RNA-Mediated Toxicity in Neurodegeneration:

The Mechanistic Role of the C9ORF72 Repeat Expansion in ALS Molecular Pathogenesis

Thesis by

Paulomi Bhattacharya

In Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY

Pasadena, California

2024

(Defended August 15th, 2024)

ACKNOWLEDGEMENTS

I am extremely grateful for my PhD experience at Caltech over the last several years. I have had the opportunity to collaborate with talented and passionate people while simultaneously engaging in cutting-edge scientific research, often beyond the realm of what I thought to be possible. I decided many years ago that I wanted to focus my scientific career on tackling neurodegenerative disease, and I am thankful to have taken steps towards this goal from the lens of an interdisciplinary genomicist. The intersection of these fields is a space filled with promise and potential, and I look forward to making further progress in coming decades.

First of all, I would like to thank my advisor, Mitch Guttman, for his guidance and mentorship throughout this PhD. I still vividly remember my first interactions with Mitch; we spoke about Thomas Kuhn's *The Structure of Scientific Revolutions* and engaged in a lengthy philosophical discussion on what it takes to shift paradigms in science. His energy and gutsiness in tackling problems was (and still is) infectious; in that first meeting, I remember already feeling that I may have found an advisor who (1) reflected some of my favorite personality traits in scientists as a broader personality type, and (2) shared many of my own scientific values. As Mitch and I continued to work together during my rotation and beyond, everything that I had initially suspected turned out to be true. He has encouraged me to be bold and brave in my approach, while staying rooted in scientific rigor. I am grateful to have trained under Mitch.

Next, I would like to extend the most heartfelt thank you to Mario Blanco. Many Guttman lab members joke that Mario is the glue that holds our lab together. I continue to be inspired by his self-sacrificing nature, incredible patience in training all of us, and deep scientific expertise. I have learned so much from Mario over the years, from experimental tools and techniques in RNA and protein biology, to broader approaches in critical thinking about complex scientific problems. I could not be more thankful for Mario's mentorship; I would not have gotten through my PhD without him.

Next, I would like to acknowledge and thank all members of my thesis committee – Mikhail Shapiro, Justin Ichida, Henry Lester, and Matt Thomson – who have provided valuable guidance over the years. In particular, I would like to thank Justin for collaborating with us on the C9 manuscript and choosing to accompany us in this undertaking. We are grateful to have his expertise and support. Along with Justin, I would like to take the opportunity to thank Jasper Rubin-Sigler for his work in generating ALS patient cells for our manuscript. I would also like to thank Shawna Hiley for her editing work but also for her support throughout the paper-writing process to help us stay focused and on track. Thank you to Inna Marie-Strahznik, for her beautiful illustrations that always succeed in explaining difficult concepts easily and aesthetically.

Next, I would like to thank my friends in the lab. In particular, I would like to thank Linlin, Prashant, Andrew, Drew, and Jimmy, who have provided support in countless

forms over the years -- from pep talks in the tissue culture room, to making sure that I was still eating and sleeping during tough times, to endless jokes and laughter over good food. I would also like to thank my LA friends outside of the lab who have provided so much support over the course of graduate school. Finally, I would like to thank my friends outside LA (New York, Boston, LA, Europe etc) for their support and companionship, even when they did not fully understand what I had been up to for so many years in the lab. You all know who you are, and I am grateful to have such a diverse group of friends in my life.

On a final, more personal note: I would like to thank my family. My mother has made endless sacrifices for my education, career, and future success, and my father has always encouraged me to strive to make the world a better place. But perhaps most importantly, my parents have instilled in me the value of integrity, honor, and humility. I will always carry this with me, wherever life takes me next. I would also like to thank both my grandparents, who have always been my greatest cheerleaders. And finally, I would like to thank my most loyal companion -- my dog Bubbles – who has been sitting next to me while I study for the last 15 years and counting. Thank you all.

PUBLISHED CONTENT AND CONTRIBUTIONS

1. Bhattacharya P., Blanco, M.R., Rubin-Sigler J., Ichida J.I.., and Guttman, M. (2024).

   The C9ORF72 Repeat Expansion Is Not Detectably Transcribed And May Instead Function As An Epigenetic Regulatory Element.

   (in review at *Nature Medicine*)

P.B. conceived of this project with M.R.B. and M.G., led the effort to develop, generate data, analyze data, generate figures, and write the manuscript.

2. Differential RNA-Protein Binding Maps Reveal Convergent Mechanisms of Splicing Dysregulation Across Genetic Subtypes of ALS.

(in preparation) P.B. conceived of this project with M.G., is leading the effort to develop, generate data, analyze data, generate figures, and write the manuscript.

TABLE OF CONTENTS

In loving memory of my grandfather, Dadai, who has always encouraged me to live life with curiosity,
pursue academic excellence and rigor,
and stay grounded in honesty, grace, and humility.

This is for you.

Chapter 1

ALS AT A GLANCE

Paulomi Bhattacharya

The following section contains unpublished data *before* we reframed my PhD thesis upon discovering that the C9-repeat RNA is not transcribed.

1.1 INTRODUCTION TO ALS AND PRELIMINARY RESULTS

1.1.1 PREVALENCE AND CLINICAL PRESENTATION

Amyotrophic lateral sclerosis (ALS) is amongst the most rapidly progressing neurodegenerative diseases, characterized by the degeneration of upper and lower neurons in the brain and spinal cord[1-3]. As the most commonly occurring disease affecting motor neurons, ALS poses an estimated lifetime risk of 1 in 350 and an average prognosis of 2-5 years, with patients ultimately succumbing to neuromuscular respiratory failure[4]. However, despite the devastating nature of its clinical progression, ALS is irreversible and incurable as the underlying etiology governing the disease still remains unknown[5].

A key challenge and source of complexity in the fight against ALS has been the heterogeneity of its presentation, from the molecular scale in the affected neuron to the clinical presentation of symptoms in patients[5]. Historically, understanding progression on the molecular level has been a challenge in investigating diseases of the brain. Until the recent development of technologies that now allow for the study of subcellular pathogenesis over time, most prior work had been limited to patient postmortem brain tissue[5-6]. As a result, the clinical presentation of ALS is arguably better characterized on a temporal scale compared to its molecular pathogenesis. At time of onset, most patients are aged between 40 and 70, while more rarely, juvenile onset cases have been reported in patients under the age of 25[1-5]. Notably, early symptoms in patients are dependent on the site of onset. The initial site of onset in patients can be the arms and legs (limb or

spinal onset), the muscles responsible for swallowing and speech (bulbar onset), or the lungs (respiratory onset)[1-5]. Therefore, early symptoms often include muscle weakness and/or spasticity, slurred speech, and difficulty breathing or swallowing[1]. As some of these symptoms can be attributed to everyday fatigue or clumsiness, time to initial diagnosis of ALS is often delayed. There is no single test currently used to diagnose the condition, so the broader diagnostic strategy includes a combination of medical history, cognitive and physical exams, electrodiagnostic tests, and lab tests to detect biomarkers in blood, urine, or spinal fluid[1-3]. However, the rapid progression of ALS and the difficulty of early detection can mean that by the time of diagnosis, the disease has progressed to a stage where preserving and slowing motor neuron damage is a significant challenge for medical professionals. Once early symptoms set in, patients experience a rapid deterioration of cognitive and motor function that culminates in a loss of the ability to swallow or breathe, quickly leading to death[1-5].

1.1.2 KEY MOLECULAR HALLMARKS OF ALS PATHOPHYSIOLOGY

The rapid rate of clinical deterioration in ALS patients provides an urgent reminder that the development of early and effective therapeutic intervention ultimately depends on our understanding of the underlying molecular etiology: specifically, an identification of the molecular phenotype(s) that are causal to neuronal death.

From a technological standpoint, the mechanistic understanding of dysregulated cellular processes in affected neurons has been recently accelerated by the development of patient-derived induced pluripotent stem cells converted to neurons (patient-derived iPSNs)[6]. In this model, fibroblasts from ALS patients are reprogrammed to induced pluripotent stem cells (iPSCs), which are then differentiated to motor neurons or cortical neurons using specific sets of transcription factors (Fig. 1A). As a disease model, iPSNs possess several key technical advantages[5-6]: (1) they are genetically identical to the affected patient, (2) they have been shown to recapitulate neuronal features as well as ALS-specific phenotypes consistent with patient postmortem cells, (3) they are relatively easy to grow and maintain at scale, and perhaps most critically, (4) they allow for the study of disease progression over time[6]. In the past two decades, the application of biochemical tools and techniques in iPSN models combined with validation of findings in postmortem brain tissue has proved to be a powerful two-pronged strategy to identify some of the aberrant processes in ALS on the DNA, RNA, and protein level[5-6].

Protein Mislocalization and Aggregation. Similar to many other neurodegenerative diseases including Alzheimer's disease, Huntington's disease, and Parkinson's disease, ALS is often characterized by the mislocalization and aggregation of proteins involved in a range of processes crucial to neuronal homeostasis[7]. One of the most important examples is TDP43, a protein that binds DNA and RNA and is critical for DNA repair, transcription, and splicing regulation in the cell[9-15]. While TDP43 carries out its wild-type function as a regulator of transcription and splicing in the nucleus, confocal imaging of

TDP43 in ALS neurons has revealed a nuclear clearance of TDP43 and its aggregation in the cytoplasm (Fig. 1B-C)[8-9,10,13]. As the concentration of TDP43 molecules sequestered in the cytoplasm increases, these molecules localize as insoluble aggregates, which are known to be a defining clinical hallmark of ALS as well as other neurodegenerative diseases (Fig. 1B-C)[1,10, 13,15-16]. Despite the heterogeneity observed in ALS pathophysiology across distinct genetic subtypes, ~97-99% of all ALS cases feature this mislocalization and cytoplasmic aggregation of TDP43[13]. Therefore, TDP43 pathology is considered to be a convergent molecular phenotype in ALS pathology across genetic backgrounds [4,13,15-16]. In addition to TDP43, another protein frequently mislocalized in ALS pathology is FUS[17]. FUS is also a powerhouse protein with a diverse set of regulatory functions; for instance, its binding interaction with the U1 spliceosomal RNA is critical for RNA splicing regulation and it is reported to couple transcription and splicing by mediating interaction between U1 and PolII[18]. Similar to TDP43, wild type FUS predominantly localizes to the nucleus in healthy cells[17-19]. However, affected ALS patient neurons feature the nuclear clearance and cytoplasmic mislocalization of FUS, which again leads to the formation of insoluble cytoplasmic aggregates visible by microscopy [7,19-20]. Taken together, TDP43 and FUS protein aggregates are considered to be pathological hallmarks of ALS, but other proteins (including the FET family of splicing regulators) are also sequestered in these cytoplasmic aggregates away from their endogenous targets, leaving them unavailable to carry out their wild type functional roles [21-22]. Therefore, even beyond TD43 and FUS, protein mislocalization and aggregation is a broader molecular phenotype observed in ALS patient cells.

Aberrant DNA Damage Repair. Another source of molecular dysregulation in ALS is the accumulation of DNA damage products paired with dysfunction in the surveillance pathways typically used to identify and repair these products at their source, known as the DNA damage response (DDR)[23-24]. In a healthy cell, DNA damage occurs as a natural byproduct of DNA replication and cell division, resulting in double-stranded breaks (DSBs) and single-stranded breaks (SSBs) which are then detected by DDR signal transduction pathways [23-24]. The repair of DSBs via the DDR consists of two main pathways: (1) non-homologous end joining (NHEJ) and (2) homologous repair (HR) (Fig. 1D)[23-24]. Because neurons as a cell type are post-mitotic and therefore nondividing, the primary DDR pathway used is NHEJ, which is known to be particularly error-prone [23-24]. Furthermore, ALS is an age-related disease in nondividing neurons where unrepaired DNA damage can accumulate over time, DDR has been increasingly implicated in ALS molecular pathophysiology in recent years [23-24]. Notably, many of the most frequently mutated proteins in ALS are implicated in DDR, including TDP43, FUS, SOD1, VCP, and NEK1[23-28]. For example, TDP43, which is the most commonly mutated protein in ALS as well as many other neurodegenerative disorders, is part of the NHEJ pathway, where it is recruited to DSB sites and recruits the XRCC-DNA ligase IV complex[25]. As TDP43 nuclear clearance and cytoplasmic mislocalization is observed in ~97% of ALS cases[13], this would prevent TDP43 from carrying out its endogenous nuclear function as a DDR regulator. Notably, TDP43 knockout cells feature a global impairment of NHEJ and an increase in DSBs[24-25], and an ALS neuron with TDP43 cleared from the nucleus would likely feature a similar phenotype. Consistent with this hypothesis, ALS cells with

mislocalized TDP43 have been shown to have increased accumulation of DNA damage products, including DSBs and R-loops[25].

RNA Splicing Dysregulation. Another critical cellular process that is aberrantly regulated in ALS is the alternative splicing of RNA[16]. As previously mentioned, the two most commonly mislocalized proteins in ALS are TDP43 and FUS, which notably both function as splicing regulators and interact with spliceosomal components [29-34]. Specifically, TDP43 is known to interact with small Cajal body specific RNAs, or scaRNAs[29], which are implicated in the biogenesis of small nuclear ribonucleoproteins (snRNPs) and localize to the Cajal body. Similarly, it is known that FUS interacts with the U1 spliceosomal RNA which comprises the U1 snRNP complex and is critical for assembly of the spliceosome[32]. In recent years, evidence of the dysregulation of these interactions (FUS-U1 and TDP43-scaRNA) has been shown in ALS patient neurons [29-34].

Other recent evidence of splicing dysregulation in ALS involves the inclusion of cryptic exons[35]. One key example is UNC13A, a gene critical for synaptic function and neuronal homeostasis[36]. In both patient-derived iPSNs as well as postmortem brain and spinal cord tissue, it has been demonstrated that TDP43 depletion leads to the inclusion of a cryptic exon in UNC13A, which triggers nonsense-mediated decay of the RNA and leads to loss of the UNC13A protein[35]. This finding directly links TDP43 nuclear depletion and more broadly, TDP43 proteinopathy, to a splicing-related molecular phenotype with critical consequences for disease pathology. Similarly, it has been shown that TDP43 binds to the

pre-mRNA of Stathmin-2 (STMN2), a neuron-specific microtubule protein implicated in axonal growth and regeneration [37-40]. Upon binding to STMN2 pre-mRNA, TDP43 prevents the inclusion of a cryptic exon (Exon2A) in the first intron. It has been shown in ALS patient cells that the nuclear loss of TDP43 results in the inclusion of cryptic Exon 2A, which includes an in-frame stop codon and thus results in the formation of a truncated mRNA product and loss of the STMN2 protein (Fig. 1E)[37-40]. Taken together, both of these key findings emphasize the causal relationship between mislocalized splicing regulators and the inclusion of cryptic exons which lead to the loss-of-function of proteins that are necessary for neuronal survival [35-41].

Finally, splicing dysregulation is observed in ALS patient cells through widespread intron retention events [42-43]. Specifically, studies in C9-ALS patient brains have reported approximately 2000 transcripts with splicing effects in intron retention[42]. Notably, these effects appeared to co-occur with the localization of the key splicing regulator HNRNP-H into insoluble aggregates visible by microscopy[42]. In addition, transcripts with retained introns were involved in pathways involved in protein clearance and quality control, processes that are mechanistically linked with the widespread protein mislocalization and aggregation often observed across most ALS cases[42]. Other notable studies have reported intron retention effects that correlate with the nuclear clearance of the splicing regulator SFPQ, a protein which binds to a retained intron in its own transcript in an autoregulatory manner[43].

Taken together, there are numerous lines of evidence that support the dysregulation of splicing as an aberrant process in ALS, though further investigation to understand causal contribution to neuronal death is still necessary.

Impaired Nucleocytoplasmic Transport. As protein mislocalization from the nucleus to the cytoplasm is a frequent ALS molecular phenotype (and in the case of TDP43, convergent across genetic subtypes), it is reasonable to consider the dysregulation of nucleocytoplasmic transport (NCT) as a mechanistic explanation [44-47]. Broadly, effective NCT is crucial for the regulation of cellular functions, because DNA, RNA, and protein molecules in distinct cellular pathways are organized spatially in specific subcellular regions or compartments in order to locate their interacting partners[44]. Briefly, NCT as a process is regulated by the nuclear pore complex (NPC), a protein complex which consists of 30 nucleoporin subunits[44]. While lower molecular weight cargo can pass through the nuclear pore via passive transport, active transport for larger cargo molecules is facilitated by the Ran gradient, powered by the hydrolysis of Ran GTPase[44].

One of the first groups to identify nucleocytoplasmic transport as an ALS phenotype conducted a genetic screen in Drosophila, where a set of nucleoporins and transport factors were categorized as suppressors or enhancers of the neurodegenerative phenotype, including all major components of the nuclear pore complex (Fig. 1F)[48]. Notably, this study demonstrated that overexpression or knockdown of these neurodegenerative phenotype ‘modifiers’ was sufficient to rescue or exacerbate the neurodegenerative

phenotype[48]. Since then, several other studies have reported evidence of transport deficits of particular cargo RNAs or proteins in ALS patient cells, even reporting global defects in protein import or RNA export, measured in aggregate[44,46-47]. In addition to the mislocalization of cargo molecules, some studies have even reported the mislocalization of NPC components[49]. One example is the mislocalization of RanGap1, which plays its endogenous role in the nucleus to maintain the Ran GTPase gradient responsible for NCT[49]. However, while there have been significant efforts to draw a causal link between NCT and other downstream neurodegenerative phenotypes that cause neuronal death in ALS, no such causality has been successfully established. However, NCT defects are indeed an aberrant cellular phenotype in ALS, along with other molecular hallmarks such as dysregulation of DDR, aberrant splicing, or protein aggregation.

1.1.3 THE GENETIC ETIOLOGY OF ALS

While there is a range of aberrant molecular processes that have been linked to ALS in a correlative manner, the specific phenotypes that are *causal* for neuronal death still remain unknown. Propelled by advances in genomics tools and techniques in recent decades, the identification of genetic mutations that are linked to or causal for ALS have led to increased efforts in understanding the genetic etiology of the disease[5,50-53].

Similar to subcellular phenotypes and clinical presentation of symptoms, the genetic etiology of ALS is also characterized by heterogeneity[50-53] (Fig. 1G). Approximately 10

percent of all cases of ALS are categorized as “familial” (fALS): having a genetic component and featuring transmission within families. The remaining are labeled “sporadic” (sALS): lacking a clear family history or defined inheritance pattern across multiple generations[50-53]. However, there are examples of genes causal to familial ALS that are also mutated in sporadic cases, as sporadic ALS cases often feature multiple pathogenic mutations[50-53].

