronic transcriptional initiation generates efficient endogenous mRNA templates for C9orf72-associated RAN translation

Shannon L. Miller[a,b], Katelyn M. Green[a,b], Bradley Crone[c], Jessica A. Switzenberg[c], Elizabeth M. H. Tank[a], Amy Krans[a,d], Karen Jansen-West[e], Clare M. Wieland[a,f,g], Eric W. Ji[a], Leonard Petrucelli[e], Sami J. Barmada[a], Alan P. Boyle[c,h], and Peter K. Todd[a,d,1]

Affiliations are included on p. 10.

Edited by Nahum Sonenberg, McGill University, Montreal, QC, Canada; received March 28, 2025; accepted July 3, 2025

Intronic GGGGCC hexanucleotide repeat expansions in *C9orf72* are the most common genetic cause of amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD). Despite its intronic location, this repeat avidly supports synthesis of pathogenic dipeptide repeat (DPR) proteins via repeat-associated non-AUG (RAN) translation. However, the template RNA species that undergoes RAN translation endogenously remains unclear. Using long-read based 5′ RNA ligase-mediated rapid amplification of cDNA ends (5′ Repeat-RLM-RACE), we identified *C9orf72* transcripts initiating within intron 1 in a C9BAC mouse model, patient-derived iNeurons, and iNeuron-derived polysomes. These cryptic m^{7}G-capped mRNAs are at least partially polyadenylated and are more abundant than transcripts derived from intron retention or circular intron lariats. In RAN translation reporter assays, intronic template transcripts–even those with short (32 nucleotide) leaders–exhibited robust expression compared to exon–intron and repeat-containing lariat reporters. To assess endogenous repeat-containing lariat RNA contributions to RAN translation, we enhanced endogenous lariat stability by knocking down the lariat debranching enzyme Dbr1. However, this modulation did not impact DPR production in patient-derived iNeurons. These findings identify cryptic, linear, m^{7}G-capped intron-initiating *C9orf72* mRNAs as an endogenous template for RAN translation and DPR production, with implications for disease pathogenesis and therapeutic development.

neurodegeneration | repeat expansion disease | ALS | translation | transcription

The most common genetic cause of both amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) is a GGGGCC hexanucleotide repeat expansion in the first intron of *C9orf72*, accounting for 30 to 50% of familial ALS and 7% of sporadic ALS (1–3), as well as up to 25% of familial FTD and 6 to 8% of sporadic FTD (4). Most healthy individuals have only 2 GGGGCC repeats in C9orf72, but patients diagnosed with ALS/FTD due to *C9orf72* mutations (C9ALS/FTD) can harbor hundreds to thousands of repeats (1, 2).

One hallmark of C9ALS/FTD is the presence of dipeptide-repeat-containing proteins (DPRs), which are generated through a noncanonical protein translation initiation mechanism termed repeat-associated non-AUG (RAN) translation. Both sense- and antisense-generated DPRs accumulate within patient brains (5–8). From the sense GGGGCC strand, a poly-glycine-alanine (GA), poly-glycine-proline (GP), and poly-glycine-arginine (GR) containing protein are translated. The antisense CCCCGG strand produces poly-proline-arginine (PR), poly-proline-alanine (PA), and a second poly-GP containing protein, although it is worth noting that the PR and PA reading frames contain an AUG codon that could be used for initiation (9). Overexpression of DPRs across multiple model systems causes neurotoxicity in the absence of repeat RNA or its native sequence context (8, 10–17), suggesting that DPRs may be sufficient for neurodegeneration upon their accumulation.

Multiple groups have utilized reporters to determine how GGGGCC repeats support RAN translation. While there are differences in how the reporters were designed, data from multiple groups suggest the following: RAN translation of the GA reading frame is the most robust (18–24), a CUG codon located 24 nucleotides upstream from the repeat is important for efficient initiation of translation in the GA reading frame (18, 19, 21, 25–29), GGGGCC repeats are most robustly translated from an m^{7}G capped mRNA (18, 19, 21, 22, 30), and *C9orf72*-associated RAN translation (C9RAN) is upregulated when the integrated stress response is activated (18, 20–22, 31).

Significance

An intronic GGGGCC repeat expansion in *C9orf72* supports an unusual translational initiation process known as repeat-associated non-AUG (RAN) translation to produce toxic dipeptide repeat (DPR) proteins that contribute to neurodegeneration in ALS and FTD. How an intronic repeat RNA engages with ribosomes to support such translation is unclear. Here, we identify a series of previously unannotated mRNA transcripts that initiate within the repeat-containing intron to create linear m^7 G-capped templates for RAN translation from GGGGCC repeats. These cryptic mRNAs are present in patient iNeurons, engage with ribosomes, and robustly support RAN translation. This finding has important implications for both our understanding of the mechanism by which RAN translation occurs and on therapeutic development in this currently untreatable class of neurodegenerative disorders.

Competing interest statement: L.P. serves as a consultant for Expansion Therapeutics P.K.T. served as a consultant with Denali Therapeutics and has licensed technology through the University of Michigan to Denali., P.K.T. holds a shared patent on ASOs related to RAN translation with Ionis Pharmaceuticals.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: petertod@umich.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2507334122/-/DCSupplemental.

Published August 4, 2025.

While these reporter assays have yielded valuable insights into disease, the exact RNA template that undergoes RAN translation in patients remains unclear. Typically, introns are spliced from mature mRNA transcripts as circularized intron lariats and then rapidly turned over via linearization by the lariat debranching enzyme Dbr1. Once linearized, they are subsequently degraded by nuclear exonucleases, precluding their export to the cytoplasm and ability to initiate translation through interaction with ribosomes. Understanding the template for C9RAN translation and how the intronic GGGGCC repeat in *C9orf72* escapes this fate are areas of active study (Fig. 1*A*). The GGGGCC repeat could theoretically escape degradation and exit to the cytoplasm by impaired splicing, by aberrantly stabilizing a correctly spliced intron, or through inclusion within transcripts with unannotated 5′ or 3′ ends (28). A splicing failure would lead to retention of the repeat and first intron within mature *C9orf72* mRNA — allowing for its export to the cytoplasm and engagement of translation machinery (32, 33). Alternatively, if the spliceosome successfully removed the intron containing the GGGGCC repeat, then the robust secondary structure formed by the repetitive GC rich sequence could aberrantly stabilize the lariat. Spliced introns have been previously reported in the cytoplasm of cells (34), and in C9ALS/FTD, C9 intron lariat RNA persists in the cytoplasm both in cells transfected with a splice-capable reporter and in patient-derived fibroblasts as measured by smFISH (35). Introns have historically been understood to be noncoding, but some circular RNAs, which arise from back splicing events, are translated in a cap-independent manner (36, 37). Finally, it is possible that the GGGGCC repeat interferes with proper *C9orf72* transcription, generating transcripts that lack the canonical 5′ or 3′ splice sites to place the repeat in a nonintronic context. This scenario could arise from altered transcription initiation from previously unmapped 5′ start sites or by premature transcription termination, leading to 3′ truncated mRNA (38).

Defining the GGGGCC repeat-containing mRNA transcript template that supports RAN translation is critical both for our understanding of this enigmatic process and for development of therapies for C9ALS/FTD. Here, we describe a set of unannotated

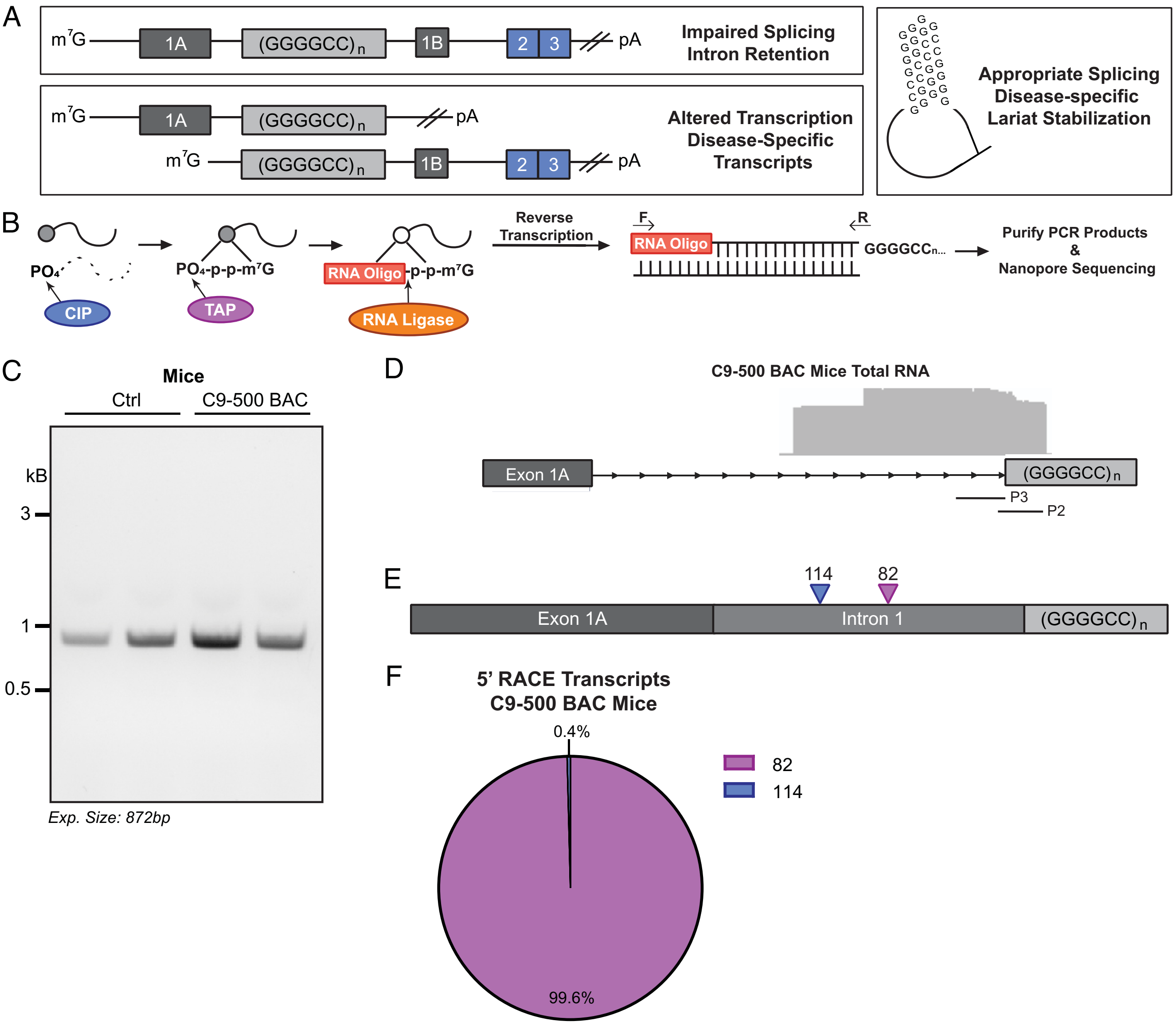

Fig. 1. 5′ RACE reveals intronic initiation of C9orf72 repeat RNAs in C9BAC mouse model. (*A*) Schematic of potential GGGGCC repeat containing mRNA templates for RAN translation in C9ALS/FTD. (*B*) Schematic of 5′ RACE-RLM experimental procedure. (*C*) Agarose gel of murine Beta-actin 5′ RACE PCR in C9-500 BAC and control mice, *n* = 2. (*D*) Aligned integrative genome viewer (IGV) plots using primers 2 and 3 in the cerebellum of C9-500 BAC mice, *n* = 3. (*E*) Annotated schematic of the most commonly identified 5′ RACE products using primers 2 and 3 in the cerebellum of C9-500 BAC mice, *n* = 3. (*F*) Pie chart of most commonly identified 5′ RACE products using primers 2 and 3 in the cerebellum of C9-500 BAC mice, *n* = 3/group.

5′ m^7G-capped *C9orf72* RNAs that are derived from cryptic initiation within intron 1 of the C9 locus. These cryptic transcripts are selectively present in C9BAC-500 mice compared to controls and are also observed in patient-derived fibroblasts, patient-derived iNeurons, and iNeuron-derived polysomes. Enrichment of poly-adenylated RNA captured these transcripts as well as additional transcripts with initiation sites in or above exon 1, indicating that some of these transcripts are poly-adenylated and some derive from intron retention, respectively. Using reporters, we find that these 5′ intron-initiating transcripts are translated much more robustly than exon–intron or lariat reporters with the same sized repeats. Moreover, efforts to enhance translation from endogenous C9 intron lariat RNA through lariat stabilization did not impact DPR production in patient-derived iNeurons. Taken together, these data define an unexpected native template for RAN translation from the GGGGCC repeat in *C9orf72*, with implications for the mechanism by which this translation occurs and for therapies aimed at mitigating this process.

Results

5′ RACE Reveals Intron-Initiating *C9orf72* Repeat-Containing mRNAs in C9BAC Mouse Model. We used a BAC transgenic mouse model engineered to express full-length human *C9orf72* with 500 GGGGCC repeats in intron 1 as well as human surrounding sequence (C9-500 BAC) (39). This mouse recapitulates key molecular and pathological features of C9ALS/FTD, including production of sense and antisense repeat containing RNA species and generation of polyGA, polyGP, and polyGR DPRs. As the cerebellum of these mice are particularly rich polyGA and polyGP aggregates (39, 40), we isolated total RNA from cerebellar homogenates from 6-wk-old mice and performed 5′ RNA ligase mediated-rapid amplification of cDNA ends (5′ Repeat-RLM-RACE) to identify the 5′ end of transcribed *C9orf72* GGGGCC repeat-containing mRNAs (Fig. 1*B*). cDNA was synthesized using random hexamers and spiked in $(CCCCGG)_4$ to increase the probability of the reverse transcriptase making it through the repeat, which is known to impair PCR (2). Unlike traditional 5′ RACE, we developed a long-read nanopore sequencing platform of our PCR amplicons rather than TOPO cloning to create a high-throughput analysis pipeline, allowing us to better capture both high and low-frequency transcripts and to determine the frequency of different products more accurately. Control RACE PCRs using primers specific for murine beta-actin and the 5′ ligated RNA oligo (*SI Appendix*, Table S1) revealed a single band of the expected size on an agarose gel, confirming that all steps of the reaction were successful (Fig. 1*C*).

We initially used three RACE primers specific to intron 1 of *C9orf72* (P1, P2, and P3, *SI Appendix*, Table S1), but found that the lack of sequence specificity of P1 did not allow for C9-specific products to be identified. However, using P2 and P3, we identified a 5′ end transcript that initiated within intron 1, 82 nucleotides (nt) upstream of the repeat in C9-500 BAC mice (Fig. 1 *D*–*F*). At low frequency (<1%), one additional product was identified, initiating at 114nt above the repeat (Fig. 1 *D*–*F*). Both of the identified transcripts contained the CUG codon located 24nt upstream of the repeat thought to drive translation of the repeats in the polyGA frame (18, 19, 21, 25). These products were not reliably detected from control mice, as P1, P2, and P3 do not have sufficient sequence homology with murine *C9orf72* to allow for binding (*SI Appendix*, Fig. S1*A*). Of note, we did not identify products mapping to known transcription start sites in exon 1A in our mouse samples despite PCR cycling parameters that would allow sufficient time for these products to be amplified, suggesting that most repeat-containing mRNAs do not arise from intron retention.

5′ RACE in C9ALS/FTD Patient-Derived Cells Identifies Additional Intron-Initiating *C9orf72* RNAs. In patient-derived cells, we initially utilized a traditional low-throughput 5′ RACE protocol to determine whether intron-initiated 5′ end products are also present in human cells. RNA was harvested from one control and two C9ALS/FTD fibroblast lines, and PCR products from P1, P2, and P3 amplification reactions were run on an agarose gel. Bands were excised, gel purified, TOPO cloned, and Sanger sequenced. Using P2 and P3, we identified a short product in both C9ALS/FTD and control fibroblasts initiating 32nt upstream of the repeat and only 8nt 5′ to the CUG near-cognate initiation codon (*SI Appendix*, Fig. S1*B*). This mRNA initiates at a −1 T, +1A motif that is enriched at transcription start sites. However, only fourteen total clones mapped to *C9orf72* over multiple reactions. While more clones came from C9ALS/FTD fibroblast lines, the presence of this product in at least one control line suggests that it is not disease specific. We identified no *C9orf72*-specific products with the P1 primer from C9ALS/FTD or control fibroblasts.

To determine whether these unannotated 5′ end transcripts undergo RAN translation in disease-relevant cells, we generated iNeurons from human induced pluripotent stem cells containing an NGN2-cassette from one control and one C9ALS/FTD line. Utilizing iNeuron lysates, we performed polysome profiling and captured RNA from 80S and polysome fractions to use as our input for 5′ RACE (Fig. 2*A*). Sequencing of clones that utilized the P1 primer did not yield any *C9orf72*-specific hits, but clones derived from the P3 primer identified sixteen *C9orf72* hits that mapped to the 32nt product seen in fibroblasts (Fig. 2*B*). We also observed a rare, longer *C9orf72* isoform in our C9ALS/FTD iNeuron line initiating near the mid-point of intron 1 (Fig. 2*B*). Thus, short, 5′-capped *C9orf72* transcripts are present within fibroblasts and actively translating fractions of patient-derived iNeuron RNA, suggesting these species could contribute to DPR production.

Due to the small number of clones derived with this low-throughput Sanger sequencing-based method, we leveraged our higher-throughput pipeline with nanopore sequencing on a larger cohort of iNeuron lines as was done with the C9-500 BAC mice. RNA from three control and three C9ALS/FTD iNeuron lines were harvested at day 10 post differentiation and processed for 5′ RACE. The presence of a single band of the expected size using control primers complementary to human beta-actin and 5′ RNA oligo primers on an agarose gel confirmed all steps of the reaction were successful (*SI Appendix*, Fig. S1*C*). In addition, nanopore sequencing of products generated using a reverse primer in exon 2 showed 5′ ends which mapped to exon 1B and 1A, respectively (*SI Appendix*, Fig. S1*D*), demonstrating the presence of correctly spliced, mature RNA in both C9ALS/FTD and control iNeurons and suggesting that the RACE process from the C9 locus is unimpeded by the repeat. Nanopore sequencing of PCR products from P2 and P3 identified the same 82nt product seen in C9-500 BAC mice which represented ~97% and 85% of the total products in C9ALS/FTD and control iNeurons, respectively, as well as the 114nt product, representing 2.5 and 12% of the total products, respectively (Fig. 2 *C*–*G*) As in the C9-500 BAC mice, we did not identify repeat-RLM-RACE reads from total RNA in C9ALS/FTD iNeurons mapping to exon 1A, suggesting that most repeat-containing mRNAs in human iNeurons do not arise from intron retention. However, a small percentage (4%) of reads from control iNeurons mapped 5′ to exon 1A (Fig. 2*G*).