To date, pathogenic variants in >30 ALS-associated genes have been implicated in familial ALS[5], and the proteins encoded by these genes play critical functional roles in the aforementioned aberrant molecular processes linked to ALS, including protein trafficking/localization/aggregation, DNA damage repair, RNA splicing, and nucleocytoplasmic transport[5,50-53]. The most frequently mutated genes in ALS are TDP43, FUS, SOD1, and C9ORF72, which together comprise approximately 60 percent of fALS and 10 percent of sALS cases[5] (Fig. 1G). Notably, FUS and TDP43 feature prominently on the protein level in ALS molecular etiology in patient neurons, so the genetic mutations encoding these proteins often provide mechanistic insight that can explain downstream molecular phenotypes. For instance, FUS-P525L and FUS-5R21C both feature mutations in their nuclear localization sequence (NLS), potentially explaining their nuclear clearance and cytoplasmic aggregation[5,50-53]. Similarly, TDP43-A382T is another NLS mutation where the mislocalization of this mutant TDP43 protein has been directly linked to R loop aggregation and accumulation of DNA damage in ALS patient derived cell lines[12].. While TDP43, FUS, SOD1, and C9ORF72 are the most commonly

mutated genes in ALS, several other genes have also been identified, including ATXN2, NEK1, TIA1, SETX, MATR3, and VCP[5,50-56]. While the broader understanding of the genetics of ALS has indeed progressed in recent years with the identification of novel mutations, the functional, mechanistic relevance of these mutations to explain their prevalence in ALS, is yet to be understood.

1.1.4 C9ORF72: A REPEAT EXPANSION IN NEURODEGENERATION

Of the >30 pathogenic mutations that have been implicated in ALS, one unique example is C9ORF72[57-59]. Specifically, C9ORF72 stands out as an exception as it is the most commonly occurring mutation linked to ALS, accounting for 40 percent of all fALS and 10 percent of sALS cases[57-59]. Moreover, C9ORF72 is the only known pathogenic mutation that is monogenic with a causal correlation to ALS; the presence of this mutation alone is sufficient to cause ALS in carriers[57-59].

Notably, C9ORF72 is a repeat expansion mutation in the first intron of the C9 gene, characterized by the 6-nucleotide repeat unit (G4C2)[57-59](Fig. 2A). Broadly, repeat expansion mutations are defined by a short sequence motif of DNA bases that is repeated an increased number of times compared to the wild-type context[60-64]. In the case of the G4C2 hexanucleotide repeat expansion (HRE) in C9ORF72, C9-ALS patients are heterozygous carriers of the HRE; the wild-type allele contains 3-20 repeats of the G4C2

motif, while the mutant allele features the expanded form of hundreds to thousands of G4C2 repeats[57-59,65](Fig. 2A).

In addition to C9ORF72, a number of neurodegenerative disorders, including Huntington's disease, Fragile X syndrome, myotonic dystrophy, Fredrich's ataxia, and frontotemporal dementia have been linked to repeat expansion mutations[60-64].. While the specific sequence of the repeat unit as well as the genetic localization (promoter, exon, intron, 5'UTR) of the insertion site may vary, it is generally the case that longer repeat lengths are associated with increased disease severity[65]. For instance, all repeat expansion-associated neurodegenerative diseases feature an expansion past some pathogenic threshold length, at which point the mutation leads to disease[60-65]. In the case of Huntington's disease, this pathogenic threshold length is well-defined at >35 repeats of the CAG motif, while in the case of Freidrich's ataxia the threshold is less strictly defined and patients often feature 600-1200 repeats[60-64].

However, the mechanism underlying this expansion of repeat sequences in genomic DNA is not well understood[61]. This phenomenon is labeled somatic instability, or repeat instability, wherein expanded repeats change size across generations of individuals, while the repeat length can also vary within a single individual across cell or tissue types. For instance, the repeat length in the blood of a mutation carrier could be significantly longer or shorter compared to repeat length in the brain[61]. This dynamic and unpredictable nature of repeat expansions results in significant phenotypic variability across generations of a

family[61]. Nevertheless, repeat expansions are typically characterized by clinical anticipation, where the severity of disease phenotypes increases with each generation[61].

Despite the mechanistic complexity and dynamic nature of repeat expansion insertions, the pathogenic contribution of some repeat expansion mutations in neurodegenerative disease have been successfully characterized, including but not limited to Fragile X syndrome or neuronal intranuclear inclusion disease[64]. However, this category of characterized and well-understood repeat expansions does *not* include the C9ORF72 HRE.

1.1.5 THE FUNCTIONAL ROLE OF THE C9 REPEAT EXPANSION

While the G4C2 HRE in the C9ORF72 gene is the most frequently occurring mutation ever linked to ALS, its role in molecular pathogenesis is still unknown[57-59,66].

The functional role of the C9 HRE has been debated since its discovery in 2011[57-59,66]. C9 ALS patients are heterozygous carriers of the HRE; in patient DNA, the wild-type allele contains 3-20 repeats of the G4C2 motif while the mutant allele can contain hundreds to thousands of repeats[57-59,66]. To date, it has been widely accepted that this G4C2 expansion mutation is transcribed in spinal and cortical neurons to produce a G4C2 expansion RNA in both iPSN models and postmortem patients, leading to the prevailing notion that the

presence of this G4C2 expansion RNA and its gain-of-function is the primary mechanism responsible for C9-ALS molecular pathogenesis[57-59,66-69].

The field has converged to two mechanistic explanations for the gain-of-function of the G4C2 HRE RNA: (1) its sequestration of RNA binding proteins (RBPs) away from their endogenous targets[57-59,66-70] and (2) its function as a template for the production of dipeptide repeat proteins (DPRs) by noncanonical RAN translation[57-59,66-70]. However, despite these proposals, the functional role of the G4C2 RNA still remains poorly understood. Additionally, the recent failure of several major clinical trials targeting either the C9 HRE RNA or the DPR proteins have added to the confusion and controversy surrounding the mechanistic role of this mutation in ALS[72].

Historically, a key challenge in the field has been that we lack robust methods to directly measure the C9 HRE, which is large in size and low in sequence complexity(>99 percent GC content)[57-59,66-70]. The HRE in the DNA of C9 patients is traditionally estimated by a combination of Southern blots and repeat-primed PCR[73], while more recent attempts have applied long-read sequencing technologies, namely through Pacific Biosciences (PacBio) and Oxford Nanopore Technologies[74]. On the other hand, the detection of the HRE in the C9 RNA has been especially challenging and has thus far relied primarily on RNA FISH and imaging[57-59,66-70]. However, there are several caveats to consider when using RNA FISH to detect the C9 HRE RNA foci. First, the sequence identity of FISH detection probes in previous work has been (G4C2)x3-4[57-59,66-70], and short oligonucleotide probes at this length

with low complexity sequences are known to bind nonspecifically to off-target molecules. Second, there are numerous other genes in the human genome (eg. RRP36, SULF2, HUWE1, RGS14) that contain 3-4 repeats of the G4C2 motif and are expressed at up to 10-fold higher levels when compared to the expression level of the 5'end end of the C9 RNA that flanks the G4C2 HRE. Without control probes that target other non-G4C2 regions of the C9 gene and quantification of colocalization with the (G4C2)x3-4 probes, it is not possible to conclude whether detected FISH foci represent the C9 HRE or these other more highly expressed non-C9 genes.

In addition to these methodological challenges, the frequency and location of reported C9 HRE RNA FISH foci further call into question their pathological relevance. Specifically, C9 HRE foci are often most frequently observed in patient cell types that are typically unaffected in ALS (e.g., cerebellum), and the percentage of cells reported with C9 HRE foci in ALS-affected brain regions or patient-derived iPSNs is extremely low, with some studies reporting 1-2 foci for only 10% of cells[57-59].Some studies have investigated the possible function of these HRE foci by investigating the RBPs that bind the C9 HRE[57-59,66-70], but these attempts have not been able to clarify the mechanistic role of the RNA in the context of relevant downstream ALS phenotypes. Furthermore, many of these RNA-RBP studies feature in-vitro approaches[70], which do not necessarily recapitulate in-vivo binding interactions, while in-vivo strategies have involved RNA pulldown techniques such as RIP, which are prone to selection of nonspecific or false interactions. Finally, many studies rely on the overexpression of the C9 HRE RNA, which does not accurately

recapitulate the physiological context because concentration and abundance of interacting molecules are critical to known mechanisms of RNA-protein interactions. Therefore, the absence of stringent and rigorous methods to detect the C9 HRE and its potential interactors have posed an immense challenge to our understanding of its mechanistic contribution to ALS molecular phenotypes.

1.1.6 SEARCHING FOR THE C9 REPEAT RNA

As previous efforts to detect the C9 G4C2 RNA foci and comprehensively characterize their RBP binding interactome have faced technical challenges in both throughput and resolution, there was a pressing need for scalable, genome-wide approaches to definitively understand the mechanistic role of this repeat RNA. Therefore, the original objective of our work was to use interdisciplinary genomics methods in C9-ALS patient cells to directly detect the C9 HRE RNA and determine its functional contribution to molecular pathogenesis, specifically by investigating RNA-protein interactions.

We set out to accomplish this goal by using two previously published methodologies developed in our research group: RNA Antisense Purification with Mass Spectrometry (RAP-MS)[75] and Split and Pool Identification of RBP Targets (SPIDR)[76]. To accomplish this, we used a two-pronged strategy, where we used patient-derived iPSNs from ALS patients and healthy controls in conjunction with an orthogonal system where we overexpressed the G4C2 HRE in HEK-293 cells (Fig. 2B-2C). This approach allowed us to conduct biochemical experiments in an overexpression model that allows for easier

execution of functional experiments, while enabling us to confirm the physiological relevance of our findings by cross-validating in patient-derived iPSNs.

In RAP-MS, we enrich for a target RNA of interest and comprehensively identify its associated proteins by using quantitative mass spectrometry (MS)[75]; this method specifically and accurately identifies the RNA binding proteins (RBPs) that directly bind an RNA in vivo (Fig. 2D). First, we used this approach in our HEK system with (+)Repeat HEK-293 cells, transfected with the plasmid containing the G4C2 HRE, in addition to a control sample, HEK-293 cells transfected with a plasmid of exactly identical sequence identity but *without* the G4C2 HRE insertion. In this experimental design, binders identified by MS that are specifically enriched for the (+)Repeat cells relative to the (-)Repeat cells must be protein binders specific to the G4C2 HRE region. After performing RAP-MS for 3 replicates of the (+)Repeat group and 3 replicates of the (+)Repeat group, we identified a set of enriched protein hits that we hypothesized were specific binders of the G4C2 repeat stretch in the HRE-transfected cells (Fig. 2E).

Next, we performed SPIDR in this same transfected HEK system in order to cross-validate the RAP-MS results and map the RNA binding sites of these candidate protein hits[76]. Briefly, SPIDR uses a combination of antibody-bead barcoding and our previously published split-pool barcoding to profile RNA binding sites of dozens to hundreds of RBPs simultaneously in a single experiment. Starting with our list of RAP-MS hit candidates, we obtained antibodies against this panel of RBPs and used our SPIDR

methodology in our overexpression HEK system (Fig 3A-C, Supp. Fig. 1). Similar to the experimental design of the RAP-MS, in a single SPIDR experiment we incorporated two replicates of RNA from the (+)Repeat HEK cells and two replicates of RNA from the (-)Repeat cells as a negative control. After sequencing and analysis, we were able to confirm the technical success of the SPIDR methodology by recapitulating known RNA binding sites of control RBPs, including PTBP1 and HNRNPH (Fig. 3D). Moreover, we were also able to identify several proteins with localization patterns over the G4C2 HRE (as annotated in the human hg38 reference genome) as defined by enrichment relative to the IgG antibody, which was embedded in the pooled immunoprecipitation reaction as a negative control (Fig. 3E). Notably, this enrichment over the G4C2 HRE for these RBPs was specific to the (+)Repeat cells (Fig. 3E).

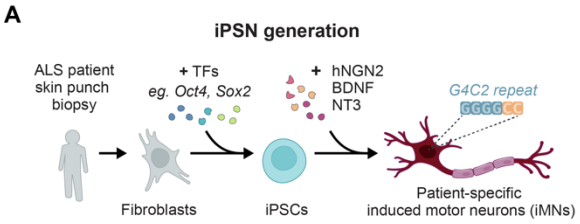
Next, we were interested in reproducing this experiment in ALS patient-derived iPSNs, with the same set of antibodies and RBP targets. However, because iPSNs had never been used for the SPIDR methodology, we first performed a small-scale experiment to ensure that the methodology would work as expected in this cell type. First, we optimized the RNA fragmentation conditions (Fig. 3B), as the fragmentation time with RNAse If must be optimized for distinct cell types. After successfully fragmenting the iPSN RNA to the target average size distribution of 200-500 bases, we used a panel of 7 previously used RBP antibodies (Supp. Fig. 2) and performed SPIDR in 1 C9-ALS patient iPSN line and 1 matched isogenic control line. After sequencing and analysis, we were once again able to recapitulate enriched binding sites of control RBPs to their known RNA targets (Fig.

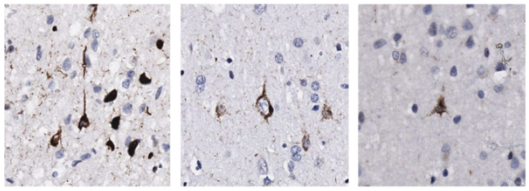
3D), confirming that SPIDR was executable in iPSNs with paired mappable RNA yields that were comparable to previously performed SPIDRs in K562s or HEK-293s.

However, we encountered a surprising finding in the iPSN SPIDR data when analyzing alignments to the C9 gene specifically. While in our HEK-293 SPIDR there was quantifiable read coverage over the G4C2 HRE with specific enrichment in the (+)Repeat cells relative to (-)Repeat cells, in our iPSN datasets there was 0 read coverage over the G4C2 region of C9. We considered that perhaps the reads had been filtered out in the alignment pipeline. We then searched for the G4C2 repeat substring, performing a liberal search for all reads containing 3 consecutive repeats of the G4C2 sequence ("GGGGCCGGGGCCGGGGCC") but were unable to find expansion reads that contained entirely G4C2s. In our HEK SPIDR results, this same analysis had yielded >3000 reads that were entirely comprised of consecutive G4C2 repeats. In the iPSN RNA, we considered that perhaps the lack of G4C2 read coverage at C9 could be an issue of sequencing depth. Nevertheless, we found this result concerning.

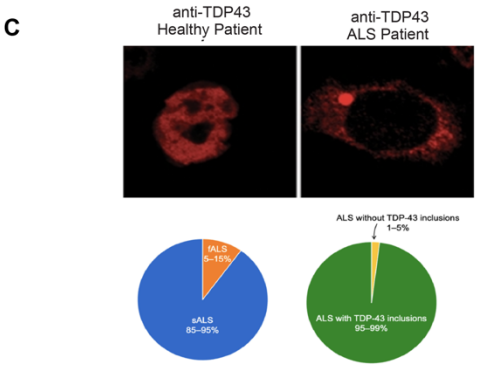
As our SPIDR was performed directly in C9-ALS patient cells, the lack of G4C2 repeat-containing reads led us to question our previously accepted mechanistic framework, as we wondered how the leading mechanism of C9-ALS pathogenesis could be the gain-of-function of an undetectable repeat RNA. We then took this opportunity to more closely investigate existing evidence of the presence of this G4C2 HRE in the C9 RNA.

1.2 MAIN FIGURES

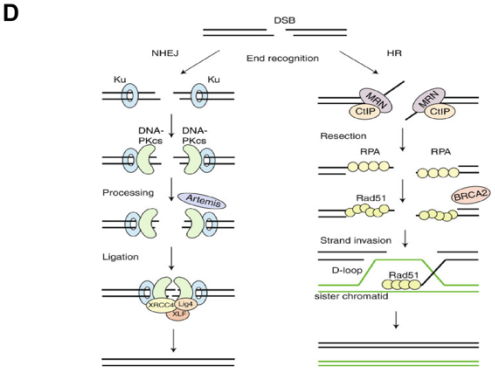




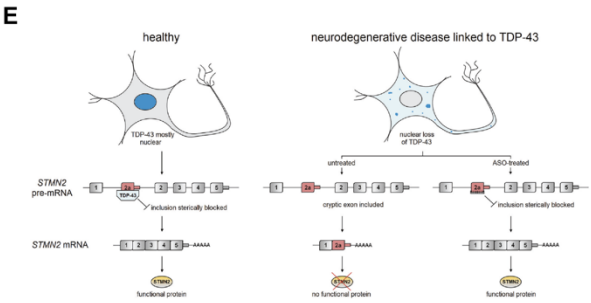
Cold Spring Harbor Laboratory. Originally from Tam et al. Cell Reports 2019



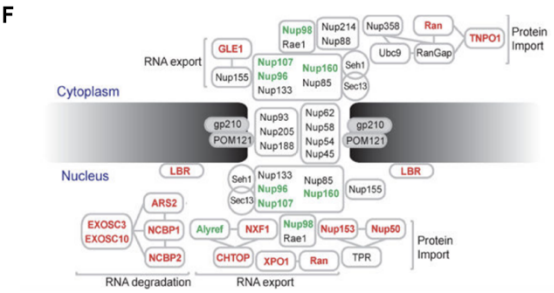
from Barkeret al. Frontier in Cellular Neuroscience 2017



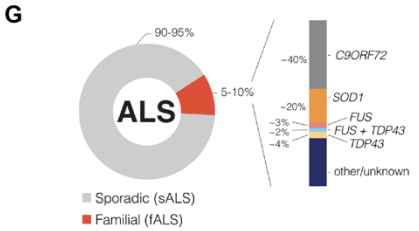
from Brandsma et al. Genome Integrity 2012



from Menge et al. Signal Transduction and Targeted Therapy 2023



from Freibaum et al. Nature 2015



Sun et al. Essays in Biochemistry 2020

Fig 1: Molecular hallmarks of ALS.
(A) Schematic showing the generation of patient-derived iPSNs from fibroblasts. (B) TDP43 inclusions (shown in dark red, nucle in blue) in ALS patient postmortem brain tissue. (C) Sample confocal microscopy (immunofluorescence) image depicting TDP43 nuclear clearance and cytoplasmic mislocalization in ALS patient compared to healthy. TDP43 cytoplasmic aggregates are seen in 98-99% of ALS patients. (D) DSBs are repaired by NHEJ and HR pathways as part of the DNA Damage Response (DDR) (Brandsma et al. Genome Integrity 2012).(E) TDP43 acts as a steric block in ALS patient cells, preventing the inclusion of a cryptic exon and functional SMN2 protein (research highlight by Menge et al). (F) Genetic screen conducted in Drosophila (Freibaum et al., Nature 2015) identifies nucleoporins and transport factors as modifiers of neurodegeneration. (G) The most common genetic mutations associated with ALS, split by sporadic sALS and familial fALS.