To assess whether these unannotated intron-initiating *C9orf72* RNAs were polyadenylated, we performed polyA-capture using

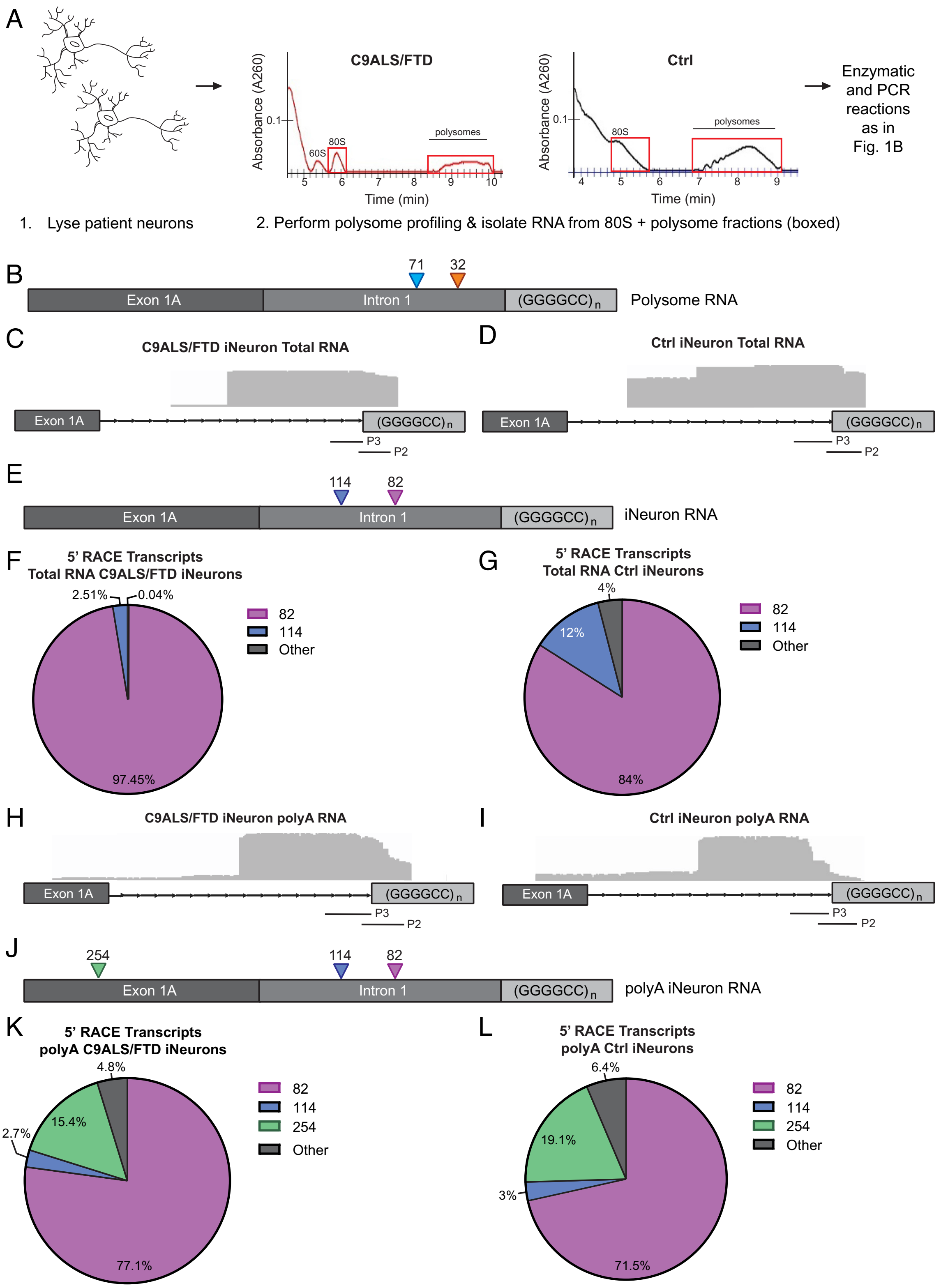

Fig. 2. 5′ RACE reveals intronic-initiating C9orf72 repeat RNAs in C9ALS/FTD patient-derived iNeurons. (*A*) Schematic of 5′ RLM-RACE experimental flow for RNA isolated from control and C9orf72 patient-derived iNeuron 80S and polysome fractions. Polysome profiles shown were obtained from ~90 µg total RNA isolated from C9ALS/FTD iNeurons (*Left*) and control iNeurons (*Right*), with 80S and polysome fractions labeled. (*B*) Annotated schematic of the most commonly identified 5′ RACE products using primers 1, 2, and 3 in polysome-associated RNA from C9ALS/FTD and control iNeurons, *n* = 1/group. (*C*) Aligned integrative genome viewer (IGV) plots using primers 2 and 3 in total RNA from C9ALS/FTD iNeuron lines, *n* = 3. (*D*) Aligned IGV plots using primers 2 and 3 in total RNA from control iNeuron lines, *n* = 3. (*E*) Annotated schematic of the most commonly identified 5′ RACE products using primers 2 and 3 in total RNA from C9ALS/FTD and control iNeuron lines, *n* = 3/group. (*F*) Frequency of most commonly identified 5′ RACE products using primers 2 and 3 in total RNA from C9ALS/FTD iNeuron lines, *n* = 3/group. (*G*) Frequency of most commonly identified 5′ RACE products using primers 2 and 3 in total RNA from control iNeuron lines, *n* = 3/group. (*H*) Aligned IGV plots using primers 2 and 3 in polyA RNA from C9ALS/FTD iNeurons, *n* = 3. (*I*) Aligned IGV plots using primers 2 and 3 in polyA RNA from control iNeurons, *n* = 3. (*J*) Annotated schematic of the most commonly identified 5′ RACE products identified using primers 2 and 3 in polyA-captured RNA from C9ALS/FTD and control iNeuron lines, *n* = 3/group. (*K*) Frequency of most commonly identified 5′ RACE products using primers 2 and 3 in polyA-captured RNA from C9ALS/FTD iNeuron lines, *n* = 3/group. (*L*) Frequency of most commonly identified 5′ RACE products using primers 2 and 3 in polyA-captured RNA from control iNeuron lines, *n* = 3/group.

oligo-dT beads prior to performing the nanopore-based repeat-RLM-RACE protocol on the same control and C9ALS/FTD iNeuron lines we harvested total RNA from. Two products in the polyA-captured RNA matched those seen in C9-500 BAC mice and total RNA from C9ALS/FTD and control iNeurons (82nt and 114nt) (Fig. 2 *H–L*). The persistence of these products in our polyA-captured 5′ RACE dataset suggests that at least some of these intron-initiating products are polyadenylated, though our experimental paradigm precludes our ability to determine whether all intron-initiating RNAs are polyadenylated. Interestingly, ~15 to 20% of polyA-captured products mapped back to exon 1A, suggesting at least some RNAs retain intron 1 and that these RNAs are polyadenylated (Fig. 2 *H–L*). Thus, across multiple cell types and systems, we identified unannotated, intron-initiating *C9orf72* RNAs, and demonstrate that at least a percentage of these RNAs are polyadenylated.

RAN Translation of Intron-Initiated mRNAs Exhibit Cap Dependence and CUG Codon Usage. To understand more about the 5′ end *C9orf72* RNAs we identified by 5′ RACE, we generated a series of nanoluciferase (NLuc) reporters to assess their translation efficiency across multiple systems (Fig. 3*A*). Briefly, the AUG start codon of NLuc was mutated to GGG, precluding translation, as previously shown (18, 41). Upstream of GGG-NLuc we inserted 70 GGGGCC repeats in the polyGA (+0), polyGP (+1), or polyGR (+2) reading frame. The 5′ ends we identified through our RACE studies were then inserted upstream of the repeats and immediately downstream of the T7 promoter, so no additional vector sequence was present in in vitro transcribed RNAs from these plasmid reporters. We also generated a "whole intron" reporter encompassing all 162nt of intron 1 to compare to the shorter within-intron initiating RNAs ("short intron" and "mid intron", respectively) (*SI Appendix*, Table S2). In line with previous results, expression of the whole intron GA70 RNA was less robust than an AUG-driven intron GA70 RNA in HEK293 cells (*SI Appendix*, Fig. S2*A*) (18). Of note, all reporters contained the CUG start codon located 24 nucleotides upstream of the GGGGCC repeat.

Using in vitro transcribed RNAs from these reporters, we observed that the mid intron reporter, which corresponds to the most common product identified by 5′ RACE, was expressed more robustly in the GA reading frame than the whole intron or the short intron RNA reporter in vitro in rabbit reticulocyte lysate (RRL), HEK293 cells, and rat hippocampal neurons (Fig. 3 *B–D*). The mid intron RNA reporter was also expressed to the highest extent in the GP and GR reading frames both in vitro and in rat hippocampal neurons (*SI Appendix*, Fig. S2 *B–D*). In HEK293 cells, while the mid intron reporter was expressed more robustly than the short intron RNA reporter in the polyGP and polyGR frames, expression was not significantly increased from the whole intron RNA reporter (Fig. 3 *E* and *F*). Next, we assessed whether the CUG codon 24nt upstream of the repeat was obligate for RAN translation from these shorter leaders. Typically, there is a 12 to 40nt region at the end of 5′ m^7G capped mRNA transcripts that is not permissive for initiation due to eIF4F binding (42–44). Despite being located only 8nt downstream from the 5′ cap, the CUG codon was important for polyGA production from the short intron reporter in both RRL and HEK293 cells, as mutation to a CCC completely abolished translation (Fig. 3 *G* and *H*). The loss of the CUG codon had a more modest inhibitory effect on translation from the polyGP in both systems (Fig. 3*H* and *SI Appendix*, Fig. S2*E*), while its loss in the polyGR frame was variable across systems (Fig. 3*H* and *SI Appendix*, Fig. S2*F*), consistent with data suggesting frameshifting may occur within the repeat (18, 19).

The 5′ RACE protocol selectively enriches for 5′ m^7G-capped mRNAs. We therefore assessed whether the translation of these intron-initiated transcripts exhibited 5′ m^7G-cap dependence. RNAs were in vitro transcribed with either a 5′ m^7G- or an A-cap analog that cannot recruit the cap binding initiating factor eIF4E but protects the mRNA from degradation. As expected, translation from Cricket paralysis virus (CrPV), which utilizes a cap-independent internal ribosome entry site (IRES) translation mechanism (45), was not impaired when A-capped, while an AUG-initiated NLuc reporter was strongly cap-dependent (Fig. 3*I* and *SI Appendix*, Fig. S3 *A–E*). In RRL and HEK293 cells, the lack of a canonical 5′ m^7G cap dramatically reduced translation of these unannotated 5′ end reporter RNAs in all reading frames and at all intron leader lengths, including the 32nt leader short intron mRNA (Fig. 3*I* and *SI Appendix*, Fig. S3 *A–E*). An intron retention reporter RNA also demonstrated strong cap dependence (Fig. 3*I* and *SI Appendix*, Fig. S3*A*). In conclusion, our findings demonstrate that intron-initiated *C9orf72* RNAs identified through 5′ RACE are efficiently translated, with the mid intron products corresponding to enhanced expression and even the short leader construct demonstrating a strong dependence on the CUG near-cognate start codon and m^7G cap, highlighting the critical role of these elements in C9RAN translation.