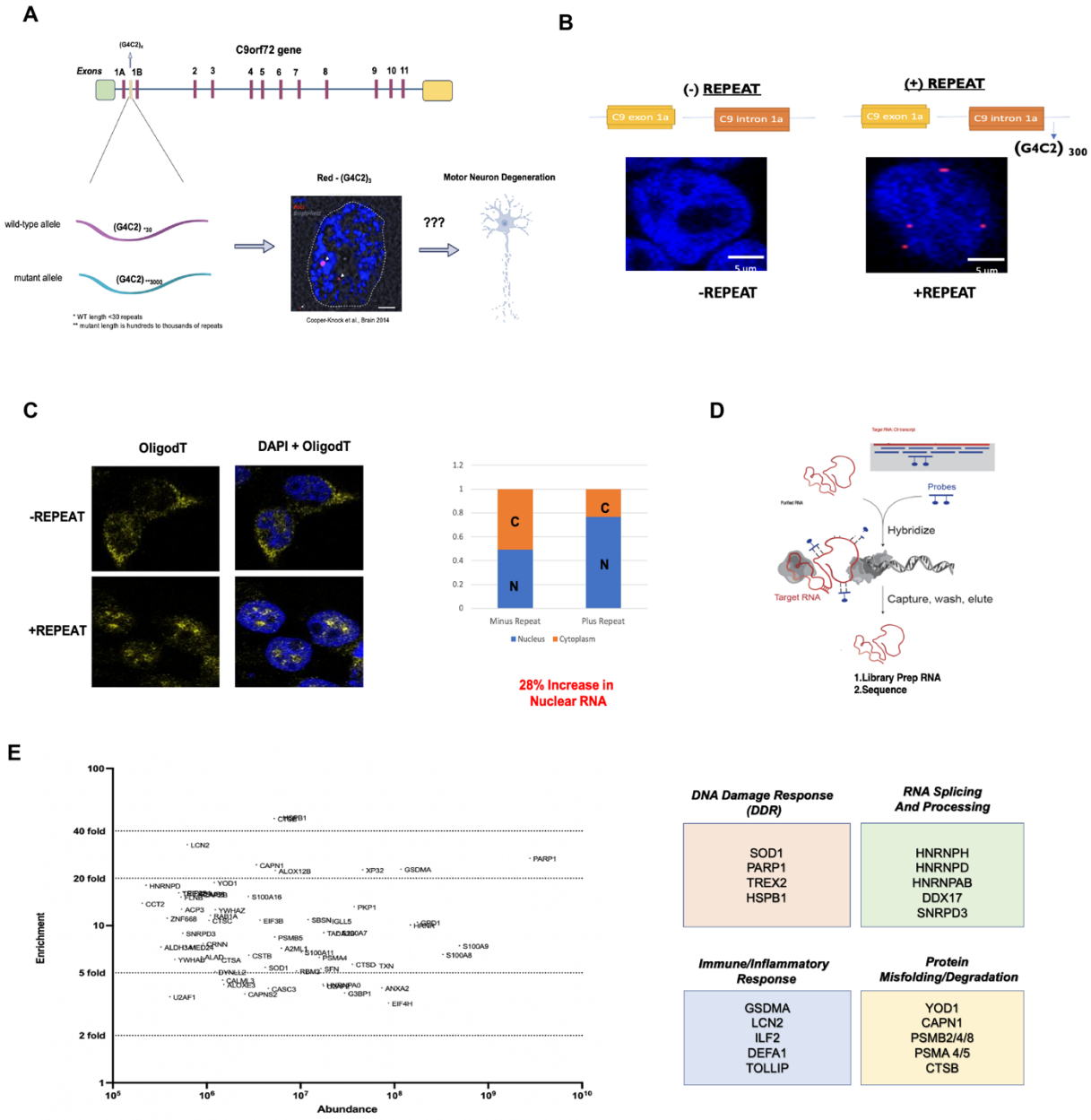


Fig 2: Characterization of G4C2-Expressing HEK System and RAP-MS.
(A) Schematic showing the G4C2 expansion site in the C9ORF72 gene. The mutation in DNA is thought to produce visible RNA foci, but the functional role of the expansion RNA in ALS pathogenesis remains unknown. (B) RNA FISH with (G4C2)x3 probes in HEK-transfected system with (-)Repeat construct and (+)Repeat construct. (C) RNA FISH with Oligo dT probes (yellow) and DAPI staining (blue) reveal a 28% increase in nuclear-retained total RNA in the (+)Repeat cells expressing the (G4C2)x300 RNA compared to (-) Repeat cells. (D) RAP-MS methodology and schematic. (E) (left) Enrichment vs abundance for top RAP-MS protein hits. Threshold set >3-fold enrichment. Selected proteins grouped by functional role (right).

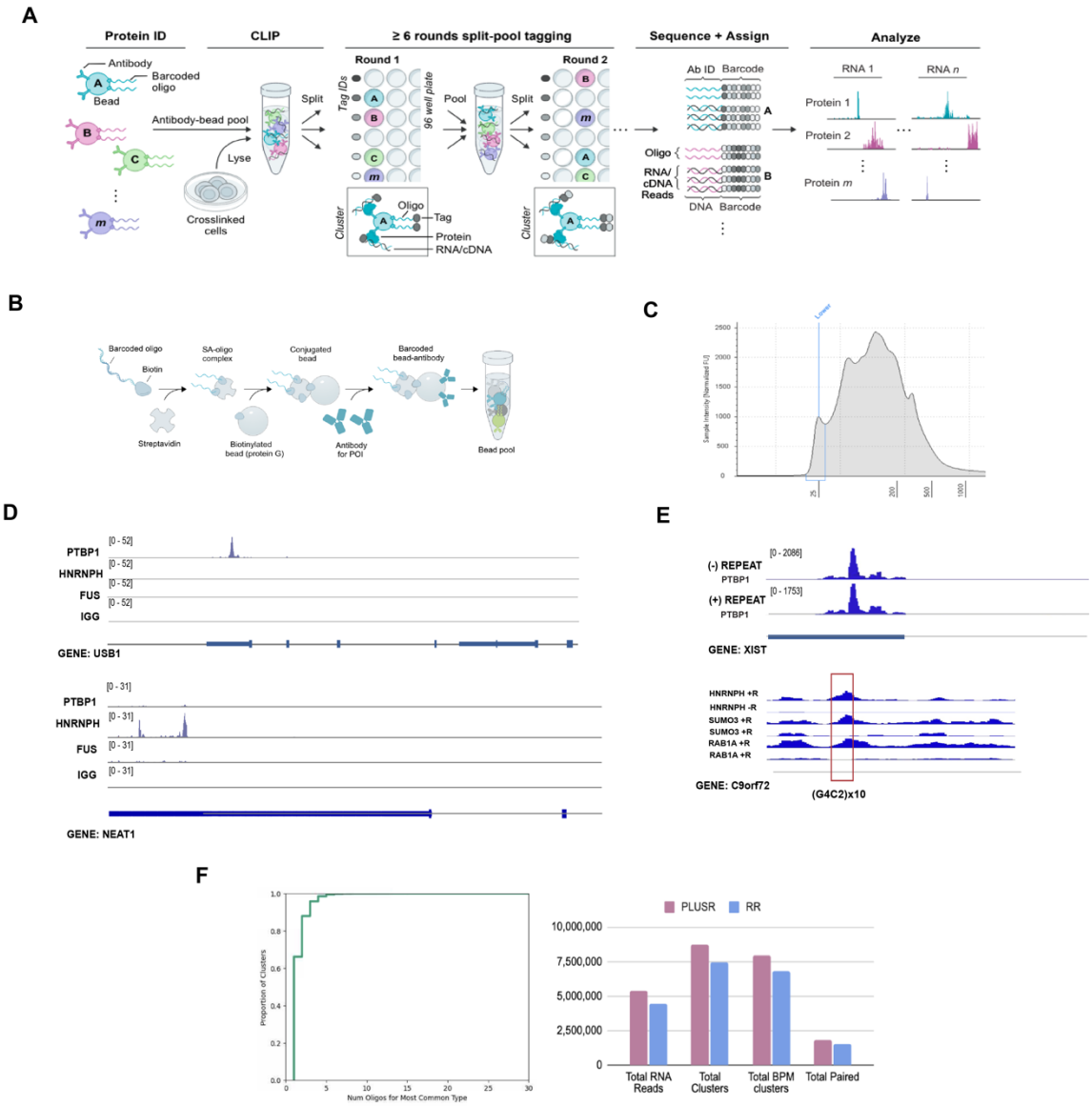


Fig. 3. SPIDR in Transfected HEK System and in C9-ALS Patient iPSNs.
(A) SPIDR methodology enables the genome-wide mapping of RNA binding proteins to their RNA targets. (B) Detailed schematic of bead-barcoding strategy used in the Protein ID step of SPIDR. (C) Sample Tapestation trace of fragmented RNA, showing average size distribution of 200-500 bases in preparation for SPIDR. (D) SPIDR in C9-ALS iPSNs, selected genomic alignment tracks shown for control proteins depict the recapitulation of known binding sites (PTBP1 to USB1, HNRNPH to NEAT1). Control tracks for other proteins show that the localization is specific. (E) SPIDR in transfected HEK system shows enrichment at the G4C2 locus that is specific to (+) Repeat cells. Control proteins (PTBP1 on Xist, top panel) are identical for both (+) Repeat and (-) Repeat. (F) SPIDR QC plots: distribution of the number of oligos of a single type per cluster (left). Statistics for the transfected HEK SPIDR (right) show quantification of total clusters, total clusters for RNA< total clusters for oligo (BPM).

1.3 SUPPLEMENTARY MATERIAL

I. Antibody Panel for RNA Binding Proteins for SPIDR in HEK-293 System

RBP Target	Manufacturer	Catalog Number
TDP43	Bethyl	A303-223A
FUS	Bethyl	A300-302A
PARP1	CST	9532S
CASC3	Abcam	ab90651
SOD1	CST	2770S
ILF2	Abcam	ab154791
G3BP1	CST	17798S
CAPN1	CST	2556S
LCN2	CST	44058S
TOLLIP	CST	4748S
RAB1A	CST	13075S
JCHAIN	Thermo Fisher	MA5-16419
PSM8	CST	13635S
S100A16	CST	13162S
HSPB1	Invitrogen	MA3-015
CTSB	CST	31718S
YOD1	Invitrogen	PA5-50157
IGLL5	Invitrogen	PA5-49022
HNRNPH	Abcam	ab10374

RBP Target	Manufacturer	Catalog Number
DEFA1	Abcam	ab9934
SNRNPD3	Invitrogen	PA5-51524
EIF3B	CST	46069S
PSMB5	CST	12919S
LGALS7	Invitrogen	PA5-115506
SUMO3	Invitrogen	700186
DHX36	Abcam	ab70269
SYF2	Abcam	ab113599
DDX17	Abcam	ab180190
EIF2S1	Abcam	ab32157
PTBP1	MBL Life Science	RN011P
IgG	Abcam	ab172730

II. Antibody Panel for RNA Binding Proteins for Pilot SPIDR in Patient iPSNs

RBP Target	Manufacturer	Catalog Number
PTBP1	MBL Life Science	RN011P
HNRNP-K	MBL Life Science	RNP019
HNRNP-H	Abcam	ab10374
TDP43	Bethyl	A303-223A
FUS	Bethyl	A300-302A
PARP-1	CST	9532S
SYF2	Abcam	ab113599
RBFOX2	Bethyl	A303-864A
IgG	Abcam	ab172730

1.4 METHODS

I. Plasmid and Reporter Cloning

The reporter construct expressing the G4C2 repeat stretch was a generous gift from the Haeusler Lab at Thomas Jefferson University. This construct was used to generate an entry clone, which was then used in a Gateway cloning reaction with an mCherry-containing destination vector to produce a final expression clone that contained mCherry as a transfection marker. BamHI and XhoI restriction sites were used to generate a (-)Repeat construct as a negative control. The length of the G4C2 repeat stretch was determined by restriction enzyme digestion of both the (+)Repeat construct and the (-)Repeat construct as control. The shared backbone of both constructs was sent out for Sanger sequencing by Primordium Labs and Laragen Inc. to verify sequence identity and generate reference maps for genome alignments and other downstream analysis.

II. HEK-293 Culture and Reporter Transfection

HEK293T cells were cultured 10 cm plates in HEK293T media consisting of 1X DMEM media (Gibco), 1 mM MEM non-essential amino acids (Gibco), 1 mM Sodium Pyruvate (Gibco), 2 mM L-Glutamine (Gibco), 1X FBS (Seradigm). HEK293T cells were transiently transfected with the G4C2 repeats-containing and repeats-removed plasmids using BioT transfection reagent (Bioland Scientific). mCherry fluorescence was monitored for 24-48 hours post-transfection as a mark of transfection efficiency.

III. RNA-FISH

RNA FISH probes were ordered ("GGGGCCGGGGCCGGGGCC" and OligodT to label PolyA RNA) from Affymetrix, Inc. RNA FISH experiments were performed using the ViewRNA ISH Cell Assay (ThermoFisher, catalog no. QVC0001). First, cells were fixed onto glass cover slips using 4% formaldehyde in PBS at RT for 20 minutes. Next, cells were permeabilized using 4% formaldehyde and 0.5% Triton X-100 in PBS at RT for an additional 10 minutes. Cells were then washed 2X with PBS and dehydrated with 70% ethanol for 20 minutes at -20C (or stored for several weeks at -20). Following dehydration, cover slips were washed 3X with PBS then incubated with the desired probe set for 3 hours at 40C. Following probe set incubation, the preamplification, amplification, and label probe incubations were performed as described in the ViewRNA ISH Cell Assay manufacturer's protocol. Finally, cover slips were incubated in 1X DAPI for 15 minutes, washed 3X in PBS, washed 1X with Ultrapure water, then mounted onto glass slides (ProLong Gold with DAPI (Invitrogen, P36935). Confocal imaging was then performed using the Leica LSM980.

IV. UV-Crosslinking

Cells were washed 1X with PBS and crosslinked directly in 10 cm culture plates. For RAP-MS, cells were crosslinked on ice to preserve RNA-protein interactions using a Spectrolinker UV Crosslinker at 0.6 J cm-2 (UV 6k) of UV at 254 nm. For SPIDR, cells

were crosslinked on ice to preserve RNA-protein interactions using a Spectrolinker UV Crosslinker at 0.4 J cm-2 (UV 4k) of UV at 254 nm. Cells were then scraped from culture dishes, washed with 1X PBS once, then pelleted by centrifugation at 1000g at 4C for 5 min. Cell pellets were then flash frozen in liquid nitrogen for storage at -80 C.

V. RAP-MS to Identify Direct Binders of C9 RNA[75]

Biotinylated antisense probes (5' - CAAGTCA + 83mer transcript specific probe - 3') were designed against the Dendra sequence, which is common to both the (+)Repeat and (-)Repeat plasmids. For each technical replicate for the (+)Repeat group, 10 million HEK-293 cells were transfected with the (+)Repeat plasmid as described in Methods Section II. As negative control, 10 million HEK-293 cells were transfected with the (-)Repeat plasmid for each technical replicate. 3 replicate sets of cells were generated for (+)Repeat and (-)Repeat each (30 million cells per group). Cells were UV-crosslinked and harvested as described in Methods Section III.

Lysis. 1 mL of Lysis Buffer (10 mM Tris pH 7.4, 5 mM EDTA, 500 mM LiCl, 0.5% Triton-X 100, 0.2% SDS, 0.1% sodium deoxycholate) supplemented with 1X final concentration Protease Inhibitor Cocktail (Sigma-#P8340-5mL) and 5 uL of Ribolock RNase Inhibitor (Thermo Fisher, #EO0382) was prepared per 10M cell pellet. A 26 gauge needle was used to further homogenize lysate if necessary by pulling the solution through the needle 3-5 times. Next, cells were sonicated using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C

for 10 cycles at 4-5 W (pulses of 0.7 s on, followed by 3.3 s off). An equal volume of 8M Urea Hybridization Buffer (8M Urea, 10 mM Tris pH 7.4, 5 mM EDTA, 500 mM LiCl, 0.5% Triton-X 100, 0.2% SDS, 0.1% sodium deoxycholate) was added to the lysate to make 4M Urea final. Insoluble cell debris was pelleted via centrifugation at 4C at 16000g for 20 mins. Supernatant was transferred to a fresh tube.

Capture 1. Beads were prepared first by phosphorylating 70 uLs of 100mM probes (per sample) with T4 PNK enzyme (NEB M0201S), PNK buffer (NEB B0201S), and 5 mM ATP (NEB P0756L) at RT for 30 minutes. Next, 350 uL Oligo-dT beads (Dynabeads, Cat #61002) per sample were washed as described in the manufacturer's protocol. PolyA bottom (5' - TGACTTG + 25-A-mer - 3' from Integrated DNA Technologies, Inc.) was hybridized to the washed beads for 30 mins shaking at RT. After washing excess PolyA bottom off the beads, the beads were then resuspended in a mix of 2X Quick Ligase (NEB, #M2200L) added to the phosphorylated probe mix. The probe ligation reaction was conducted at 1 hour at RT shaking at 1000 RPM. Then, beads were washed 3X at RT and 3X heating at 95C with TE elution buffer (20 mM Tris-HCl pH 7.4, 1 mM EDTA, 0.1% SDS) to remove excess probe. After two additional washes with 4M Urea Hybridization Buffer, beads were resuspended in 4M Urea Hybridization Buffer. For the first capture, lysate was then added to 100 uL of prepared beads and incubated for 1 hour at 42C at 1000 RPM. After 1 hour, beads were washed with the following solutions for 2 mins each at 37C: 2 washes in 4M Urea Hybridization buffer, 2 washes in SDS wash buffer (50 mM HEPES, 10% SDS, 10 mM EDTA), 2 washes in oligo-dT Wash Buffer

(50 mM HEPES, 300 mM NaCl, 2.5 mM EDTA, 0.1% Triton-X 100). The first capture was then eluted off beads at 95C for 3 minutes then adjusted to 500 mM LiCl final and 4M Urea final in preparation for the second capture.

Capture 2 and Elution. For the second capture, 50 uL of previously prepared oligo dT beads were added to the elution from Capture 1 for 1 hour at 65C at 1000 RPM. Washes and elution were repeated as after the first capture.

Mass Spectrometry. Elutions were then submitted for mass spectrometry analysis at the Caltech Protein Expression Laboratory for further characterization of enriched RNA binding proteins. Specifically, for all protein hits across all samples and replicates, the enrichment value was calculated relative to inputs and then normalized by protein abundance across all samples. Then, scores for individual proteins for the (+) Repeat samples were normalized to their corresponding scores for the (-) Repeat samples to identify candidate protein binders specific to the G4C2 HRE sequence stretch present only in the (+) Repeat.

VI. iPSC Reprogramming

Human lymphocytes from ALS patients and healthy controls were obtained (NINDS Biorepository at the Coriell Institute for Medical Research/Loma Linda University Neurology Clinic) and reprogrammed into iPSCs using episomal plasmids [REF]. The Adult Dermal Fibroblast Nucleofector Kit and Nucleofector 2b Device (Lonza) were

used to introduce mammalian vectors expressing Oct4, Sox2, Klf4, L-Myc, Lin28, and a p53 shRNA into the lymphocytes according to the manufacturer's protocol. Cells were cultured on a MEF feeder layer until the appearance of iPSCs (> 26–30 days). Then, colonies were selected and expanded in mTESR1 medium on Matrigel.

VII. Generation of iPSC derived neurons (iPSNs)[78-79]

Lipofectamine Stem (Thermofisher) was used to transfected iPSCs with both the Super piggyBac transposase expression vector (SBI) and a dox-inducible h*NGN2* expression cassette (Addgene 172115). mTESR1 medium was replaced 24 hours after transfection. At 48 hours after transfection, iPSCs were replated at spare density using Accutase + 10μM ROCK-inhibitor (Ri, Selleckchem). Ri was removed 24 hours after seeding, and 1μg/mL puromycin (Caymen 13884) was added to the media for an initial selection. Puromycin concentrations were increased until a highly pure population of BFP expressing cells was visible, and these purified transgenic iPSCs were used for further experimentation. iPSCS expressing the h*NGN2* piggybac expression cassette were dissociated with Accutase + 10μM Ri to generate iPSN cultures. Then, dissociated iPSCs were seeded into Matrigel coated 6-well plates at ~150,000-200,000 cells as single cells, directly in induction medium (IM) containing DMEM+Glutamax (Thermofisher), NEAA (Gibco 100x), 1% penicillin streptomycin, N2 supplement (Gibco), and 10ng/ml each of BDNF(R&D) and NT-3 (Peprotech), and doxycycline (Caymen , 2μg/mL) to induce the expression of the h*NGN2* transgene. Fresh IM was added to cells 48 hours post-induction and an additional 1μg/mL of puromycin was added if non-transgenic (non-converting)

cells were visible. To eliminate dividing cells, 40μM BrDu (Millipore Sigma) was added 72 hours post induction. Pure populations of early iPSNs were observed 5 days after induction, and cell media was then switched to neuronal maintenance medium (MM), containing neurobasal (thermofisher), N2 and B27 supplements (Gibco), 1% penicillin/streptomycin, glutamax (Thermofisher, 100x), and NEAA (Gibco 100x, BDNF and NT-3 (10ng/mL). Cell media was replaced every 72 hours.

VIII. SPIDR for Genome-Wide Mapping of RBPs to RNA[76]

Lysis and Fragmentation. Cells were generated and crosslinked with 4K UV crosslinking as previously described. The HEK-293 SPIDR experiment was performed with 10M cells transfected with the (+)Repeat construct and 10M cells transfected with (-)Repeat construct. The iPSN SPIDR experiment was performed with 10M cells for patient line #6769 and 10M cells for its isogenic control.

For both experiments, each 10M cell pellet was lysed with 1 mL RIPA buffer (50mM HEPES pH 7.4, 100mM NaCl, 1% NP-40, 0.5% Na-Deoxycholate, 0.1% SDS) with 1X final concentration of Protease Inhibitor Cocktail (Sigma-#P8340-5mL), 5 uL of Ribolock RNase Inhibitor (Thermo Fisher, #EO0382)), 10 uL of Turbo DNAse ((Invitrogen, #AM2238), and 1X of Manganese/Calcium mix (2.5 mM MnCl2 , 0.5 mM CaCl2). Cells were incubated for 10 mins on ice in this lysis reaction. Next, cells were sonicated using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C for a total of 1.5 min at 4-5 W (pulses of 0.7 s

on, followed by 3.3 s off). Cells were then incubated at 37C for 10 minutes to allow for DNase digestion, then the reaction was quenched with 0.25 M EDTA/EGTA mix for a final concentration of 10 mM EDTA/ EGTA. To fragment the RNA, a 1:500 dilution of RNAse If ((NEB, #M0243L) was added and lysate was incubated at 37C for 10 mins to fragment RNA to a final size distribution of 300-400 bps. The RNAse reaction was quenched with 500 uL of ice-cold RIPA buffer supplemented with 5 uLs of Ribolock RNAse Inhibitor and 1X Protease Inhibitor Cocktail, followed by 3 minute incubation on ice for complete quenching. Centrifugation was performed at 15000g at 4C for 2 mins to clear the lysates. Supernatant was transferred to fresh tubes. Size distribution of the fragmented RNA was verified by cleaning the samples using the Zymo IC RNA Clean and Concentrator kit and quantification using the Tapestation High Sensitivity RNA kit. The final fragmented lysate was then stored on ice until the antibody-coupled beads were ready for immunoprecipitation.

Preparation of Antibody-Coupled Beads. The bead labeling strategy is adopted from the Guttman lab methodology ChIP DIP, which enables the multiplexed mapping of proteins to DNA (https://guttmanlab.caltech.edu/technologies/). First, antibody ID oligos were designed and ordered from Integrated DNA Technologies, Inc. (IDT). Each antibody ID oligo contains a 5'phosphate group to enable ligation, a 3'biotin group to allow binding to streptavidin beads, a UMI (blue), a sticky end to ligate to subsequent ODD barcodes (red), and a unique antibody barcode sequence (green). First, the biotinylated antibody ID oligos were coupled with purified streptavidin (BioLegend,

280302) to make a stock of 909 nM streptavidin conjugated oligo, then diluted 4X to make a final dilution plate of 227 nM. In parallel, Protein G beads were biotinylated as described in the Guttman Lab ChIP-DIP protocol [REF] with 5 mM EZ-Link Sulfo-NHS-Biotin (Thermo, 21217) at room temperature for 30 minutes.

10 uLs of oligo-coupled Protein G beads were prepared for each capture antibody in the SPIDR experiment. As described in the original SPIDR protocol, biotinylated Protein G beads were first washed then aliquoted into a 96-well plate. Then, 14 μL from the from 227nM stock plate of streptavidin-coupled antibody ID oligos was added to each well. The streptavidin coupled antibody ID oligos and biotinylated Protein G beads were bound at room temperature for 30 minutes at 1200 RPM. Beads were then washed twice in M2 buffer (20 mM Tris 7.5, 50 mM NaCl, 0.2% Triton X-100, 0.2% Na-Deoxycholate, 0.2% NP-40), and twice in PBST. The number of oligos loaded per bead was quantified as a QC step before proceeding to immunoprecipitation. The “Terminal” tag from the split pool barcoding scheme was ligated onto a 20% fraction of the conjugated beads, and this ligated product was then PCR amplified for 10 cycles, and purified using 1X SPRI beads. The purified product was quantified using Tapestation D1000, and the concentration of the final library and the number of PCR cycles was used to quantify the pre-PCR oligo complexity. The pre-PCR oligo complexity was divided by the number of beads to obtain the bead loading ratio. Conditions were optimized such that 250-300 oligos/bead were loaded for each well of capture antibody.

Next, 2.5 ugs of each antibody was added to each well of the 96-well plate of oligo-coupled Protein G beads. The plate was incubated at RT for 30 mins, then washed twice with 1X PBST. Then, each well of beads was resuspended in 200 µL of 1x PBSt and 4mM biotin and incubated at room temperature for 15 minutes to quench free Protein G or free streptavidin binding sites.

Pooled Immunoprecipitation. Next, all wells of oligo labeled and antibody coupled beads were washed 2x with PBST + 2 mM biotin. In preparation for the immunoprecipitation reaction, the volume from all wells were pooled together into one tube. In parallel, the fragmented lysate was diluted with RIPA buffer such that the final volume of the reaction corresponded to 1 mL of RIPA buffer for every 100 uL Protein G beads. The antibody coupled bead pool was then split by volume and added to all tubes of diluted lysate. 1M biotin was added to a final concentration of 10mM biotin to quench any remaining free streptavidin-oligo. Each immunoprecipitation reaction (lysate + pooled antibody-oligos) were incubated on a HulaMixer at 4C overnight . The following day, beads were washed 2x with RIPA buffer, 2X with High Salt Wash Buffer (50 mM HEPES pH 7.4, 1 M NaCl, 1% NP-40, 0.5% Na-Deoxycholate, 0.1% SDS), and 2X with CLAP Tween buffer (50 mM HEPES pH 7.4, 0.1% Tween-20).

Tagging of RNA Molecules and Preparation for Barcoding. Beads were incubated at 37C for 10 mins with T4 Polynucleotide Kinase (NEB, #M0201L) to modify the 3'ends of RNA to have 3'OH groups for subsequent ligation. Beads were washed 2x with High

Salt Wash Buffer and 2x with CLAP Tween buffer after end repair of the RNA. Next, RNA was ligated with "RNA Phosphate Modified" (RNA) adaptor (Quinodoz et al., 2021[77]) and High Concentration T4 RNA Ligase I (NEB, M0437M) shaking on a Thermomixer for 1 hour 15 mins at 24C. Beads were then washed 3x with CLAP Tween buffer. RNA was converted to cDNA using Maxima RT 42C for 20 minutes with "RPM Bottom" as an RT primer, enabling ligation during split-pool barcoding by adding a 5'sticky end. Next, Exonuclease I (NEB, #M0293L) was used at 37C for 15 minutes to digest excess primer from the RT reaction.

Split Pool Barcoding Split-pool barcoding was performed as described in previous publications with some modifications. Specifically, beads were split-pool-tagged with 6 rounds with sets of "Odd," "Even," and "Terminal" tags. 6 rounds were chosen to ensure that all beads used in the experiment could be resolved, such that almost all barcode clusters (>95%) represented molecules belonging to unique, individual beads. All barcode ligation steps were supplemented with 1:40 Ribolock RNAse Inhibitor and 2mM biotin and were performed at room temperature for 4 minutes. After the final round of barcoding, beads were resuspended in CLAP Tween buffer and aliquots (5% of total beads) were stored for library preparation.

Library Preparation. First, RNA in each aliquot was degraded by incubating with RNase cocktail (Invitrogen, #AM2286) and RNase H (NEB, #M0297L) for 20 minutes at 37C. "Splint ligation" (as described in Quinodoz et al 2021[77]) was used to attach double

stranded oligonucleotides to the 3'ends of the resulting cDNA. 1X Instant Sticky End Master Mix (NEB #M0370) was used for the splint ligation reaction for 1 hour at 24°C at 1400 RPM on a ThermoMixer. Elution of the biotinylated oligo and barcoded cDNA were performed by boiling in NLS elution buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 2% N-lauroylsarcosine, 2.5 mM TCEP) for 6 minutes at 91C, at 1350 RPM shaking.

Then, biotinylated oligo was captured by eluting the eluate in 0.5X PBST, 5 mM Tris pH 8.0, 0.5 mM EDTA, 1M NaCl. This reaction was then bound to MyOne Streptavidin C1 Dynabeads (Invitrogen, #65001) for 30 minutes at room temperature. Beads were placed on a magnet and the cDNA-containing supernatant was stored in a fresh tube. 2X Q5 Hot Start Master Mix (NEB #M0494) was used to PCR amplify the biotinylated oligo on-bead, using indexed Illumina adaptor primers. In parallel, the cDNA was incubated with a "anti-RPM," a biotinylated antisense ssDNA that removes empty insert products by hybridizing between the reverse transcription primer and the splint. After incubation with anti-RPM, the reaction was bound to MyOne Streptavidin C1 Dynabeads (Invitrogen, #65001) for 30 minutes at room temperature. The supernatant was purified using silane beads ((Invitrogen, #37002D) as described in the manufacturer's protocol. 2X Q5 Hot Start Master Mix (NEB #M0494) was used to PCR amplify the cDNA, again using primers with indexed Illumina adaptor sequences.

For both the oligo and cDNA libraries, 1.2X SPRI beads (Bulldog Bio CNGS500) were used to clean the PCR product and DNA High Sensitivity D1000 Tapestation was used to quantify library size distribution. Before sequencing, a 2% agarose gel was used to gel purify libraries to remove excess primer.

Sequencing. Paired-end sequencing was performed using an Illumina NextSeq 2000 using reads lengths of 100x200 (Read1xRead2) nucleotides. Sequencing depth for each aliquot was calculated from the number of unique RNA molecules and number of barcoded beads in each aliquot. On average, cDNA was sequenced to 2X saturation and beads were sequenced to 5X saturation.

IX. SPIDR Analysis

Trim Galore! V0.6.2 was used to trim adaptor sequences, and trimming quality was then evaluated with FastQC v0.11.8. Cutadapt v3.4 was used to trimmed the RPM sequence from both the 5' and 3' end of reads. The Guttman lab's previously published Barcode ID v1.2.0[77] (https:// github.com/GuttmanLab/sprite2.0-pipeline) was used to identify barcodes and assess ligation efficiency for each round of split-pool barcoding. RNA and oligo tag reads were separated to two output files (using RPM sequence to identify RNA reads). For RNA reads, Bowtie2 was first used to align to a reference genome and contains the sequences for repetitive RNAs (eg. rRNAs, snRNAs, tRNAs, 45S pre-rRNAs, snoRNAs). For the remaining unaligned reads, STAR aligner was then used to

align to the hg38 human genome. PCR duplicates were removed in the genomic alignment by identifying reads with identical start and stop positions in the genome.. Similarly, the unique molecular identifier (UMI) on the oligo reads were used to remove duplicates.

Aligned RNA reads were then merged with oligo reads to generate cluster files as previously published (https:// github.com/GuttmanLab/sprite2.0-pipeline) [77], incorporate filtering of barcode strings that were not in the correct order. From the filtered cluster files, RNA reads were then split into separate files by oligo IDs that corresponded to each protein. As each cluster in SPIDR represents a unique bead, the distribution of oligo tags (each tag represents a unique protein) was measured to determine the metrics by which to split the cluster file. Specifically, we set a threshold that >80% of all tags in a cluster needed to correspond to a protein-oligo tag in order to designate that cluster as belonging to a given protein. We then visualized the genomic alignment files for each protein in Integrated Genomics Viewer (IGV).

1.5 REFERENCES

1. Masrori, P. & Van Damme, P. Amyotrophic lateral sclerosis: a clinical review. *Eur J Neurol* 27, 1918–1929 (2020).

2. Amyotrophic lateral sclerosis (ALS) - Symptoms and causes. *Mayo Clinic* https://www.mayoclinic.org/diseases-conditions/amyotrophic-lateral-sclerosis/symptoms-causes/syc-20354022.

3. Amyotrophic Lateral Sclerosis (ALS) | National Institute of Neurological Disorders and Stroke. https://www.ninds.nih.gov/health-information/disorders/amyotrophic-lateral-sclerosis-als.

4. Mead, R. J., Shan, N., Reiser, H. J., Marshall, F. & Shaw, P. J. Amyotrophic lateral sclerosis: a neurodegenerative disorder poised for successful therapeutic translation. *Nat Rev Drug Discov* 22, 185–212 (2023).

5. Amyotrophic lateral sclerosis: translating genetic discoveries into therapies | Nature Reviews Genetics. https://www.nature.com/articles/s41576-023-00592-y.

6. Lee, S. & Huang, E. J. Modeling ALS and FTD with iPSC-derived Neurons. *Brain Res* 1656, 88–97 (2017).

7. Soto, C. & Pritzkow, S. Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases. *Nat Neurosci* 21, 1332–1340 (2018).

8. Tollervey, J. R. *et al.* Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience* 14, 452–458 (2011).

9. Loss of nuclear TDP-43 in amyotrophic lateral sclerosis (ALS) causes altered expression of splicing machinery and widespread dysregulation of RNA splicing in motor neurones - Highley - 2014 - Neuropathology and Applied Neurobiology - Wiley Online Library. https://onlinelibrary.wiley.com/doi/abs/10.1111/nan.12148.

10. Mackenzie, I. R. A. & Rademakers, R. The role of TDP-43 in amyotrophic lateral sclerosis and frontotemporal dementia. *Curr Opin Neurol* 21, 693–700 (2008).

11. Rothstein, J. D., Warlick, C. & Coyne, A. N. Highly variable molecular signatures of TDP-43 loss of function are associated with nuclear pore complex injury in a population study of sporadic ALS patient iPSNs. *bioRxiv* 2023.12.12.571299 (2023) doi:10.1101/2023.12.12.571299.

12. Giannini, M. *et al.* TDP-43 mutations link Amyotrophic Lateral Sclerosis with R-loop homeostasis and R loop-mediated DNA damage. *PLOS Genetics* 16, e1009260 (2020).

13. Scotter, E. L., Chen, H.-J. & Shaw, C. E. TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. *Neurotherapeutics* 12, 352–363 (2015).

14. Bright, F., Chan, G., van Hummel, A., Ittner, L. M. & Ke, Y. D. TDP-43 and Inflammation: Implications for Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Int J Mol Sci* 22, 7781 (2021).

15. Jo, M. *et al.* The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Exp Mol Med* 52, 1652–1662 (2020).

16. Ling, S.-C., Polymenidou, M. & Cleveland, D. W. Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis. *Neuron* 79, 416–438 (2013).

17. Tyzack, G. E. *et al.* Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. *Brain* 142, 2572–2580 (2019).

18. Jutzi, D. *et al.* Aberrant interaction of FUS with the U1 snRNA provides a molecular mechanism of FUS induced amyotrophic lateral sclerosis. *Nat Commun* 11, 6341 (2020).

19. Sharma, A. *et al.* ALS-associated mutant FUS induces selective motor neuron degeneration through toxic gain of function. *Nat Commun* 7, 10465 (2016).

20. Scekic-Zahirovic, J. *et al.* Cytoplasmic FUS triggers early behavioral alterations linked to cortical neuronal hyperactivity and inhibitory synaptic defects. *Nat Commun* 12, 3028 (2021).

21. Mackenzie, I. R. A. & Neumann, M. FET proteins in frontotemporal dementia and amyotrophic lateral sclerosis. *Brain Res* 1462, 40–43 (2012).

22. Svetoni, F., Frisone, P. & Paronetto, M. P. Role of FET proteins in neurodegenerative disorders. *RNA Biol* 13, 1089–1102 (2016).

23. Sun, Y., Curle, A. J., Haider, A. M. & Balmus, G. The role of DNA damage response in amyotrophic lateral sclerosis. *Essays Biochem* 64, 847–861 (2020).

24. Wang, H., Kodavati, M., Britz, G. W. & Hegde, M. L. DNA Damage and Repair Deficiency in ALS/FTD-Associated Neurodegeneration: From Molecular Mechanisms to Therapeutic Implication. *Frontiers in Molecular Neuroscience* 14, (2021).

25. Mitra, J. *et al.* Motor neuron disease-associated loss of nuclear TDP-43 is linked to DNA double-strand break repair defects. *Proc Natl Acad Sci U S A* 116, 4696–4705 (2019).

26. Higelin, J. *et al.* FUS Mislocalization and Vulnerability to DNA Damage in ALS Patients Derived hiPSCs and Aging Motoneurons. *Front. Cell. Neurosci.* 10, (2016).

27. Kannan, A., Cuartas, J., Gangwani, P., Branzei, D. & Gangwani, L. Mutation in senataxin alters the mechanism of R-loop resolution in amyotrophic lateral sclerosis 4. *Brain* 145, 3072–3094 (2022).