Intron-Initiated mRNAs Are an Efficient Template for RAN Translation. Prior studies have suggested that the major template for RAN translation is either a linear mRNA containing a retained intron (32, 33), or a spliced, circular C9 intron lariat that is exported from the nucleus to the cytoplasm (35). To directly test the efficiency of translation from these potential templates, we generated constructs that included both exon1A and all intronic sequence 5′ to the repeat, followed by 70 GGGGCC repeats in the polyGA frame upstream of NLuc. This construct contains a predicted AUG initiated uORF that would terminate 79nt from the start of the repeat (19, 46). We next modified this construct to directly assess a role for C9 intron lariat-based RAN translation by expressing a NLuc reporter plasmid with 70 GGGGCC repeats in the polyGA frame positioned between the 5′ and 3′ ends of intron 1, flanked by exon 1A upstream and roughly 50 nucleotides of exon 2 fused with a V5 c-terminal tag, excluding the canonical AUG start codon downstream (lariat GA70) (Fig. 4*A*). When transfected into cells, this reporter undergoes splicing to create a circular C9 intron lariat containing the repeat (Fig. 4*A*). To confirm the presence of a spliced circular C9 intron lariat, we designed primers to amplify the region of the lariat that crosses the branch point. Importantly, this sequence is only present if the intron is circularized, and sequencing of HEK293 cells expressing lariat GA70 confirmed our reporter generated the anticipated product (*SI Appendix*, Fig. S4*A*).

To compare the translation efficiency of these reporters to the intron-initiated products, we transfected these linear and lariat GA70 plasmid reporters into HEK293 cells and rat hippocampal neurons. The mid-intron reporter corresponding to the most common intronic-initiating RNA identified by 5′ RACE was by far the most efficiently expressed construct in both systems (Fig. 4 *B* and *C*). Consistent with prior studies in rabbit reticulocyte lysate (19), inclusion of exon1A impaired expression compared to the intron-only GA70 reporter (Fig. 4 *B* and *C*). In both HEK293 cells and rat hippocampal neurons, linear GA70 reporters were expressed between 10- to 100-fold more robustly than the lariat GA70 lariat reporter (Fig. 4 *B* and *C*). Despite this, NLuc signal of lariat GA70 was significantly greater than that of mock transfection (*SI Appendix*, Fig. S4*B*), suggesting that the lariat reporter is engaging with translational machinery and generating products, consistent with previous results (22, 35).

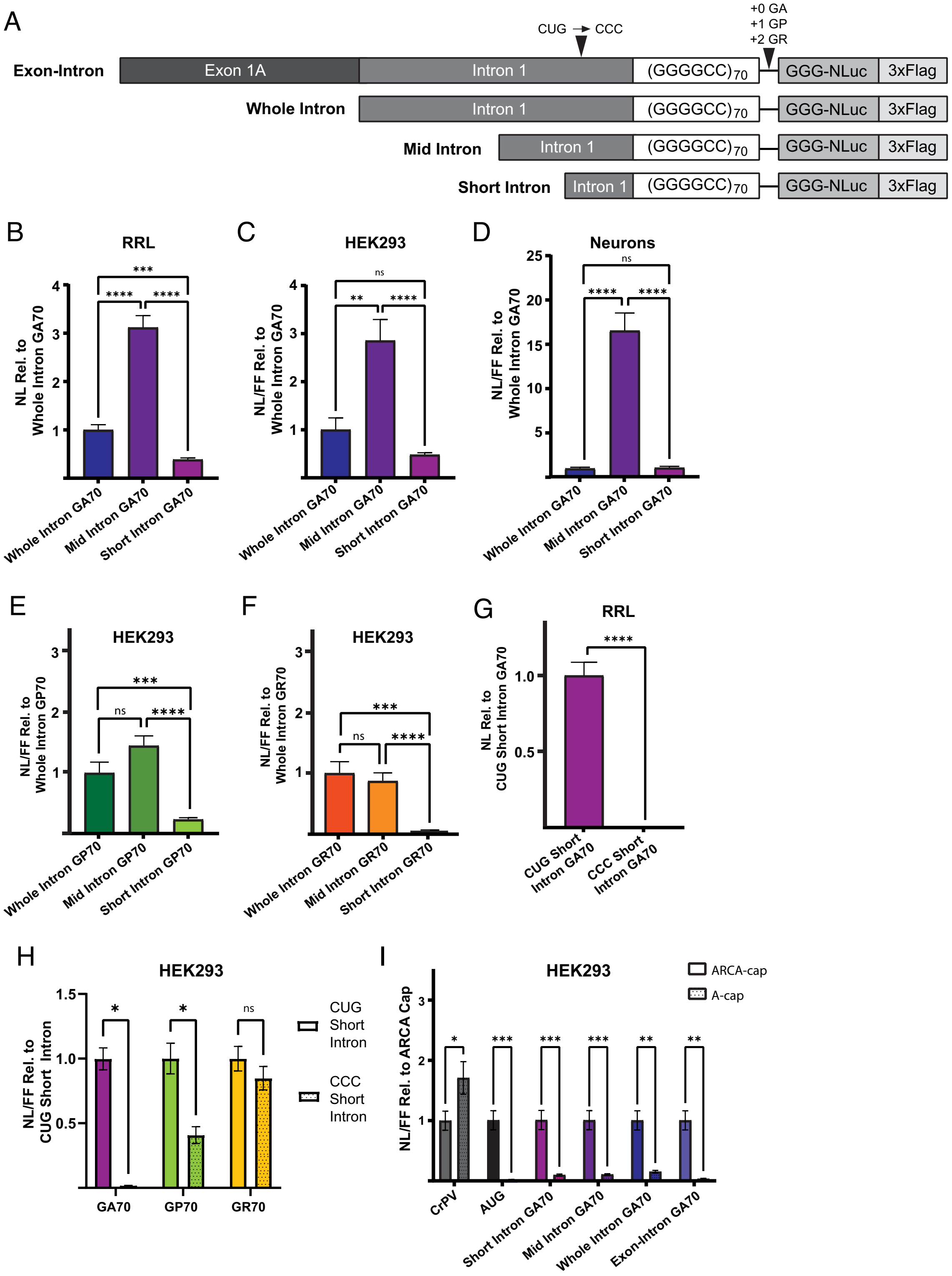

Fig. 3. Intronic-initiating reporter RNAs are translated in a cap dependent manner and more robustly than intron inclusion RNAs. (*A*) Schematic of linear nanoluciferase (NLuc) reporter mRNAs, including mRNA with short (32nt) leader. (*B*) Expression of short, mid, and whole intron GA70 reporters in rabbit reticulocyte lysate (RRL), expressed as NLuc (NL) relative to whole intron GA70, $n = 9$. (*C*) Expression of short, mid, and whole intron GA70 reporters in HEK293 cells, expressed as NL/FF relative to whole intron GA70, $n = 9$. (*D*) Expression of short, mid, and whole intron GA70 reporters in rat hippocampal neurons, expressed as NL/FF relative to whole intron GA70, $n = 15$. (*E*) Expression of short, mid, and whole intron GP70 reporters in HEK293 cells, expressed as NL/FF relative to whole intron GP70, $n = 9$. (*F*) Expression of short, mid, and whole intron GR70 reporters in HEK293 cells, expressed as NL/FF relative to whole intron GR70, $n = 9$. (*G*) Expression of short intron GA70 reporters in RRL, expressed as NL relative to CUG short intron GA70. polyGA frame CUG was mutated to CCC, $n = 9$. (*H*) Expression of short intron GA70, GP70, and GR70 reporters in HEK293 cells, expressed relative to CUG short intron for each frame. polyGA frame CUG was mutated to CCC, $n = 9$. (*I*) Expression of ARCA m^7G-capped and A-capped short intron, mid intron, whole intron, and exon–intron GA70 reporters in HEK293 cells, expressed as the ratio of NL/FF signal in A-capped reporters to m^7G-capped reporters, $n = 9$. Graphs represent mean + SEM, $*P < 0.05$, $***P < 0.001$, $****P < 0.0001$. 3xF = 3x FLAG tag; GA = glycine–alanine; GP = glycine–proline; GR = glycine–arginine; CrPV = cricket paralysis virus. (*B*–*F*) Brown–Forsythe one-way ANOVA with Welch's unpaired *t* test multiple comparison correction, (*G*) Two-tailed Student's *t* test with Welch's correction, (*H*–*I*) Multiple Holm–Šídák's two-tailed Student's *t* test with Welch's correction.