28. Naumann, M. *et al.* Impaired DNA damage response signaling by FUS-NLS mutations leads to neurodegeneration and FUS aggregate formation. *Nat Commun* 9, 335 (2018).

29. Izumikawa, K. *et al.* TDP-43 regulates site-specific 2'-O-methylation of U1 and U2 snRNAs via controlling the Cajal body localization of a subset of C/D scaRNAs. *Nucleic Acids Res* 47, 2487–2505 (2019).

30. Arnold, E. S. *et al.* ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proceedings of the National Academy of Sciences* 110, E736–E745 (2013).

31. Zhou, Y., Liu, S., Liu, G., Öztürk, A. & Hicks, G. G. ALS-Associated FUS Mutations Result in Compromised FUS Alternative Splicing and Autoregulation. *PLOS Genetics* 9, e1003895 (2013).

32. Sun, S. *et al.* ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nat Commun* 6, 6171 (2015).

33. Reber, S. *et al.* Minor intron splicing is regulated by FUS and affected by ALS-associated FUS mutants. *The EMBO Journal* 35, 1504–1521 (2016).

34. Tsuiji, H. *et al.* Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. *EMBO Molecular Medicine* 5, 221–234 (2013).

35. Mehta, P. R., Brown, A.-L., Ward, M. E. & Fratta, P. The era of cryptic exons: implications for ALS-FTD. *Molecular Neurodegeneration* 18, 16 (2023).

36. Brown, A.-L. *et al.* TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* 603, 131–137 (2022).

37. Menge, S., Decker, L. & Freischmidt, A. Restoring expression of Stathmin-2: a novel strategy to treat TDP-43 proteinopathies. *Sig Transduct Target Ther* 8, 1–2 (2023).

38. Melamed, Z. *et al.* Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. *Nat Neurosci* 22, 180–190 (2019).

39. Prudenceio M. *et al.* Truncated stathmin-2 is a marker of TDP-43 pathology in frontotemporal dementia. *J Clin Invest* 130, 6080–6092 (2020).

40. Baugh, M. W. *et al.* Mechanism of STMN2 cryptic splice-polyadenylation and its correction for TDP-43 proteinopathies. *Science* 379, 1140–1149 (2023).

41. Perron, B. *et al.* Alternative Splicing of ALS Genes: Misregulation and Potential Therapies. *Cell Mol Neurobiol* 40, 1–14 (2020).

42. Wang, Q., Conlon, E. G., Manley, J. L. & Rio, D. C. Widespread intron retention impairs protein homeostasis in C9ORF72 ALS brains. *Genome Res.* 30, 1705–1715 (2020).

43. Luisier, R. *et al.* Intron retention and nuclear loss of SFPQ are molecular hallmarks of ALS. *Nat Commun* 9, 2010 (2018).

44. Zhang, K., Grima, J. C., Rothstein, J. D. & Lloyd, T. E. Nucleocytoplasmic transport in C9ORF72-mediated ALS/FTD. *Nucleus* 7, 132–137 (2016).

45. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* 495, 467–473 (2013).

46. Chou, C.-C. *et al.* TDP-43 pathology disrupts nuclear pore complexes and nucleocytoplasmic transport in ALS/FTD. *Nat. Neurosci.* 21, 228–239 (2018).

47. McGoldrick, P. & Robertson, J. Unraveling the impact of disrupted nucleocytoplasmic transport systems in C9ORF72-associated ALS. *Front Cell Neurosci* 17, 1247297 (2023).

48. Freibaum, B. D. *et al.* GGGGCC repeat expansion in C9ORF72 compromises nucleocytoplasmic transport. *Nature* 525, 129–133 (2015).

49. Zhang, K. *et al.* The C9ORF72 repeat expansion disrupts nucleocytoplasmic transport. *Nature* 525, 56–61 (2015).

50. Butti, Z. & Patten, S. A. RNA Dysregulation in Amyotrophic Lateral Sclerosis. *Frontiers in Genetics* 9, (2019).

51. Mejzini, R. *et al.* ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci* 13, 1310 (2019).

52. Kim, G., Gautier, O., Tassoni-Tsuchida, E., Ma, X. R. & Gitler, A. D. ALS Genetics: Gains, Losses, and Implications for Future Therapies. *Neuron* 108, 822–842 (2020).

53. Ghasemi, M. & Brown, R. H. Genetics of Amyotrophic Lateral Sclerosis. *Cold Spring Harb Perspect Med* 8, a024125 (2018).

54. Bennett, C. L. *et al.* Senataxin mutations elicit motor neuron degeneration phenotypes and yield TDP-43 mislocalization in ALS4 mice and human patients. *Acta Neuropathol* 136, 425–443 (2018).

55. Richard, P. *et al.* SETX (senataxin), the helicase mutated in AOA2 and ALS4, functions in autophagy regulation. *Autophagy* 17, 1889–1906.

56. Xue, Y. C. *et al.* Dysregulation of RNA-Binding Proteins in Amyotrophic Lateral Sclerosis. *Frontiers in Molecular Neuroscience* 13, (2020).

57. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron* 72, 245–256 (2011).

58. Xu, Z. *et al.* Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. *Proceedings of the National Academy of Sciences* 110, 7778–7783 (2013).

59. Lee, Y.-B. *et al.* Hexanucleotide Repeats in ALS/FTD Form Length-Dependent RNA Foci, Sequester RNA Binding Proteins, and Are Neurotoxic. *Cell Reports* 5, 1178–1186 (2013).

60. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics* 108, 764–785 (2021).

61. Paulson, H. Repeat expansion diseases. *Handb Clin Neurol* 147, 105–123 (2018).

62. Ellerby, L. M. Repeat Expansion Disorders: Mechanisms and Therapeutics. *Neurotherapeutics* 16, 924–927 (2019).

63. Ramakrishnan, S. & Gupta, V. Trinucleotide Repeat Disorders. in *StatPearls [Internet]* (StatPearls Publishing, 2023).

64. Zhou, Z.-D., Jankovic, J., Ashizawa, T. & Tan, E.-K. Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. *Nat Rev Neurol* 18, 145–157 (2022).

65. van Blitterswijk, M. *et al.* Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol* 12, 978–988 (2013).

66. Zu, T. *et al.* RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proceedings of the National Academy of Sciences* 110, E4968–E4977 (2013).

67. Targeted degradation of sense and antisense C9ORF72 RNA foci as therapy for ALS and frontotemporal degeneration | PNAS. https://www.pnas.org/doi/full/10.1073/pnas.1318835110.

68. Wood, H. C9ORF72 RNA foci—a therapeutic target for ALS and FTD? *Nat Rev Neurol* 9, 659–659 (2013).

69. Cooper-Knock, J. *et al.* C9ORF72 GGGGCC Expanded Repeats Produce Splicing Dysregulation which Correlates with Disease Severity in Amyotrophic Lateral Sclerosis. *PLoS ONE* 10, e0127376 (2015).

70. Conlon, E. G. *et al.* The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife* 5, e17820 (2016).

71. Barker, H. V., Niblock, M., Lee, Y.-B., Shaw, C. E. & Gallo, J.-M. RNA Misprocessing in C9ORF72-Linked Neurodegeneration. *Front Cell Neurosci* 11, 195 (2017).

72. Sattler, R. *et al.* Roadmap for C9ORF72 in Frontotemporal Dementia and Amyotrophic Lateral Sclerosis: Report on the C9ORF72 FTD/ALS Summit. *Neurol Ther* **12**, 1821–1843 (2023).

73. Cleary, E. M. *et al.* Improved PCR based methods for detecting C9ORF72 hexanucleotide repeat expansions. *Mol Cell Probes* **30**, 218–224 (2016).

74. Ebbert, M. T. W. *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).

75. McHugh, C. A. & Guttman, M. RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells. *Methods Mol Biol* 1649, 473–488 (2018).

76. Wolin, E. *et al.* SPIDR: a highly multiplexed method for mapping RNA-protein interactions uncovers a potential mechanism for selective translational suppression upon cellular stress. *bioRxiv* 2023.06.05.543769 (2023) doi:10.1101/2023.06.05.543769.

77. Quinodoz, S. A. *et al.* SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. *Nat Protoc* 17, 36–75 (2022).

78. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat Methods* **8**, 409–412 (2011).

79. Pfisterer, U. *et al.* Direct conversion of human fibroblasts to dopaminergic neurons. *Proceedings of the National Academy of Sciences* **108**, 10343–10348 (2011).

Chapter 2

THE ALS-ASSOCIATED REPEAT EXPANSION IN C9ORF72 IS NOT TRANSCRIBED AND MAY INSTEAD ACT BY DISRUPTING AS A REGULATORY ELEMENT

(in review at *Nature Medicine*)

Paulomi Bhattacharya[1], Mario Blanco[1], Jasper Rubin-Sigler[2,3], Justin Ichida[2,3*], and Mitchell Guttman[1*]

1. Division of Biology and Bioengineering, California Institute of Technology, Pasadena CA 10 91125, USA

2. Department of Stem Cell Biology and Regenerative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

3. Eli and Edyth Broad CIRM Center for Regenerative Medicine and Stem Cell Research at USC, Los Angeles, CA, USA

ABSTRACT

The G4C2 hexanucleotide repeat expansion in the first intron of the C9ORF72 gene is the most common genetic mutation linked to ALS, accounting for ~40 percent of familial and 10 percent of sporadic cases. Yet, its functional contribution to molecular pathogenesis remains unknown. The prevailing model is that this expansion leads to transcription of a novel RNA (C9-repeat RNA) that leads to disease either through its RNA product or translation of dipeptide repeat proteins it encodes ("gain-of-function"). However, recent attempts to degrade the C9-repeat RNA in several major clinical trials have failed to show any improvement in C9-ALS patients, raising questions about what role, if any, the C9-repeat RNA plays in ALS pathogenesis. Here, we demonstrate that the C9-repeat RNA is not detectable in C9-ALS patient-derived iPSNs or postmortem brain tissue. We show that transcription of the C9ORF72 gene initiates downstream of the G4C2 repeat sequence with the repeat expansion residing at a promoter-proximal region and displaying chromatin signatures of an enhancer. Because this region is GC-rich and has been reported to be preferentially methylated in C9-ALS patients, we explored whether this repeat expansion might lead to reduced C9ORF72 gene expression. We show that the C9-repeat is associated with reduced allele-specific expression of the C9ORF72 gene, consistent with the GC-rich features of the repeat expansion and previous reports of preferential DNA methylation in C9-ALS patients. Taken together, our findings challenge the prevailing gain-of-function models in C9-ALS

and instead suggest that the repeat expansion region may function as a regulatory element that silences C9ORF72 expression from the mutant allele.

INTRODUCTION

Many neurodegenerative disorders, including Huntington's disease, Fragile X syndrome, Fredrich's ataxia, frontotemporal dementia (FTD), and amyotrophic lateral sclerosis (ALS) are caused by repeat expansion mutations[1-4]. One example is the expansion of the G4C2 hexanucleotide repeat in the intron of the C9ORF72 gene (C9-repeat expansion), which is the most common mutation leading to ALS, accounting for ~40 percent of familial and 10 percent of sporadic cases [5-8]. C9-ALS patients are generally heterozygous for the C9-repeat expansion, with the wild-type allele containing 3-20 copies of the G_4C_2 repeat and the mutant allele containing hundreds to thousands of copies[9-10].

Yet, despite the clear genetic importance of this mutation in leading to ALS[5-10], its molecular role in pathogenesis has been debated since its initial discovery in 2011[5-8]. Two opposing models have been proposed [12-18]: (i) the C9-repeat leads to a reduction in the levels of the C9ORF72 mRNA and protein product, leading to disease ("loss-of-function") and (ii) the presence of the C9-repeat leads to production of a novel RNA or protein product that leads to disease ("gain-of-function").

In support of the gain of function model, early studies reported the detection of an RNA product produced from the C9-repeat (C9-repeat RNA) in neurons differentiated from patient-derived induced pluripotent stem cells (iPSNs, Fig. 1A) and postmortem patient tissue [19-24]. There are two proposals for how the C9-repeat RNA could lead to pathogenesis: (1) the C9-repeat RNA itself is a toxic molecular species that sequesters RNA binding proteins (RBPs) away from their endogenous targets [25-37], and (2) non-canonical translation of the C9-repeat RNA gives rise to toxic dipeptide repeat proteins (DPRs) [12-18, 38-43]. Based on these gain-of-function models, therapeutic strategies have been developed that target the C9-repeat RNA to reduce both the RNA itself and the DPR proteins they encode [44-46].

Yet, attempts to degrade the C9-repeat RNA in recent clinical trials have failed to show any improvement in C9-ALS patients, raising questions about the role of the C9-repeat RNA in ALS pathogenesis [47]. In addition, there are multiple reasons to question this model. **First**, measuring the C9-repeat RNA has been challenging because of its large size, low sequence complexity, and high GC content. While the presence of the C9-repeat in genomic DNA is traditionally measured by Southern blots, repeat-primed PCR [48], and long read sequencing (e.g. PacBio and Nanopore) [49], measurement of the C9-repeat RNA has relied primarily on imaging of the low complexity repeat sequence (e.g., RNA FISH) using short probes against 3-4 copies of the G_4C_2 repeat [19-24]. Yet, there are many other RNAs in the human genome (e.g., RRP36, SULF2, HUWE1, RGS14) that contain this short repeat sequence, and imaging does not directly measure whether these detected G_4C_2-containing RNA

arises from the C9-RNA or other RNAs containing this same repeat stretch. **Second**, the reported C9-repeat RNA foci are most frequently observed in patient cell types that are typically unaffected in ALS (e.g., cerebellum), and the percentage of cells reported with C9-repeat RNA foci in ALS-affected brain regions or patient-derived iPSNs is extremely low, with some studies reporting 1-2 foci for only 10% of cells [12-24].

Based on these observations, we revisited the gain-of-function model and investigated the expression of the C9-repeat RNA in ALS patient samples. Here, we show that the C9-repeat RNA is not transcribed and retained in the mature C9-RNA, nor is it transcribed in the nascent pre-mRNA and spliced out. Rather, transcription at the C9 locus is initiated downstream of the C9-repeat such that the repeat is not transcribed from the wild-type or the mutant allele. Instead, this genomic region has chromatin signatures of enhancer activity, suggesting that the repeat expansion may instead act to modulate the expression level of C9ORF72 and/or neighboring genes. Together, our work argues for a critical reconsideration of the prevailing notion that the gain-of-function of the C9-repeat RNA or its encoded DPR proteins is responsible for the molecular pathogenesis of ALS.

2.1 RESULTS

2.1.1 THE C9-REPEAT RNA IS UNDETECTABLE IN C9-ALS PATIENT-DERIVED IPSNS AND POSTMORTEM TISSUE

To investigate the expression of the C9-repeat RNA in ALS patient samples, we first explored RNA-Seq datasets of 10 C9-ALS patient iPSN[50-51] lines (Supp. Table 1, Figure 1A) containing >50 G4C2 repeats within the Answer ALS Data Consortium database[52]. In all 10 C9-ALS patient samples, we observed no RNA reads aligning to or spanning the G4C2 repeat annotated within the first intron of the C9ORF72 gene (Fig. 1B).

To ensure that the lack of read coverage over the G4C2 repeat was not due to an alignment error, we extracted all reads that contained at least three copies of the G4C2 repeat from the unaligned sequence reads, aligned them to the human genome, and a custom reference genome containing all transcript variants of the C9ORF72 gene (see **Methods**). Out of a total of 2,317 Read 1 reads containing three consecutive copies of the G4C2 repeat, 96.42% (2234 reads) aligned to the human genome, while from a total of 3,207 Read 2 reads containing three G4C2 repeats, 97.75% (3135 reads) aligned to the human genome (Fig. 1C, see Supp. Fig. 1 for results per patient). However, only two reads from Read 1 and five reads from Read 2 align to C9ORF72 (Fig. 1C). Notably, across all ten C9 patients, the seven reads that align to C9 do not map to the region spanning or including the C9 repeat. In addition, the 5,369 G4C2-containing reads that mapped to the human genome

aligned to other human genes containing multiple copies of the G4C2 repeat (Fig. 1D), including RRP36 (4 copies), RGS14 (9 copies), CDKN1C (4 copies), PRRC2B (4 copies), and HUWE1 (7 copies). The large number of genome-wide (but non-C9) matches of the G4C2 repeat may explain the previously reported RNA FISH results using probes targeting 3-4 copies of the G4C2 repeat[19-24], in particular because these other G4C2 repeat-containing genes are each at least 4-fold more highly expressed than C9ORF72.

Next, we considered the possibility that the few reads (3.58% from Read 1, 2.25% from Read 2) that failed to align to the human genome were comprised entirely of G4C2 repeats. To explore this, we visually inspected each of these unmapped reads and verified that none of these contained G4C2 expansions (Supp. Fig 1D). To account for potential antisense transcription, we performed the same analysis for reads containing three repeats of the antisense sequence (G2C4) and similarly observed that these reads fail to align to the C9ORF72 gene (Supp Fig. 1E).

Finally, to ensure that our observations are not due to a specific limitation of C9-ALS patient iPSN models, we investigated the expression of the C9-repeat RNA in postmortem brain tissue collected from C9-ALS patients. Because ALS is predominantly a disease associated with neurodegeneration of the motor cortex[14-16], we analyzed previously published RNA-seq data generated from neurons and glia from the motor cortex of C9-ALS patients [57]. We explored the read coverage over the C9-repeat in neurons from the motor cortex of 5 C9-ALS patients by aligning all reads containing G4C2 sequences to the C9ORF72 gene and the human

genome and observed no reads overlapping the C9-repeat in any of the postmortem samples from neurons (Fig. 1E). Because there is evidence implicating glial cells in ALS pathology [13-16], we explored whether the C9-repeat RNA was expressed in glia from the motor cortex and similarly observed no read coverage overlapping the C9-repeat (Fig 1E).

Together, these data indicate that there is no detectable read coverage for the C9-repeat RNA in C9-ALS patient iPSN or postmortem models.

2.1.2 FAILURE TO DETECT THE C9-REPEAT RNA IS NOT DUE TO TECHNICAL ISSUES WITH RNA-SEQUENCING

We next explored whether the lack of detectable C9-repeat RNA reads may be a result of technical issues in one of the steps of RNA sequencing library preparation. First, we investigated the possibility that the C9-repeat cannot be amplified by PCR, a necessary step in library preparation. If true, DNA sequencing from these same patients should also exhibit a lack of DNA read coverage over the C9-repeat. However, we observed read coverage over the genomic DNA region containing the C9-repeat that was comparable to the coverage observed across the remainder of the C9ORF72 genomic locus (Fig 2A).This demonstrates that, when present, the C9-repeat can be amplified by PCR and successfully sequenced.

An alternative possibility is that another aspect of RNA library preparation, such as reverse transcription, might explain the lack of detection of the C9-repeat in RNA,

but not DNA, sequencing. To explore this, we expressed a reporter construct containing Exon 1A and Intron 1A of C9ORF72 followed by ~300 copies of the G4C2 repeat ($G4C2_{300}$, gift from Haeusler lab) in HEK293T cells and performed RNA-seq[53] on total RNA extracted from these cells alongside three C9-ALS patient iPSN lines (with ~1000-4000 repeats each[11]; Fig. 2B, Supp. Table 2, see **Methods**). In the patient-derived iPSN lines, we observed 0 reads mapping over the C9-repeat within the C9ORF72 gene (Fig. 2C). In contrast, we observed high-levels of read coverage over the C9-repeat within the ($G4C2_{300}$)-transfected HEK sample (~360 reads/base, Fig. 2C). These results demonstrate that the lack of C9-repeat RNA detection in iPSN lines is not due to technical issues during the RNA library preparation protocol.

Finally, to ensure that the lack of detection of the C9-repeat was not simply because of its low expression level, we used RNA antisense purification (RAP)[54] to enrich for the C9ORF72 RNA with probes concentrated around the genomic location of the G4C2 repeat (Supp. Fig 2). Using this strategy, for all captures we achieved a >1,000 fold enrichment of the C9ORF72 gene relative to other genes (Fig. 2D-E), yet we were still unable to detect any RNA reads aligning to the G4C2 repeat. In all, these data demonstrate that the C9-repeat RNA is not detectable in patient iPSN models nor post-mortem tissue, and that this is not due to technical limitations in library preparation or sequencing of the C9-repeat sequence.

2.1.3 TRANSCRIPTION OF THE C9 GENE PRIMARILY ORIGINATES AFTER THE C9-REPEAT

Because the C9-repeat RNA is proposed to occur due to the retention of this first intron (which contains the repeat expansion), the expression level of the first intron (estimated from the non-repeat sequences) should reflect the levels of any retained mature C9-repeat RNA even if there is an issue mapping the repeat itself. However, we observe the opposite; specifically that the number of reads overlapping the non-repeat regions of the first intron are significantly lower than that observed for the other introns within the C9ORF72 gene.

To explore the reasons for this, we investigated the relative usage of the different annotated transcript variants of the C9ORF72 gene. The C9ORF72 gene is annotated as containing two alternative transcription start sites (TSS) on either side of the C9-repeat (Figure 3A) [55]. Because transcription of the C9-repeat RNA would require transcription from the TSS upstream of the C9-repeat (uC9-repeat), we first assessed the expression level of transcripts initiating from this TSS across 13 different brain regions using data generated by GTeX (Genotype-Tissue Expression)[56] in healthy (non-ALS) brain samples (Supp. Fig. 3). Across the 13 brain regions, we observed >120-fold lower coverage over Exon 1A (included only in the uC9-repeat variant) relative to Exon 2 (included in both variants) in the GTeX data[56] (Supp. Fig. 3). Notably, for several brain regions in the GTeX data (including amygdala, hippocampus, and putamen), there are zero RNA sequencing reads over

Exon 1A[56] (Supp. Fig. 3D). These data highlight the extremely low expression of the uC9-repeat transcript variant in brain tissue, even in the non-ALS context.

We next investigated the expression of the uC9-repeat variant (Exon 1A) relative to the downstream of C9 (dC9-repeat) variant that initiates at Exon 1B. Specifically, we aligned RNA-sequencing data from C9-ALS patients directly to the two spliced variants of C9ORF72 (Figure 3B). For each patient, we computed the difference in coverage of the Exon 1A to Exon 2 junction (uC9) relative to the Exon 1B to Exon 2 junction (dC9) and normalized each of these to the coverage spanning the Exon 2 to Exon 3 junction (both). When we performed this analysis for the Answer ALS data from C9-ALS patient-derived iPSNs, we observed ~8-fold lower expression of the uC9-repeat variant (Exon 1A-Exon2) relative to the dC9-repeat variant (Exon 1B-Exon2) (Fig. 3B-3C). Similarly, we observed from 12 to 50-fold lower expression of the uC9-repeat variant relative to the dC9-repeat variant in RNA-sequencing data that we generated from our two patient-derived iPSN lines (Fig. 3B-3C).

To further investigate the low expression of the uC9-repeat variant, we mapped the 5’ start site of transcription in C9-ALS patient iPSNs. To do this, we purified total RNA, fragmented the RNA such that cleaved products contain a 5’-monophosphate (5’-P), and treated with a 5’-3’ exonuclease that specifically digests RNA with 5’-P but not RNA containing a 5’-methylguanasine cap (5’-cap) (Fig. 4A). The resulting library should be enriched for fragments containing the 5’-cap. We sequenced this library (where the first read corresponds to the 3’-end of the

randomly fragmented RNA and the second read corresponds to the protected 5'-cap on the RNA fragment) and quantified the enrichment of the 5'-end of genes compared to their RNA-sequencing profiles. We observed a clear enrichment at the annotated 5'-ends of individual control genes (Fig 4B) and globally (Fig 4C).

We then quantified the coverage of the 5'-cap location of each RNA fragment (defined by the start of the second read) to define the TSS of the C9ORF72 gene (Fig 4D). We observed that ~90% of reads (110/123) were located over 1B and ~10% of reads overlapped Exon 1A. This corresponds to an ~9-fold increase in fragments spanning the Exon1B-Exon2 junction relative to the Exon1A-Exon2 junction (Fig 4E).

Together, these data demonstrate that the dC9-repeat variant, which excludes the C9-repeat, is the predominant variant transcribed in C9-ALS patient-derived iPSNs.

2.1.4 THE C9-REPEAT OCCURS WITHIN A PUTATIVE ENHANCER REGION

Even though the C9-repeat RNA is not transcribed, the C9-repeat expansion in patient DNA is the most frequently occurring genetic cause of ALS[5]. Therefore, we explored whether an alternative regulatory mechanism might explain the role of this repeat expansion in C9-ALS. Specifically, because C9-ALS patient cells have been reported to undergo hypermethylation of DNA at Exon 1A and a ~50%

reduction of C9ORF72 mRNA levels[9,11,58-63], we hypothesized that Exon 1A might function as an enhancer that regulates the transcription of the C9ORF72 mRNA.

Enhancers are generally associated with the localization of histone acetylation (e.g. H3K27ac) and low levels of bidirectional transcription [64-67]. To investigate the possible regulatory role of Exon1A as an enhancer, we used our ChIP-DIP (ChIP done-in-parallel) method [68] to measure a panel of histone modifications that demarcate distinct regulatory activity including promoters (H3K4me3), enhancers (H3K27ac), and gene bodies (H3K79me1/3) (Supp. Fig. 4). We observed strong enrichment of H3K4me3 and H3K27ac over the 5' end of the C9ORF72 gene containing the repeat expansion (Fig 5A, Supp. Fig 4). We also observed similar H3K27ac enrichment over this region in ALS patient postmortem tissue from the motor cortex (postmortem data re-aligned and peak-calling analysis performed from raw sequencing data previously published[57]), emphasizing the physiological relevance of this observation (Fig 5A). Consistent with the fact that enhancers are often associated with bidirectional transcription, we observed bidirectional transcription at Exon 1A for all C9-ALS patient samples (Fig. 5B, Supp. Fig. 5). Exon 1A is the only region within the C9ORF72 gene that contains comparable sense and antisense transcription (Fig. 5B, Supp. Fig. 5). Importantly, this antisense transcription does not produce antisense transcripts of the C9-repeat RNA, indicating that the antisense C9-repeat RNA is not expressed.

Together, these results suggest that the genomic region containing the C9-repeat expansion ("Exon 1A") is likely a promoter-proximal enhancer element. Because

enhancers are often transcribed at low levels [64-67], this would explain the low levels of transcription observed over Exon 1A.

2.1.5 THE C9-REPEAT IS ASSOCIATED WITH REDUCED ALLELE-SPECIFIC EXPRESSION OF THE C9ORF72 GENE

The C9-repeat is contained within a CpG island and the repeat expansion itself is a GC rich sequence. CpG islands can undergo DNA methylation, a process which generally acts to silence transcription[59]. Previous studies have observed an increased level of DNA methylation over the genomic region containing Exon 1A that is specific to the allele containing the C9-repeat expansion [9,11,58-63]. If the Exon 1A-containing region indeed functions as an enhancer, then increased DNA methylation could explain the ~50% reduction observed in total C9ORF72 mRNA levels and indicate that the C9-repeat allele may be epigenetically silenced, consistent with a loss-of-function mechanism for pathogenesis [11]. Indeed, this is the mechanism by which FMR1 is silenced via methylation of a GC-rich repeat expansion, leading to Fragile X Syndrome [69-71].

To explore this possibility, we analyzed the allele specific expression of the C9ORF72 mRNA in patient and control samples. Specifically, we quantified allele-specific expression by first identifying all SNPs within each of the 10 C9-ALS

patient and 10 healthy control samples from the Answer ALS consortium [52]. For each SNP, we computed the proportion of each allelic variant in the genomic DNA sequence (DNA variation) and in the RNA sequence (RNA variation) within each sample (Fig. 6A). Focusing on the genomic region containing the C9ORF72 gene, we observed that the allelic variation in DNA across all SNPs and ALS patients was largely centered at 50%, as expected (Supp. Fig 6). In contrast, within the RNA reads we observed a strong bias towards expression from only one allele (Figure 6A), with an ~2-fold decrease in RNA-to-DNA allelic usage (Fig 6B). We did not observe this RNA-to-DNA allelic usage shift in the healthy control samples where both DNA and RNA allelic frequencies were largely centered at ~50% across all SNPs (Fig. 6B, Supp. Fig 6).

To ensure that this effect is not a general property of these patient samples, we computed the DNA and RNA allele frequencies across the entire genome for both ALS and healthy patient groups. We observed no global difference in the RNA-to-DNA allele ratio in the ALS patient samples relative to controls (Supp. Fig 6). As an example, focusing on other specific control genes, such as GAPDH, we observe comparable DNA and RNA allelic usage in patient and control samples (Fig. 6A-B). We similarly observed comparable DNA and RNA allelic usage for genes flanking the C9ORF72 gene (Supp. Fig 6A).

This significant shift in allele-specific expression of the C9ORF72 mRNA suggests that the presence of the C9-repeat in patients acts to suppress allele-specific transcription from the mutant expanded allele.

2.2 DISCUSSION

Our observations that the C9 repeat RNA is not transcribed (in either the sense or antisense orientation) challenge the prevailing gain-of-function models of pathogenesis in C9-ALS, because neither the repeat RNA itself nor dipeptide repeat proteins (DPRs) translated from this RNA can explain disease pathology. These observations may explain why several prominent clinical trials targeting the C9 repeat RNA have failed to show therapeutic effects [47].

Specifically, our findings call into question the source of previously reported DPRs, as the C9 repeat RNA cannot serve as a template for their translation if it is not transcribed. Since many prior studies rely on the overexpression of these DPRs to measure their proposed pathological effects[38-43], a critical reevaluation of their endogenous expression levels in the correct physiological context is necessary. It is possible that the reagents used to detect DPRs are measuring the same amino acid repeat sequences present within endogenous proteins (Supp. Fig. 7) and that some of these endogenous proteins might even be overexpressed or upregulated specifically in the ALS context. Furthermore, production of isolated DPRs may be a pathogenic phenotype specific to C9-ALS compared to control, but their production may be due to other aberrant cellular processes that are independent of the C9 repeat RNA.

Taken together, our results suggest that the DNA region containing the C9-repeat may function as a regulatory element that, upon expansion, leads to increased DNA methylation and transcriptional silencing of C9ORF72 expression from the mutant allele (Fig. 6C). This allele-specific silencing would explain the previously reported ~50% reduction of C9ORF72 mRNA and protein levels in C9-ALS patients and our previous demonstration that a 50% reduction in C9ORF72 protein leads to neurodegeneration in C9-ALS patient cells [11]. This proposed mechanism closely parallels the mechanism underlying other neurodegenerative disorders such as Fragile X Syndrome, where a CGG repeat expansion (>200 CGG repeats) in the 5'-UTR of the FMR1 gene leads to DNA methylation and transcriptional silencing of the FMR1 gene [69-71]. As the G4C2 expansion in C9-ALS is similar in position (promoter-proximity) and sequence composition (CG-rich), this hypermethylation and transcriptional silencing mechanism might be a shared pathological mechanism.

While the current gain-of-function models based on the C9-RNA (toxic RNA or DPRs) cannot explain ALS pathogenesis, it remains possible that ALS pathogenesis is due to a gain-of-function due to the repeat expansion of the DNA element. Specifically, the G4C2 expansion in DNA could create new functions for this enhancer in regulating the expression of other gene targets that are important for ALS (e.g., nuclear transport, RNA splicing and processing, or protein trafficking/aggregation/turnover). Although this remains to be explored, we note

that the C9ORF72 gene is neighbored by the gene encoding interferon K (IFNK), a protein that is involved in inflammation, a known hallmark of ALS.

Together, our results argue for a critical re-evaluation of the mechanisms by which repeat expansions leads to neurodegeneration in ALS and other contexts.

2.3 MAIN FIGURES

FIGURE 1

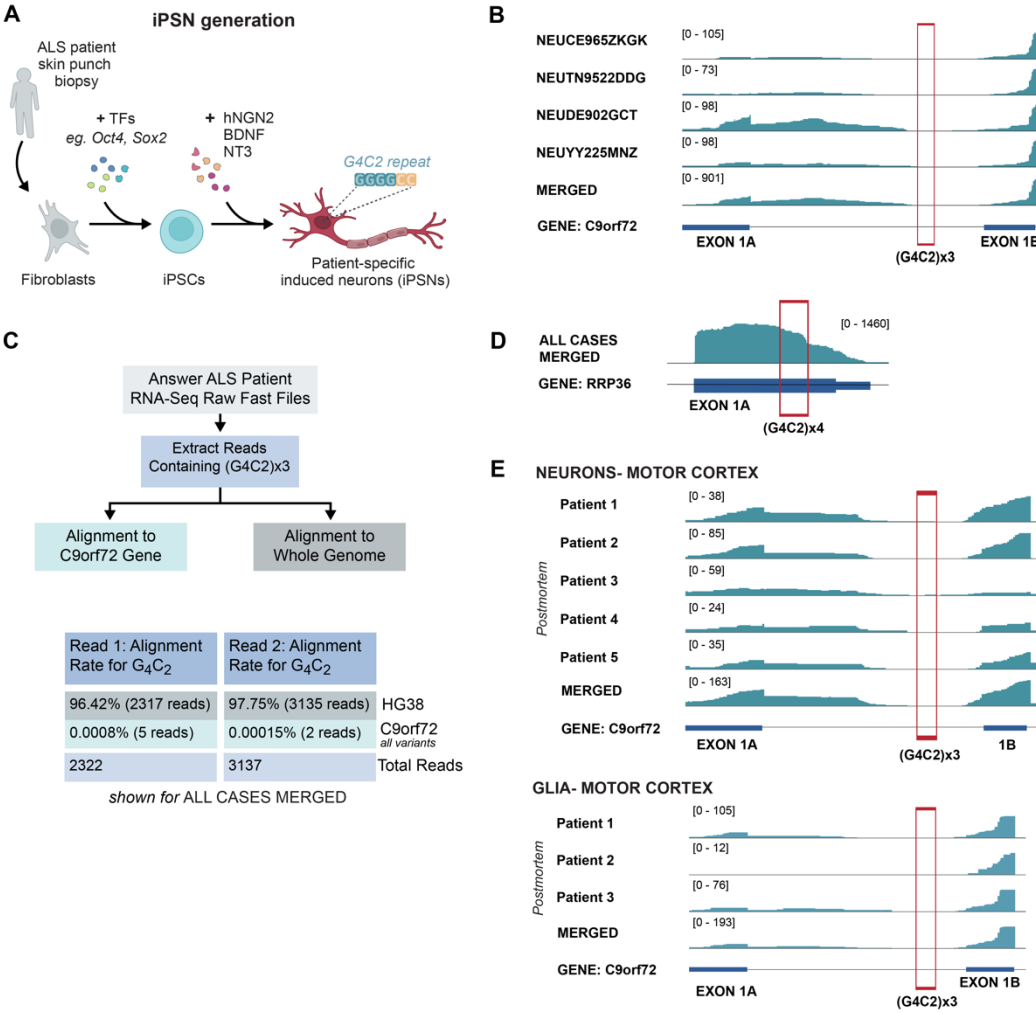


FIGURE 1. The C9-repeat RNA is undetectable in C9-ALS patient-derived iPSNs and postmortem tissue.

A. Schematic of iPSN generation from patient fibroblasts. Fibroblasts from C9-ALS patients and healthy control patients are obtained by skin punch biopsy, reprogrammed into iPSCs using a defined cocktail of transcription factors (eg. Oct4, Sox2), and differentiated into iPSNs using hNGN2 (+BDNF, NT3). Final iPSNs are genetically identical to the patients from which they are derived.

B. RNA-sequencing tracks for the C9ORF72 gene (zoomed in on 5'end of the gene) in selected individual C9 patients (Top, 4 individual patients shown) obtained from Answer ALS. Bottom track shows RNA-sequencing tracks merged for all 10 C9 patients with >50 repeats of G4C2. Red box denotes the position of the G4C2 sequence in the C9ORF72 gene (denoted in the hg38 reference genome as 3 consecutive G4C2 repeats).

C. Schematic of G4C2 subset reads analysis pipeline. First, reads containing at least 3 copies of G4C2 are extracted. Then these reads are aligned to the 3 known transcription variants of the C9ORF72 gene AND the whole human hg38 genome. Alignment rates are shown for all C9-ALS patient cases merged. (Unmapped reads were saved for downstream analysis.)

D. RNA sequencing track from the merge of all C9-ALS cases from Answer ALS for the RRP36 gene shows detectable read coverage over the G4C2 sequence stretch. In addition to RRP36, other non-C9 genes that include G4C2 repeat stretches (containing sequence stretches of 3 or more consecutive G4C2 repeats include GS14, HUWE1, CDKN1C, and PRRC2B.

E. (Top) RNA sequencing tracks in neurons from postmortem motor cortex over the C9ORF72 gene. RNA alignment tracks (top 5 tracks) are shown for 5 individual patients over C9ORF72 alongside the merged RNA alignment track across all patients. Red box denotes G4C2 insertion site. (Bottom) RNA sequencing tracks in glia from postmortem motor cortex over the C9ORF72 gene. RNA tracks (top 3 tracks) are shown for 3 individual patients over C9ORF72 alongside the merged RNA track across all patients. Red box denotes G4C2 insertion site.

FIGURE 2

FIGURE 2. Failure to detect C9-repeat RNA is not due to technical issues with RNA-sequencing.