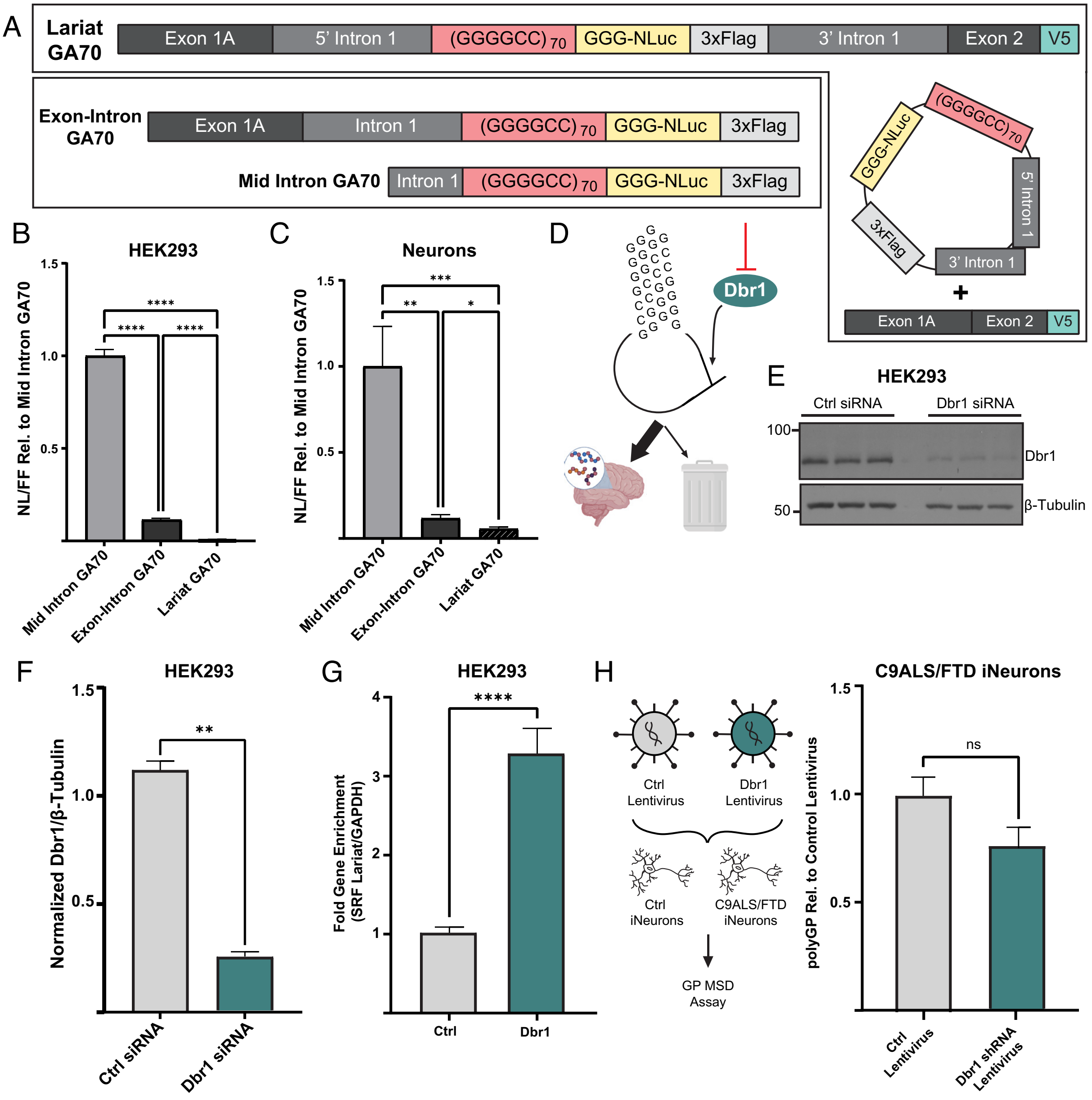

Fig. 4. Linear intron-initiating repeat RNAs drive C9RAN translation. (*A*) Schematic of mid intron GA70, exon–intron GA70, and C9 lariat (lariat GA70) nanoluciferase reporters. (*B*) Expression of mid intron, exon–intron, and lariat GA70 reporters in HEK293 cells expressed as NL/FF relative to mid intron GA70 reporter, *n* = 9. (*C*) Expression of mid intron, exon–intron, and lariat GA70 reporters in rat hippocampal neurons expressed as NL/FF relative to mid intron GA70 reporter, *n* = 15. (*D*) Schematic of predicted Dbr1 impact a C9 lariat RNA. (*E*) Western blot of Dbr1 protein levels in HEK293 cells transfected with nontargeting control or Dbr1 siRNA, 48 hours postknockdown (KD). Tubulin was used as a loading control, *n* = 3. (*F*) Quantification of western blot (*E*) of Dbr1 proteins levels in HEK293 cells transfected with nontargeting control or Dbr1 siRNA, 48-h post-KD. Expressed as relative intensity normalized to nontargeting control siRNA, *n* = 3. (*G*) Fold enrichment in gene expression of SRF lariat in HEK293 cells transfected with nontargeting control or Dbr1 siRNA, 48-hours post-KD. Normalized to GAPDH and expressed as relative to nontargeting control siRNA, *n* = 17. (*H*) Quantification of GP by MSD assay from C9ALS/FTD iNeurons treated with control or Dbr1 shRNA expressing lentivirus, expressed as arbitrary units relative to control shRNA lentivirus, *n* = 9. Graphs represent mean + SEM, **P* < 0.05; ****P* < 0.001; *****P* < 0.0001. GA = glycine–alanine; GP = glycine-proline. (*B* and *C*) Brown–Forsythe one-way ANOVA with Welch's unpaired *t* test multiple comparison correction (*F*–*H*) Two-tailed Student's *t* test with Welch's correction.

Dbr1 is responsible for the cleavage of the 2′ to 5′ phosphodiester bond formed at the branch point of circularized lariats. Upon cleavage, the circular lariat becomes linearized, and due to its lack of protective 5′ cap or 3′ polyA tail, is rapidly degraded by exonucleases (47–49). We hypothesized that inhibition of Dbr1 would inhibit the cell's ability to clear lariats, thus stabilizing these typically highly transient species, and that if a circularized lariat is the main driver of C9 RAN translation, expression of our lariat GA70 reporter would increase (Fig. 4*D*). To assess this, we took advantage of an endogenous lariat RNA that is stable across multiple human cell types (34). The serum response factor (SRF) lariat is detectable in HEK293 cells at baseline (*SI Appendix*, Fig. S4*C*), and knockdown of Dbr1 by siRNA in HEK293 cells led to a fourfold enrichment the SRF lariat RNA compared to a nontargeting control siRNA (*SI Appendix*, Table S4 and Fig. 4 *E*–*G*). As expected, knockdown of Dbr1 did not impact translation of linear GA70 reporters (*SI Appendix*, Fig. S4*D*). However, Dbr1 reduction had no impact on lariat GA70 elicited NLuc expression, despite a 75% reduction in Dbr1 at the protein level (Fig. 4 *E* and *F* and *SI Appendix*, Fig. S4*D*). While Dbr1 knockdown trended toward enhancement of lariat GA70 reporter-derived lariats as measured by RT-PCR across the branchpoint, this change was not significant due to high variance and low signal, suggesting that even reporter-based expanded repeat C9 lariats are difficult to stabilize in cells (*SI Appendix*, Table S3 and Fig. S4*E*).

We next assessed whether we could measure circular C9 intron lariats endogenously in patient-derived iNeurons. RNA was extracted from three control and three C9ALS/FTD patient-derived iNeuron lines and treated with RNaseR, an RNase that selectively degrades linear RNA while leaving circular RNA intact. This methodology significantly enhanced our detection of the SRF lariat compared to mock treated patient-derived iNeurons and significantly reduced the levels of linear mature *C9orf72*, as measured by exon2-exon3 levels (*SI Appendix*, Fig. S4*G*). However, we were unable to detect the presence of a circular C9 intron lariat despite allowing for 45 cycles of qRT-PCR. To confirm that our primers could bind their complementary sequence and generate a PCR product despite crossing a lariat branch point, we developed a reporter plasmid containing the unique branch point crossing sequence, which is only present when the intron is circularized in a lariat and arranged it linearly (linear C9 lariat). (*SI Appendix*, Fig. S4*G*). PCR of the branch point crossing primers on the linear C9 lariat reporter was successful at inputs down to 0.005 ng over 40 cycles (*SI Appendix*, Fig. S4*G*). This linearly arranged sequence was also readily detectable in transfected HEK293 cells (*SI Appendix*, Fig. S4*H*). The lack of readily detectable C9 lariat is in-line with previous work aimed at measuring the circular C9 intron lariat, which required multiple rounds of 40-cycle PCR to identify the circular C9 intron boundary (35).

We next sought to investigate whether reducing Dbr1 levels in patient-derived iNeurons might impact endogenous DPR levels. We hypothesized that if an endogenous C9 intron lariat with a large repeat were to act as a critical endogenous RAN translation template, then Dbr1 knockdown should increase the levels of toxic DPRs (Fig. 4*D*). We measured baseline polyGP levels in four C9ALS/FTD and four control patient-derived iNeuron lines, two of which were isogenic pairs (*SI Appendix*, Fig. S5*A*). One line expressed significantly higher levels of polyGP compared to the others tested (10-fold or greater), which we used for our Dbr1 studies moving forward (*SI Appendix*, Fig. S5*A*). Using a lentivirus expressing a Dbr1 shRNA (*SI Appendix*, Table S5), we transduced DIV3 C9ALS/FTD and control iNeurons and harvested cells at DIV10 to assess Dbr1 protein knockdown by western blot. Transduction with Dbr1 shRNA lentivirus reduced protein levels by 50% in control and C9ALS/FTD iNeurons (*SI Appendix*, Fig. S5 *B* and *C*). We next measured polyGP levels in C9ALS/FTD iNeurons treated with a Dbr1 shRNA lentivirus or a control lentivirus and saw no significant change in polyGP levels in C9ALS/FTD iNeurons (Fig. 4*H*). Taken together, these data suggest that a circular C9 intron lariat is unlikely to be an abundant or efficient template for RAN translation in patient-derived neurons.

Discussion

A central question in C9ALS/FTD pathogenesis is how an intronic repeat sequence is translated into toxic DPRs. Evidence suggests that RAN translation is important in disease pathogenesis and knowledge of its endogenous template in patient neurons could inform the development of therapeutics, such as antisense oligonucleotides (ASOs) and small molecules that target DPR production. Here, we show that the repeats are found in linear, 5′ m^7G-capped and sometimes polyadenylated mRNAs generated from transcriptional initiation within the intron itself in both a BAC transgenic mouse model of the disease and patient iNeurons. RAN translation of these cryptic mRNAs is highly efficient across all reading frames compared to lariat RNAs or retained exon-intron RNAs—both of which also are less abundant. Based on these findings, we propose that these intron-initiating repeat containing linear RNAs are a major endogenous template for RAN translation in C9ALS/FTD, with important implications for therapy development.

We utilized a relatively new approach that couples nanopore sequencing with 5′ RACE as opposed to traditional molecular cloning followed by Sanger sequencing of individual clones (Fig. 1*B*) (50). This methodology increased both the throughput and sensitivity of the assay to detect low frequency transcripts, thus generating a more accurate view of the transcriptional landscape of the 5′ end of C9orf72. However, the absence of allele-specific SNPs within the region of interest precludes our ability to distinguish whether the transcripts we identified in our patient-derived iNeurons derive from the expanded disease allele or the nonexpanded wildtype allele. While this cannot be resolved in our iNeuron model, the BAC transgenic C9 mouse model exclusively contains the expanded human allele, and insufficient sequence homology from the murine *C9orf72* gene prevents amplification by the *C9orf72*-specific 5′ RACE primers used in these experiments (*SI Appendix*, Fig. S1*A*). Thus, intron-initiated transcripts do occur in the presence of a large GGGGCC repeat (Fig. 1 *D*–*F*).

Identification of the same intron-initiated 5′ end transcripts by 5′ RACE of poly-adenylated mRNA suggests that, for at least a portion of these mRNAs, transcription proceeds through the repeat and to a polyA site (Fig. 2 *H*–*L*). While it is possible for pre-mRNA to be prematurely poly-adenylated, the stability of transcripts with this hallmark is markedly reduced. As such, our results are most consistent with the poly-adenylation of these transcripts occurring at nonpremature sites (51). This is consistent with recent reports of cytoplasmic repeat-containing RNA mapping to C9orf72 exons and suggests that intron-initiated transcripts may be "exonized" and contain the expected downstream exons (38). Further, "exonized" transcripts are known to be actively translated, as siRNAs against exon 2 in C9ALS/FTD fibroblasts exhibit significantly decreased abundance of polyGA and polyGP DPRs (38).

Importantly, we observe 5′ intron-initiated mRNAs within polysomes from C9 patient-derived iNeurons, suggesting that these transcripts undergo translation in a disease relevant cell type and contribute to DPR production (Fig. 2 *A* and *B*). While our nanopore-coupled 5′ RACE technique increased our ability to detect 5′ RACE products compared to traditional cloning and Sanger sequencing, we did not identify the shorter 32 nucleotide product seen in fibroblasts and iNeurons, likely due to a size-recovery limitation of PCR clean up. Over the past several years, long-read sequencing through repetitive regions has improved dramatically (52–55), and future 5′ RACE endeavors may be able to utilize primers that bind 3′ to the repeat to more reliably capture even very short products initiating just above the repeat.

In the classical scanning model of translational initiation, the 43S preinitiation complex made up of eIF2, GTP, Met-tRNA$_i^{Met}$, and 40S ribosome binds to capped mRNA through interaction with eIF4F, which includes the cap binding factor eIF4E (56). As such, steric hindrance induced by eIF4E is thought to preclude access of the 40S ribosome to the very 5′ end of capped mRNAs, creating a region of mRNA that is "blind" to the ribosome (57). Previous studies have suggested that a typical blind spot could range from 12 to 40 nucleotides (43, 44). Despite this, using in vitro transcribed mRNA nanoluciferase reporters that mimic the intron-initiated transcripts we identified through 5′ RACE, we observed robust translation from all three reading frames even when a very short leader was utilized that started just eight nucleotides above the CUG near-cognate codon previously shown to be important for polyGA reading frame translation (18, 19, 21, 25–27, 29, 46) (Fig. 3 *B–D* and *SI Appendix*, Fig. S2 *A–C*). Mutating this CUG codon to CCC markedly reduced production from this short intron reporter (Fig. 3 *G* and *H* and *SI Appendix*, Fig. S2 *E* and *F*), and its translation was cap-dependent (Fig. 3*I* and *SI Appendix*, Fig. S3 *A–E*). There are prior precedents for translation initiation at AUG codons within this typically "blind" region. Translation Initiator of Short 5′ UTR (TISU) is a short (twelve nucleotide median length) regulatory element located in 4.5% of protein-encoding genes. This element is typically close to the transcription start site and plays important roles in the translation of mRNAs in the absence of ribosome scanning (58). Translation of in vitro transcribed GFP reporter TISU mRNAs remained m^{7}G cap-dependent and was not significantly altered when the 5′ UTR was shortened to only five nucleotides, consistent with our short intron GA70 reporter findings (Fig. 3*I* and *SI Appendix*, Fig. S3 *A–E*). Similarly, histone H4 mRNAs also have short 5′ UTRs that utilize a hybrid cap-dependent scanning and IRES-driven initiation process (59). In this context, the histone ORF contains a highly structured helix region that directs positioning of the ribosome on the cognate start codon in the absence of ribosomal scanning (59). The GGGGCC repeat is known to adopt a strong secondary structure as a hairpin or g-quadruplex (60–63), suggesting it could act in a similar manner to support translation from a short leader sequence. However, translation from CUG and GUG codons in a short, 24nt UTR context has been previously reported in the absence of any such repeat structure (64). Interestingly, the CUG codon 24nts upstream of the repeat in *C9orf72* is in a relatively strong Kozak sequence, with both a purine at the −3 position and a G in the +4 position, both of which are thought to be strong predictors of efficient translation from non-AUG codons in mammals (65, 66). Future studies will be needed to formally assess the role of the repeat structure and leader sequence in RAN translation initiation.

The linear repeat-containing RNAs we observe in C9ALS/FTD patient neurons are much more efficient templates for RAN translation than either retained intronic repeats or spliced lariats that contain repeats. Previous data using a splice-capable C9 intron lariat reporter demonstrate that a spliced, circular intron can be exported to the cytoplasm and serves as the template for RAN translation (35). This work primarily utilized a reporter containing highly structured exogenous sequences (MS2 and PP7 binding sites) to track the mRNA, which could potentially influence the behavior of these reporters (67). Therefore, we sought to determine the contribution of a circular C9 intron lariat from a reporter lacking these sequences. Consistent with previous findings where the percentage of translating introns as measured by SunTag signal was markedly reduced compared to AUG-dependent translation (35), translation of our lariat reporter is inefficient in both HEK293 and rat hippocampal neurons (Fig. 4 *B* and *C*). This is likely due to inefficient, cap-independent translation mechanism by which a lariat would have to engage with ribosomes. Consistent with this theory, RAN translation of intron 1 is also inefficient when included as part of a bicistronic reporter system, where translation of a monocistronic reporter was ~20 to 30 fold higher than with a bicistronic reporter (22).

We were unable to measure a repeat-containing intron lariat in C9ALS/FTD patient-derived iNeurons despite post hoc enrichment for lariats by treating total RNA with RNaseR and assaying to 45 cycles by qRT-PCR (*SI Appendix*, Fig. S4*F*). Importantly, this technique did significantly enrich for a known stable lariat (SRF). Prior studies required deep sequencing following two rounds of PCR to reliably detect the branch point junction on the C9 lariat (35), suggesting that endogenous C9 intron lariats are maintained at a very low abundance within patient cells. To counter that limitation, we enriched for lariats by using a Dbr1 shRNA lentivirus in C9ALS/FTD patient-derived iNeurons. However, this strategy had no impact on endogenous DPR production (Fig. 4*H* and *SI Appendix*, Fig. S5 *B–D*). While Dbr1 knockdown did not impact polyGP production, future analysis of additional reading frames including polyGA and GR could reveal frame-specific effects.

In summary, our data demonstrate that intron-initiated repeat containing transcripts are more efficiently translated than either exon–intron or spliced lariat reporters and are more abundant in patient-derived iNeurons. These findings suggest that intron-initiated mRNAs represent the major endogenous RAN translation template in C9orf72-associated ALS and FTD. This finding has significant therapeutic implications, as some, but not all, currently designed ASOs and other therapeutic approaches target this species. We propose that these cryptic intronic initiation sites could be viewed as unique therapeutic targets. Future studies will be needed to define what approach would be most effective.

Methods

5′ Repeat-RLM-RACE. 5′ RACE was performed using the GeneRacer kit (Invitrogen, L150201) according to kit protocol. Total RNA from HEK293 cells and iNeurons was harvested using the RNeasy Mini Kit (Qiagen, 74106) according to manufacturer protocol. RNA was extracted from 6-wk-old mouse cerebellum using TRIzol (Thermo Fisher, 15596026) according to manufacturer protocol. For mouse experiments, 2 to 3 μg of DNase-treated total RNA was used for the initial CIP reaction. For iNeuron experiments, 1ug of DNase-treated total RNA, 40 ng of DNase-treated polyA captured RNA, or 0.2 μg of monosome/polysome fraction DNase-treated RNA was used for the initial CIP reaction. For fibroblast experiments, ~2.5 to 4.5 μg of DNase-treated total RNA was used in the initial CIP reaction. 1 pmol of C_4G_2 × 4 was spiked into the first step of cDNA synthesis, and cycling parameters were as follows (5 min at 25 °C, 60 min at 55 °C, 15 min at 70 °C). Control reactions with beta-actin primers were performed as per kit protocol. All PCRs utilized Platinum PCR SuperMix High Fidelity (Thermo Fisher, 12532016) with 1uL cDNA and touchdown protocol recommended in the kit protocol. For P1 and P2 primers, annealing was performed at 68 °C with 25 s extension. For the P3 primer, annealing was performed at 66 °C with 25 s extension. PCR products were cleaned with DNA Clean and Concentrator Magbead Kit (Zymo Research) according to manufacturer instructions. Purified PCR products were used for nanopore long read sequencing.

polyA Capture. DIV10 iNeurons were harvested with the RNeasy Mini Kit (Qiagen), treated with TURBO DNase (Thermo Fisher, AM2238), and clean and concentrated with the RNA Clean and Concentrate-25 kit (Zymo Research, R1018) as described above. polyA isolation was performed on 2 μg of DNase-treated total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs, E7490L) according to the manufacturer's express protocol.