A. (Top) Whole genome DNA sequencing (orange) alongside RNA-sequencing (blue) zoomed in to the 5'-end of the C9ORF72 gene a re shown for the same C9 patient sample (Case NEUCE965ZKGK). (Bottom) Whole genome DNA sequencing for all C9 patients (>50 repeats of G4C2) merged (orange), compared to RNA sequencing for all C9-ALS cases merged (blue). The red box denotes the position of the G4C2 insertion sequence in the C9ORF72 gene (denoted in the hg38 reference genome as 3 consecutive G4C2 repeats).

B. Schematic of previously published RNA-tag seq protocol[53] for pooled RNA library preparation for C9 patient-derived iPSNs (2 iPSN patients included, 2 replicates each) and HEK-293s expressing (G4C2)x300 (2 replicates).

FIGURE 2 (cont.)

C. (Top) RNA alignment to the (G4C2)x300 plasmid sequence, containing pCMV promoter, Exon1A (same sequence as in human genome), Intron1A (same sequence as in human genome), and the G4C2 sequence stretch. Reads are extracted over the Intron 1A-containing region of transfected plasmid (immediately 5' upstream of the G4C2 stretch) and the G4C2 region. Coverage values (reads/base) for RNA sequencing over the G4C2 containing region relative to coverage over Intron 1A-containing region. (Bottom) RNA tracks over C9ORF72 gene for patient-derived iPSNs (2 lines shown).

D. (Top) RNA tracks over C9ORF72 gene for RNA sequencing (input) compared to RAP captures. (Bottom) RNA alignments tracks over GAPDH housekeeping gene for RNA sequencing (input) compared to RAP captures. These experiments were repeated twice (n=2) with similar results.

E. RAP-RNA enrichment values for the C9ORF72 transcript in RAP captures normalized relative to input.

FIGURE 3

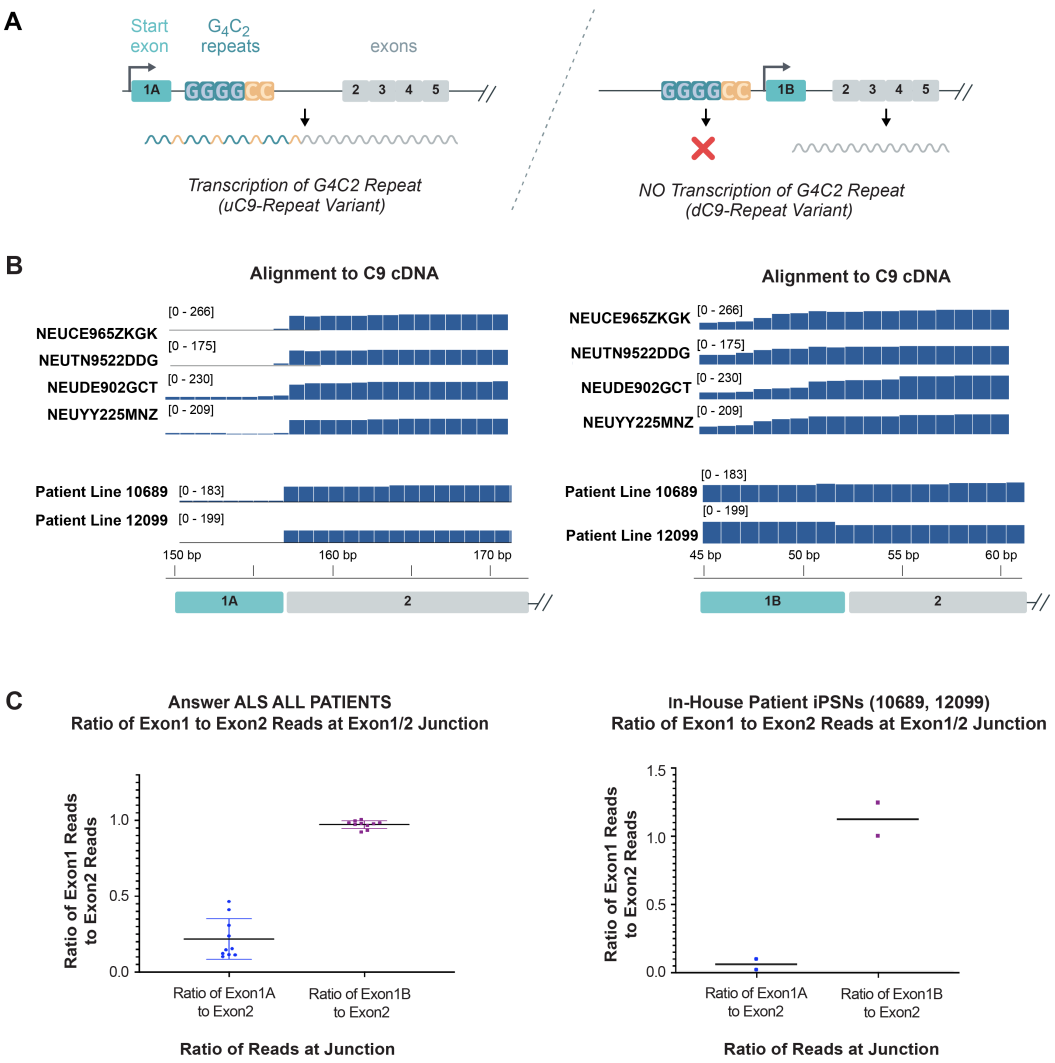


FIGURE 3. Transcription of the C9ORF72 variant containing the C9-repeat is extremely low compared to the variant lacking the C9-repeat.

A. Annotated transcript variants of C9ORF72 gene. Top schematic features TSS upstream of Exon1A, resulting in transcription of the C9 repeat expansion (uC9 variants). The bottom variant features TSS at Exon1B (dC9 variant), thus the C9 repeat expansion is not transcribed.

B. RNA sequencing alignments to the cDNA of C9ORF72 variants (shown for both Exon1A and Exon1B-initiating transcription variants). (Top, left) C9 cDNA alignments for 4 representative Answer ALS patients for Exon1A-initiating variant. (Top, right) C9 cDNA alignments for 4 representative Answer ALS patients for Exon1B-initiating variant. (Bottom, left) C9 cDNA alignments for 2 in-house C9-ALS patient iPSN lines for Exon1A-initiating variant. (Bottom, right) C9 cDNA alignments for 2 in-house C9-ALS patient iPSN lines for Exon1B-initiating variant.

C. Summary plots showing Exon1B relative to Exon1A expression in all Answer ALS lines (left) and all in-house patient iPSN lines (right). For each patient, read coverage (of the end of Exon1A relative to the start of Exon2) at the Exon1A-Exon2 junction was quantified and then normalized to the read coverage at the Exon2-Exon3 junction. Similarly, the read coverage at the Exon1B-Exon2 junction was quantified and normalized to read coverage at the Exon2-Exon3 junction. Then, for each individual patient line, the normalized read coverage at Exon1A relative to Exon2 was plotted and compared to the normalized read coverage at Exon1B relative to Exon2. .

FIGURE 4

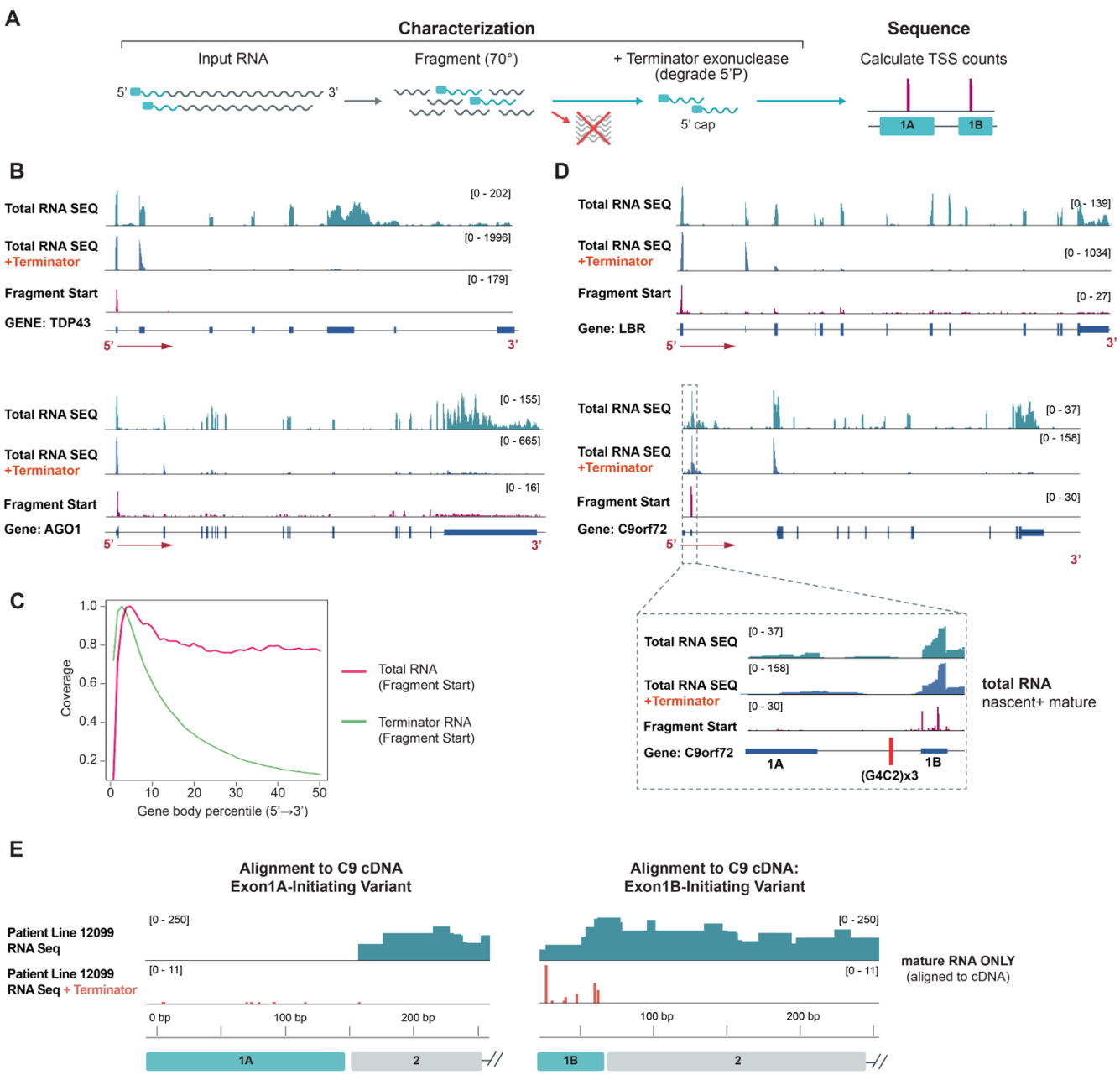


FIGURE 4. Transcription of C9ORF72 primarily originates after the C9-repeat .

A. Schematic of experiment to map transcription start sites. First, total (input) RNA from C9 patient iPSNs is is fragmented at 70C. Then the Terminator exonuclease is used to degrade 5'P and enrich for the 5'ends of all transcripts from fragmented total RNA.

B. RNA tracks for C9-ALS patient iPSN line 12099 (generated for this study); total RNA sequencing is shown for control genes TDP43 and AGO1 compared to alignment tracks for Terminator treated RNA. The bottom track for each gene shows the plotted start of the RNA fragment from the Terminator treated RNA data, thus denoting the TSS at each gene. These experiments were repeated 3 times with similar results. Tracks for replicates were merged for final analysis.

FIGURE 4 (cont.)

C. Signal profile plots across all genes for input RNA (pink) and Terminator treated RNA (green) for the same patient line (12099). Profile plots were generated from the RNA fragment start data, from total RNA sequencing and Terminator treated RNA sequencing.

D. RNA tracks for C9-ALS patient iPSN line (12099); total RNA vs Terminator treated RNA is shown for control gene LBR (Top) compared to C9ORF72 (Bottom). As shown for control genes in Fig 4B, the location of the start of the RNA Fragment corresponds to the TSS. Zoom-in shows C9ORF72 TSS at Exon1B. Red denotes G4C2 insertion site.

E. RNA sequencing alignments to C9ORF72 cDNA variants (shown for both Exon1A and Exon1B-initiating transcription variants). (Left) C9 cDNA alignments for RNA sequencing input library (top) alongside fragment start plot of Terminator-treated RNA library (bottom) for Exon1A-initiating variant. (Right) C9 cDNA alignments for RNA sequencing input library (top) alongside fragment start plot of Terminator-treated RNA library (bottom) for Exon1B-initiating variant. All data corresponds to patient #12099..

FIGURE 5

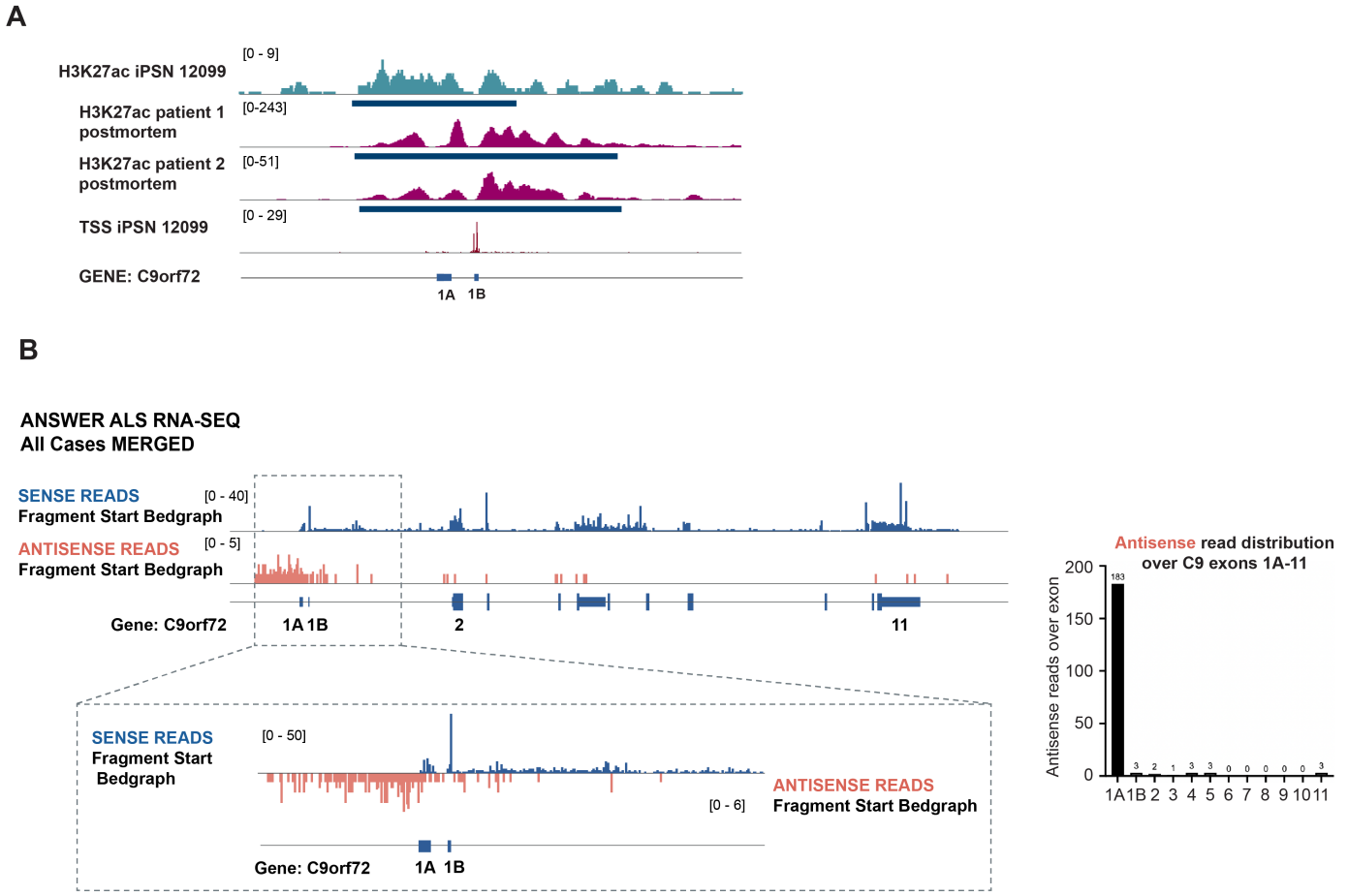


FIGURE 5. The C9-repeat occurs within a putative enhancer region.

A. (Top) ChIP-DIP coverage for H3K27ac over the 5'end of C9ORF72. (Bottom, middle two tracks) Re-aligned and re-analyzed H3K27ac profiles in C9-ALS postmortem neurons from motor cortex (raw data previously published[57]) shown over the 5'end of C9ORF72. (Bottom, bottom track) Terminator-treated RNA sequencing data in the same in-house C9 patient line #12099 as used for ChIP-DIP. Peaks (called using HOMER) are underlined for all tracks(dark blue).

B. (Left, top) Second read pileups (calculated from total RNA sequencing) for all Answer ALS C9-ALS cases (>50 repeats of G4C2) merged, demonstrating sense (blue) and antisense (orange) transcription over the C9ORF72 gene. (Left, bottom) Zoom-in of sense (blue) and antisense (orange) transcription at the 5'-end of the C9ORF72 gene at Exon1B and Exon1A. (Right) Distribution of antisense reads across all Exons of C9ORF72 gene demonstrates only background-level of antisense reads at Exon1B. .

FIGURE 6

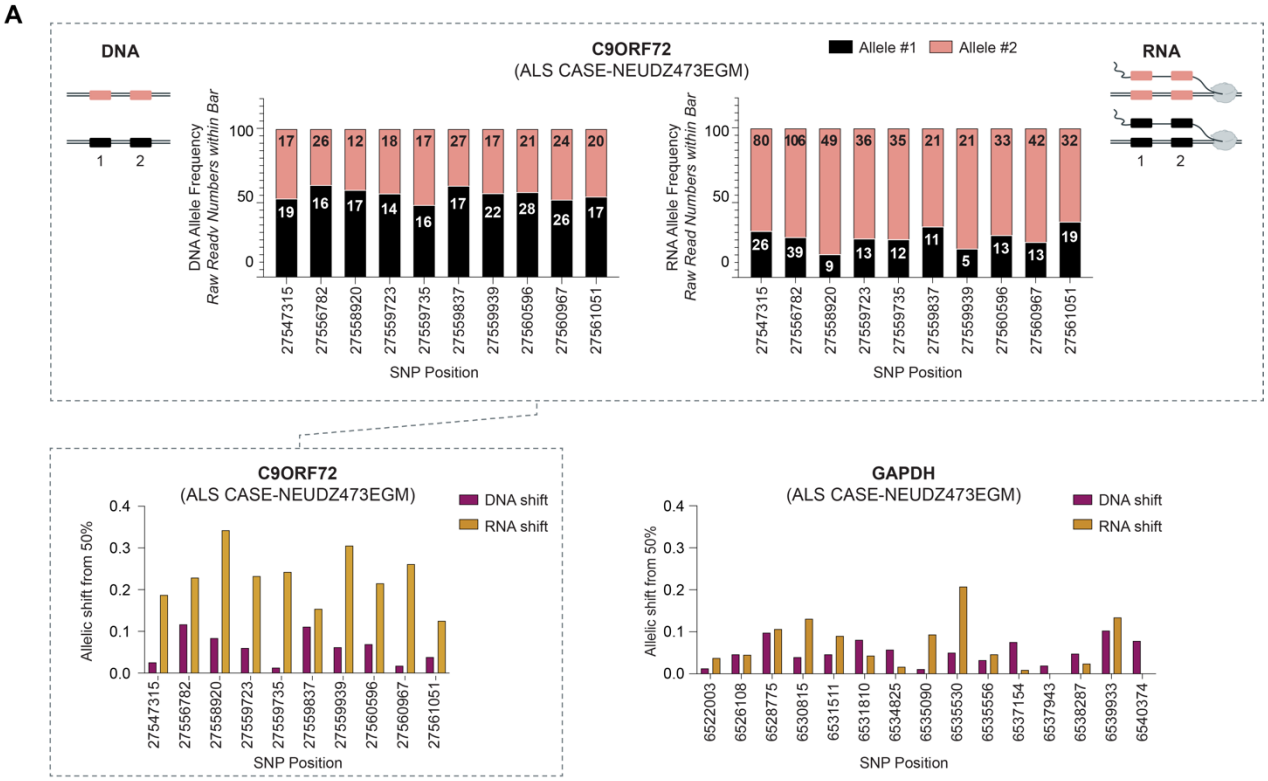


FIGURE 5. The C9-repeat is associated with reduced allele-specific expression of the C9ORF72 gene..