ONT Library Preparation of Amplicons and Analysis. Sequencing libraries consisting of amplicons were prepared using the ONT Native Barcoding kit 24 V14 (SQK-NBD114.24, ONT) as described here with the following modifications. 10uL of each amplified product was phosphorylated using 0.5μL of T4 PNK (M0201S, NEB) and 1.5μL 10x T4 DNA ligase buffer (B0202S, NEB) and 3μL of molecular biology grade water in a total of 15μL. The phosphorylated products were ligated in separate PCR tubes with different barcodes in a 20 μL reaction with 0.5μL Native Barcode (NB01-24), 1μL T4 DNA ligase (M0202L, NEB), 1μL of T4 DNA ligase buffer and 2.5 μL of water for 20 min at RT. 2μL of 0.5 M EDTA was added to each tube to stop the barcode ligation and all reactions were pooled into a single 1.5 mL microcentrifuge tube. The barcoded reactions were incubated with 1.1X CleanNGS beads (CNGS005, Bulldog Bio) for 10 min at RT with rotation. The samples were then placed on a magnet and washed twice with 700 μL of 80% ethanol. Following the washes, the pooled samples were eluted in 36 μL of water, and 1μL was used to quantify the DNA concentration on the Qubit. The Native Adapter (NA) was ligated to the samples in a 50 μL ligation reaction (5 μL T4 DNA ligase, 12.5 μL LNB, 5 μL NA) and rotated for 20 min at RT. Next, 1.1X CleanNGS beads were added and incubated for an additional 10 min at RT with rotation. The library was placed on a magnet and the supernatant was removed followed by two 150 μL washes with Small Fragment Buffer (SFB). After the final wash, the bead pellet was allowed to air dry for 30 s and the library was eluted in 16 μL EB, and 1 μL was used to quantify the DNA on the Qubit. The sequencing library was prepared with at least 500 ng of barcoded-adapted sample, 15 μL Sequencing Buffer (SB), 5 μL Library Beads (LIB), and

sequenced on a Flongle R10.4.1 flow cell following the ONT Flongle loading method. ONT sequencing was performed for up to 24 h using ONT Flongle flowcells. The data were base called with Dorado version 7.6.7 using the dna_r10.4.1 _ e8.2 _ 400bps_5khz_sup model and a minimum qscore of 10. Sequence alignment was performed using minimap2, aligning reads to the human reference assembly GRCh38. On-target reads for each barcode were filtered and counted by transcript length leveraging the Pysam module in custom Python software.

Lentivirus. A lentiviral shRNA plasmid against Dbr1 was purchased from Horizon Discovery. GFP control vector (pLLEV-GFP) was purchased from the University of Michigan Vector Core. Lentiviruses were packed at the University of Michigan Vector Core with HIV lentiviruses and then concentrated to 10x concentration in 10 mL of DMEM. Knockdown of Dbr1 was confirmed by western blot. To transduce iNeurons, media was removed from control and C9 iNeurons on DIV3, and a full media change with diluted control GFP lentivirus or Dbr1 shRNA lentivirus in B27 was applied. A half media change with B27 media was performed on DIV6, and iNeurons were harvested on DIV10 according to the protocol detailed below for GP MSD.

GP MSD Assay. From a 6-well plate, media from DIV10 iNeurons was removed manually, and cells were harvested by scraping with 200 μL cold co-IP buffer (50 mM Tris-HCl, 300 mM NaCl, 5 mM EDTA, 0.1% triton-X 100, 2% SDS, protease inhibitor, phosphoSTOP) on ice. Lysates were passed through a 21G syringe 12 times, spun at 16,000 × *g* for 20 min at 15 °C, and supernatant was collected. polyGP protein levels were measured using the Meso Scale Discovery (MSD) electrochemiluminescence detection technology as previously described (8). Detection was performed as described in ref. 5. Additional information is detailed in the *SI Appendix, Materials and Methods*.

Data, Materials, and Software Availability. Most data are included in the manuscript and supplemental appendices, including sequences of identified 5′ RACE products. Aligned BAMs are available on Deep Blue Data (https://doi.org/10.7302/bmsz-v233) (68). Code for custom Python software used for nanopore based 5′ RACE transcript length determination using the Pysam module is available on GitHub at https://github.com/bcrone/5-RACE (69). All study data are included in the article and/or *SI Appendix*.

ACKNOWLEDGMENTS. This work was supported by NIH (F31127371 to S.L.M.; F31NS100302 to K.M.G.; R01NS097542, P01NS08497, and R35NS097273 to L.P.; R01NS113943 and 1R56NS128110-01 to S.J.B.; R21HG011493 and R01GM144484 to A.P.B.; R01NS099280 to P.K.T.; P30AG072931 to the University of Michigan Brain Bank and Alzheimer's Disease Research Center), The Mayo Clinic Foundation to L.P., Target ALS (Amyotrophic Lateral Sclerosis) Foundation to L.P., Robert Packard Center for ALS Research at Johns Hopkins University to L.P., S.J.B., and P.K.T., A. Alfred Taubman Medical Research Institute at the University of Michigan to A.P.B. and P.K.T., and the Veteran's Administration (VA BLRD BX004842 to P.K.T.). We acknowledge support from the Bioinformatics Core of the University of Michigan Medical School's Biomedical Research Core Facilities (RRID:SCR_019168), and from the Vector Core of the University of Michigan Medical School's Biomedical Research Core Facilities for packaging lentivirus for this study.

Author affiliations: [a]Department of Neurology, University of Michigan, Ann Arbor, MI 48109; [b]Cellular and Molecular Biology Graduate Program, University of Michigan, Ann Arbor, MI 48109; [c]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [d]Ann Arbor Veterans Administration Healthcare, Ann Arbor, MI 48109; [e]Department of Neuroscience, Mayo Clinic, Jacksonville, FL 32224; [f]Neuroscience Graduate Program, University of Michigan, Ann Arbor, MI 48109; [g]Medical Scientist Training Program, University of Michigan Medical School, University of Michigan, Ann Arbor, MI48109; and [h]Department of Human Genetics, University of Michigan, Ann Arbor, MI 48109

Author contributions: S.L.M., K.M.G., A.P.B., and P.K.T. designed research; S.L.M., K.M.G., B.C., J.A.S., E.M.H.T., A.K., K.J.-W., C.M.W., and E.W.J. performed research; S.L.M., K.M.G., B.C., J.A.S., A.K., K.J.-W., L.P., S.J.B., and A.P.B. contributed new reagents/analytic tools; S.L.M., K.M.G., B.C., E.M.H.T., K.J.-W., S.J.B., A.P.B., and P.K.T. analyzed data; and S.L.M., K.M.G., and P.K.T. wrote the paper.

1. A. E. Renton *et al.*, A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
2. M. DeJesus-Hernandez *et al.*, Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. *Neuron* **72**, 245–256 (2011).
3. P. Masrori, P. Van Damme, Amyotrophic lateral sclerosis: A clinical review. *Eur. J. Neurol.* **27**, 1918–1929 (2020).
4. E. Majounie *et al.*, Frequency of the C9orf72 hexanucleotide repeat expansion in patients with amyotrophic lateral sclerosis and frontotemporal dementia: A cross-sectional study. *Lancet Neurol.* **11**, 323–330 (2012).
5. K. Mori *et al.*, Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. *Acta Neuropathol. (Berl.)* **126**, 881–893 (2013).
6. K. Mori *et al.*, The C9orf72 GGGGCC repeat is translated into aggregating dipeptide-repeat proteins in FTLD/ALS. *Science* **339**, 1335–1338 (2013).
7. P. E. A. Ash *et al.*, Unconventional translation of C9ORF72 GGGGCC expansion generates insoluble polypeptides specific to c9FTD/ALS. *Neuron* **77**, 639–646 (2013).
8. T. Zu *et al.*, RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E4968–E4977 (2013).
9. Y. Sonobe *et al.*, Translation of dipeptide repeat proteins in C9ORF72 ALS/FTD through unique and redundant AUG initiation codons. *eLife* **12**, e83189 (2023).
10. S. Mizielinska *et al.*, C9orf72 repeat expansions cause neurodegeneration in Drosophila through arginine-rich proteins. *Science* **345**, 1192–1194 (2014).
11. X. Wen *et al.*, Antisense proline-arginine ran dipeptides linked to C9ORF72-ALS/FTD form toxic nuclear aggregates that initiate *in vitro* and *in vivo* neuronal death. *Neuron* **84**, 1213–1225 (2014).
12. A. Jovičić *et al.*, Modifiers of C9orf72 dipeptide repeat toxicity connect nucleocytoplasmic transport defects to FTD/ALS. *Nat. Neurosci.* **18**, 1226–1229 (2015).
13. B. D. Freibaum *et al.*, GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. *Nature* **525**, 129–133 (2015).
14. Y.-J. Zhang *et al.*, Poly(GR) impairs protein translation and stress granule dynamics in C9orf72-associated frontotemporal dementia and amyotrophic lateral sclerosis. *Nat. Med.* **24**, 1136–1142 (2018).
15. Z. Hao *et al.*, Motor dysfunction and neurodegeneration in a C9orf72 mouse line expressing poly-PR. *Nat. Commun.* **10**, 2906 (2019).
16. Y.-J. Zhang *et al.*, Aggregation-prone c9FTD/ALS poly(GA) RAN-translated proteins cause neurotoxicity by inducing ER stress. *Acta Neuropathol. (Berl.)* **128**, 505–524 (2014).
17. S. May *et al.*, C9orf72 FTLD/ALS-associated Gly-Ala dipeptide repeat proteins cause neuronal toxicity and Unc119 sequestration. *Acta Neuropathol. (Berl.)* **128**, 485–503 (2014).
18. K. M. Green *et al.*, RAN translation at C9orf72-associated repeat expansions is selectively enhanced by the integrated stress response. *Nat. Commun.* **8**, 2005 (2017).
19. R. Tabet *et al.*, CUG initiation and frameshifting enable production of dipeptide repeat proteins from ALS/FTD C9ORF72 transcripts. *Nat. Commun.* **9**, 152 (2018).
20. T. Westergard *et al.*, Repeat-associated non-AUG translation in C9orf72-ALS/FTD is driven by neuronal excitation and stress. *EMBO Mol. Med.* **11**, e9423 (2019).
21. Y. Sonobe *et al.*, Translation of dipeptide repeat proteins from the C9ORF72 expanded repeat is associated with cellular stress. *Neurobiol. Dis.* **116**, 155–165 (2018).
22. W. Cheng *et al.*, C9ORF72 GGGGCC repeat-associated non-AUG translation is upregulated by stress through eIF2α phosphorylation. *Nat. Commun.* **9**, 51 (2018).
23. F. He *et al.*, The carboxyl termini of RAN translated GGGGCC nucleotide repeat expansions modulate toxicity in models of ALS/FTD. *Acta Neuropathol. Commun.* **8**, 122 (2020).
24. Y. Sonobe *et al.*, A C. elegans model of C9orf72-associated ALS/FTD uncovers a conserved role for eIF2D in RAN translation. *Nat. Commun.* **12**, 6025 (2021).
25. S. Almeida *et al.*, Production of poly(GA) in C9ORF72 patient motor neurons derived from induced pluripotent stem cells. *Acta Neuropathol. (Berl.)* **138**, 1099–1101 (2019).
26. M. Boivin *et al.*, Reduced autophagy upon C9ORF72 loss synergizes with dipeptide repeat protein toxicity in G4C2 repeat expansion disorders. *EMBO J.* **39**, e100574 (2020).
27. A. Lampasona, S. Almeida, F.-B. Gao, Translation of the poly(GR) frame in C9ORF72-ALS/FTD is regulated by cis-elements involved in alternative splicing. *Neurobiol. Aging* **105**, 327–332 (2021).
28. H. M. Van't Spijker, S. Almeida, How villains are made: The translation of dipeptide repeat proteins in C9ORF72-ALS/FTD. *Gene* **858**, 147167 (2023).
29. S. Gotoh *et al.*, *EIF5* stimulates the CUG initiation of RAN translation of poly-GA dipeptide repeat protein (DPR) in C9orf72 FTLD/ALS. *J. Biol. Chem.* **300**, 105703 (2024).
30. H. Ito *et al.*, Reconstitution of C9orf72 GGGGCC repeat-associated non-AUG translation with purified human translation factors. *Sci. Rep.* **13**, 22826 (2023).
31. T. Zu *et al.*, Metformin inhibits RAN translation through PKR pathway and mitigates disease in C9orf72 ALS/FTD mice. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 18591–18599 (2020).
32. M. Niblock *et al.*, Retention of hexanucleotide repeat-containing intron in C9orf72 mRNA: Implications for the pathogenesis of ALS/FTD. *Acta Neuropathol. Commun.* **4**, 18 (2016).
33. Ł.J. Sznajder *et al.*, Intron retention induced by microsatellite expansions as a disease biomarker. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4234–4239 (2018).
34. G. J. S. Talhouarne, J. G. Gall, Lariat intronic RNAs in the cytoplasm of vertebrate cells. *Proc. Natl. Acad. Sci.* **115**, E7970–E7977 (2018).
35. S. Wang *et al.*, Nuclear export and translation of circular repeat-containing intronic RNA in C9ORF72-ALS/FTD. *Nat. Commun.* **12**, 4908 (2021).

36. C. Wang *et al.*, Characterization of distinct circular RNA signatures in solid tumors. *Mol. Cancer* **21**, 63 (2022).
37. I. Legnini *et al.*, Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell* **66**, 22–37.e9 (2017).
38. S. Yang *et al.*, Aberrant splicing exonizes C9ORF72 repeat expansion in ALS/FTD. bioRxiv [Preprint] (2023). 10.1101/2023.11.13.566896.
39. Y. Liu *et al.*, C9orf72 BAC mouse model with motor deficits and neurodegenerative features of ALS/FTD. *Neuron* **90**, 521–534 (2016).
40. L. Nguyen *et al.*, Survival and motor phenotypes in FVB C9-500 ALS/FTD BAC transgenic mice reproduced by multiple labs. *Neuron* **108**, 784–796.e3 (2020).
41. M. G. Kearse *et al.*, CGG repeat-associated non-AUG translation utilizes a cap-dependent scanning mechanism of initiation to produce toxic proteins. *Mol. Cell* **62**, 314–322 (2016).
42. N. M. Kaye, K. J. Emmett, W. C. Merrick, E. Jankowsky, Intrinsic RNA binding by the eukaryotic initiation factor 4F depends on a minimal RNA length but not on the m7G cap. *J. Biol. Chem.* **284**, 17742–17750 (2009).
43. Y. Gu, Y. Mao, L. Jia, L. Dong, S.-B. Qian, Bi-directional ribosome scanning controls the stringency of start codon selection. *Nat. Commun.* **12**, 6604 (2021).
44. J. Brito Querido *et al.*, Structure of a human 48S translational initiation complex. *Science* **369**, 1220–1227 (2020).
45. I. S. Fernández, X.-C. Bai, G. Murshudov, S. H. W. Scheres, V. Ramakrishnan, Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).
46. H. M. Van't Spijker, *et al.*, Ribosome profiling reveals novel regulation of C9ORF72 GGGGCC repeat-containing RNA translation. *RNA N. Y.* **28**, 123–138 (2022).
47. S. L. Ooi *et al.*, *RNA Lariat Debranching Enzyme Methods in Enzymology* (Academic Press Inc., 2001), pp. 233–248.
48. N. E. Clark *et al.*, Metal content and kinetic properties of yeast RNA lariat debranching enzyme Dbr1. *RNA N. Y. N* **28**, 927–936 (2022).
49. L. Buerer *et al.*, The debranching enzyme Dbr1 regulates lariat turnover and intron splicing. *Nat. Commun.* **15**, 4617 (2024).
50. P. G. Adamopoulos, P. Tsiakanikas, I. Stolidi, A. Scorilas, A versatile 5′ RACE-Seq methodology for the accurate identification of the 5′ termini of mRNAs. *BMC Genomics* **23**, 163 (2022).
51. L. A. Passmore, J. Coller, Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. *Nat. Rev. Mol. Cell Biol.* **23**, 93–106 (2022).
52. M. T. W. Ebbert *et al.*, Long-read sequencing across the C9orf72 'GGGGCC' repeat expansion: Implications for clinical use and genetic discovery efforts in human disease. *Mol. Neurodegener.* **13**, 46 (2018).
53. S. Miyatake *et al.*, Rapid and comprehensive diagnostic method for repeat expansion diseases using nanopore sequencing. *NPJ Genomic Med.* **7**, 62 (2022).
54. S. E. Salomonsson *et al.*, Validated assays for the quantification of C9orf72 human pathology. *Sci. Rep.* **14**, 828 (2024).
55. Y.-C. Tsai, K. A. Brown, M. T. Bernardi, J. Harting, C. D. Clelland, Single-molecule sequencing of the C9orf72 repeat expansion in patient iPSCs. *Bio-Protoc.* **14**, e5060 (2024).
56. R. J. Jackson, C. U. T. Hellen, T. V. Pestova, The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.* **11**, 113–127 (2010).
57. P. Kumar, C. U. T. Hellen, T. V. Pestova, Toward the mechanism of eIF4F-mediated ribosomal attachment to mammalian capped mRNAs. *Genes Dev.* **30**, 1573–1588 (2016).
58. R. Elfakess, R. Dikstein, A translation initiation element specific to mRNAs with very short 5′UTR that also regulates transcription. *PLoS ONE* **3**, e3094 (2008).
59. F. Martin *et al.*, Cap-assisted internal initiation of translation of histone H4. *Mol. Cell* **41**, 197–209 (2011).
60. A. Ursu *et al.*, Structural features of small molecules targeting the RNA repeat expansion that causes genetically defined ALS/FTD. *ACS Chem. Biol.* **15**, 3112–3123 (2020).
61. Y.-J. Tseng *et al.*, The RNA helicase DHX36-G4R1 modulates C9orf72 GGGGCC hexanucleotide repeat-associated translation. *J. Biol. Chem.* **297**, 100914 (2021).
62. F. Raguseo *et al.*, The ALS/FTD-related C9orf72 hexanucleotide repeat expansion forms RNA condensates through multimolecular G-quadruplexes. *Nat. Commun.* **14**, 8272 (2023).
63. A. Taghavi, J. T. Baisden, J. L. Childs-Disney, I. Yildirim, M. D. Disney, Conformational dynamics of RNA G4C2 and G2C4 repeat expansions causing ALS/FTD using NMR and molecular dynamics studies. *Nucleic Acids Res.* **51**, 5325–5340 (2023).
64. L. Tang *et al.*, Competition between translation initiation factor eIF5 and its mimic protein 5MP determines non-AUG initiation rate genome-wide. *Nucleic Acids Res.* **45**, 11941–11953 (2017).
65. S. J. Grzegorski, E. F. Chiari, A. Robbins, P. E. Kish, A. Kahana, Natural variability of Kozak sequences correlates with function in a zebrafish model. *PLoS ONE* **9**, e108475 (2014).
66. C. Ambrosini *et al.*, Translational enhancement by base editing of the Kozak sequence rescues haploinsufficiency. *Nucleic Acids Res.* **50**, 10756–10771 (2022).
67. S. Heinrich, C. L. Sidler, C. M. Azzalin, K. Weis, Stem-loop RNA labeling can affect nuclear and cytoplasmic mRNA processing. *RNA N. Y. N* **23**, 134–141 (2017).
68. P. K. Todd, S. L. Miller, B. Crone, A. P. Boyle, BAM files related to "Cryptic intronic transcriptional initiation generates efficient endogenous mRNA templates for C9orf72-associated RAN translation". Deep Blue Data. https://doi.org/10.7302/bmsz-v233. Deposited 24 June 2025.
69. B. Crone, 5′ Race. Github. https://github.com/bcrone/5-race. Deposited 26 June 2025.