A.(Top) Schematic depicts possibilities for allele 1 vs. allele 2 for DNA and RNA for any given patient. (Middle) Percentage of reads aligning to allele 1 or allele 2 for each SNP within the C9ORF72 gene is shown for DNA (left) or RNA (right) for all SNPs for one sample C9-ALS line (NEUDZ473EGM from Answer ALS). (Bottom) For this same patient line, allelic shift in DNA and RNA calculated for SNPsover C9ORF72 (left) compared to SNPs over GAPDH (right).

(legend continued on next page)

FIGURE 6 (cont.)

B. (Allelic shift in RNA relative to DNA (shown separately) computed for all SNPs, shown for 10 C9-ALS lines pooled compared to 10 healthy control lines pooled, over C9ORF72 (left) and GAPDH (right).

C. Schematic of proposed mechanism. Upon expansion to pathogenic length, the G4C2 DNA repeat leads to increased DNA methylation and transcriptional silencing of C9ORF72 expression from the mutant allele.

2.4 SUPPLEMENTARY MATERIAL

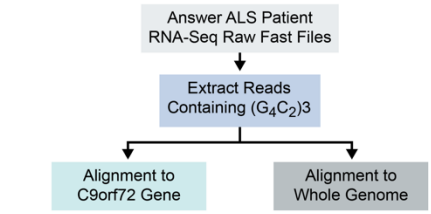
SUPPLEMENTARY FIGURE 1

A

PATIENT LINE ID	Age at Onset	G4C2 Repeat Length
CASE-NEUCE965ZGK	51	131
CASE-NEUDE902GCT	63	901
CASE-NEUDZ473EGM	64	779
CASE-NEUEM720BUU	60	858
CASE-NEUFV237VCZ	NA*	993
CASE-NEUGH995TFK	50	284
CASE-NEUJX990GR5	54	679
CASE-NEUNL415AW	51	321
CASE-NEUTN952DDG	55	741
CASE-NEUYY225MNZ	NA	NA

*NA indicates "Not Available" in Answer ALS Metadata

B



	Read 1: Alignment Rate for G_4C_2	Read 2: Alignment Rate for G_4C_2
HG38	96.61%	99.60%
C9orf72 *all variants*	0%	0%

shown for CASE-NEUDE902GCT

C

ALIGNMENT RATE FOR (G4C2)x3

PATIENT LINE ID	READ1 to HG38	READ2 to HG38	READ1 to C9 CUSTOM	READ2 to C9 CUSTOM
CASE-NEUCE965ZGK	97.12%	98.75%	0%	0.25% (1 read/399 reads)
CASE-NEUDE902GCT	96.61%	99.60%	0%	0%
CASE-NEUDZ473EGM	97.67%	99.01%	0.29% (1read/344 reads)	0%
CASE-NEUEM720BUU	95.13%	96.93%	0%	0.38% (1read/261 reads)
CASE-NEUFV237VCZ	90.74%	94.87%	1.85% (1 read/ 54 reads)	0%
CASE-NEUGH995TFK	97.92%	95.68%	0%	0.81% (3reads/370 reads)
CASE-NEUJX990GR5	94.76%	95.76%	0%	0%
CASE-NEUNL415AW	96.88%	98.71%	0%	0%
CASE-NEUTN952DDG	99.07%	98.97%	0%	0%
CASE-NEUYY225MNZ	98.97%	97.54%	0%	0%

Note: Aligned reads (>0%) are not aligned at G4C2 locus

D

EXAMPLE OF UNMAPPED READS (shown for CASE-NEUCE965ZGK)

Unmapped Reads from READ1 (to HG38)

CCCCGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGG
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACTACTGCTCATCTCGTATGCC

Unmapped Reads from READ2 (to HG38)

CCGGAGCGGGTCGCGGCCCGGCGCGGGCGGGGCCGGGGCCGGGGCCCCCAAG
CGCCAACCCCCCCGGGGGGGCCCCCCCCCCCCCCCCGGGACGGGGACAA — **(chromosome 10)**

GTCCTCCTCCTCCCCCGTCTCCGCCCCCCGCCCCCGGGCCAGCCGCCGGGGGGCGCCCG
GGGGGGGGGGGGGGCCGGGGCCGGGGCCGGCGCGGGCGCCGG — **(chromosome 22)**

GTACTTGGGAGGCTGGAGAGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCG
GGGCCGGGGTCAGGGTCGGACCCCGAGTGGGTGAGGCTGGCC — **(chromosome 10)**

GCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCC
TGTTGGGGGGGGCGACGGGTAGGGGGGGGGGGGTGGGCGGTGGG — **(chromosome X)**

GCTGTACTTGGGAGGCTGGAGAGCCGGGGCCGGGGCCGGGGCCGGGGCCGGGGCCGG
GGCCGGGGTCAGGGTCGGACCCCGAGTGGGTGAGGCTGGCCCGG — **(chromosome 13)**

SUPPLEMENTARY FIGURE 1 (cont.)

E

	ALIGNMENT RATE FOR (G4C2)x3			
PATIENT LINE ID	**READ1 to HG38**	**READ2 to HG38**	**READ1 to C9 CUSTOM**	**READ2 to C9 CUSTOM**
CASE-NEUCE965ZGK	98.45%	96.7%	0%	0%
CASE-NEUDE902GCT	99.40%	97.91%	0%	0%
CASE-NEUDZ473EGM	98.83%	94.47%	0%	0.16% (1reads/614 reads)
CASE-NEUEM720BUU	98.98%	93.8%	0%	0.19% (1read/519 reads)
CASE-NEUFV237VCZ	98.41%	96.35%	0%	1.46% (2reads/137 reads)
CASE-NEUGH995TFK	97.12%	95.59%	0.82% (2reads/243 reads)	0%
CASE-NEUJX990GR5	94.59%	92.48%	0.45% (1reads/222 reads)	0%
CASE-NEUNL415AW	100%	93.12%	0%	0%
CASE-NEUTN952DDG	98.7%	95.92%	0%	0%
CASE-NEUYY225MNZ	98.7%	98.13%	0%	0%

Note: Aligned reads (>0%) are not aligned at G4C2 locus

SUPP. FIG 1. Alignment of G4C2x3 and G2C4x3-containing reads to hg38 genome and custom C9 genome of all transcription variants.

S1A. Patient line ID numbers for all C9-ALS cases (>50 G4C2 repeats) used from Answer ALS for G4C2 subset analysis. ID numbers are shown alongside age of onset and G4C2 repeat length for each patient (obtained from Answer ALS metadata download; NA = not available). .

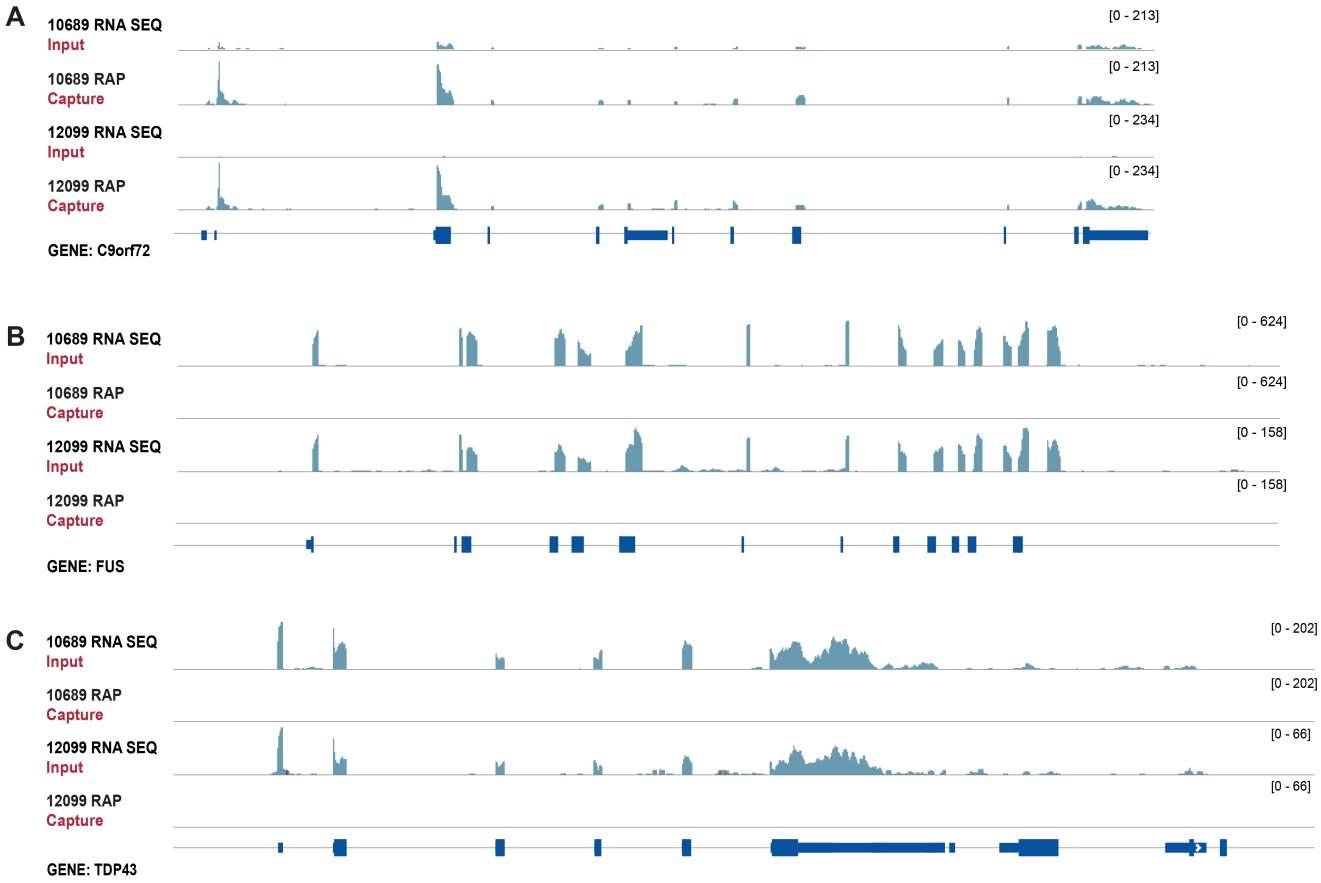
S1B. (Top) Schematic of analysis pipeline for G4C2 subset analysis. A custom genome was built including all 3 known transcript variants of the C9ORF72 gene. Alignments for the G4C2 subset reads were run against this custom C9 genome map and hg38 genome (whole human genome). (Bottom) Percentage of aligned reads are shown from Read 1 and Read 2 for one sample ALS patient (Case #NEUDE902GCT, 901 G4C2 repeats).

S1C. Alignment percentages for the (G4C2)x3 string motif. From the raw sequencing files, reads containing this motif were extracted from the Read 1 raw fastq file and Read 2 raw fastq file for every patient. Those extracted reads were then aligned to the custom C9 genome map and hg38 genome. Any reads that aligned to the C9 custom genome map were each visually inspected and none aligned to the G4C2 locus.

S1D. Unmapped reads were all inspected in case that G4C2 expansion reads were excluded in the alignment process and filtered out to unmapped reads. Upon closer inspection, none of the unmapped reads for all patients contained entirely G4C2 expansions. The unmapped reads from Read 1 and Read 2 are shown for a sample patient. The unmapped reads align to other genes in the human genome as denoted by the chromosome numbers on the right.

S1E. Alignment percentages for the (G2C4)x3 string motif. From the raw sequencing files, reads containing this motif were extracted from the Read 1 raw .fastq file and Read 2 raw .fastq file for every patient. Those extracted reads were then aligned to the custom C9 genome map and hg38 genome. Any reads that aligned to the C9 custom genome map were each visually inspected- none aligned to the G4C2 locus.

SUPPLEMENTARY FIGURE 2



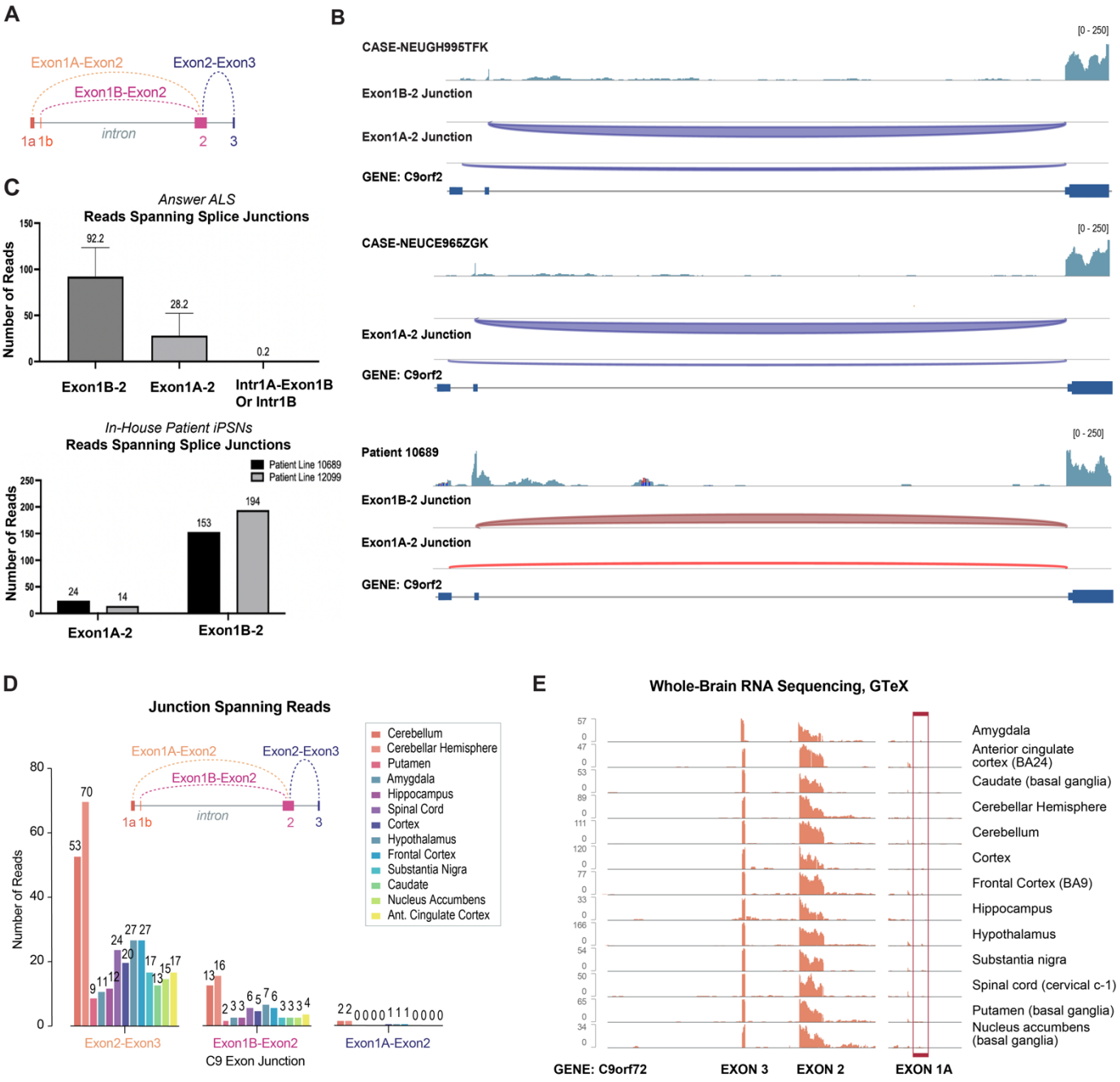
SUPP. FIG 2. RAP enrichment of C9-RNA specifically enriches for C9 and depletes control genes.

S2A. In RAP, biotinylated antisense probes are hybridized to the C9 target RNA and captured on beads with denaturing washes to select against nonspecific interactions. Captured RNA is prepared as a sequencing library, sequenced, and analyzed relative to inputs. RNA-seq (Input) tracks over C9ORF72 gene compared to C9 capture tracks over C9ORF72 gene, shown for 2 in-house C9-ALS patient iPSN lines 10689 and 12099.

S2B. Control genes for RAP-RNA experiment; RNA-seq (input) and C9 capture tracks shown over FUS control genes for 2 in-house C9-ALS patient iPSN lines 10689 and 12099. Specificity of captures is shown by lack of FUS reads in RAP captures relative to inputs.

S2C.Control genes for RAP-RNA experiment; RNA-seq (input) and C9 capture tracks shown over TDP43 control genes for 2 in-house C9-ALS patient iPSN lines 10689 and 12099. Specificity of captures is shown by lack of TDP43 reads in RAP captures relative to inputs.

SUPPLEMENTARY FIGURE 3



SUPP. FIG 3. Splice junction-spanning reads and RNA-sequencing data from C9-ALS patient iPSNs and whole-brain RNA sequencing highlight low usage of Exon1A in physiological contexts.

S3A.Schematic of possible splice junctions at 5' end of C9ORF72 gene (Exon1A-Exon2, Exon1B to Exon2, Exon2 to Exon3).

S3B.(Top two tracks) Splice junction-spanning reads for C9ORF72 gene from RNA sequencing for 2 sample C9-ALS cases from Answer ALS. (Bottom) Splice-junction spanning reads from RNA sequencing for 1 in-house C9-ALS iPSN line. Data shows very few reads spanning Exon1A-Exon2 even for spliced reads, highlighting low usage of Exon1A.

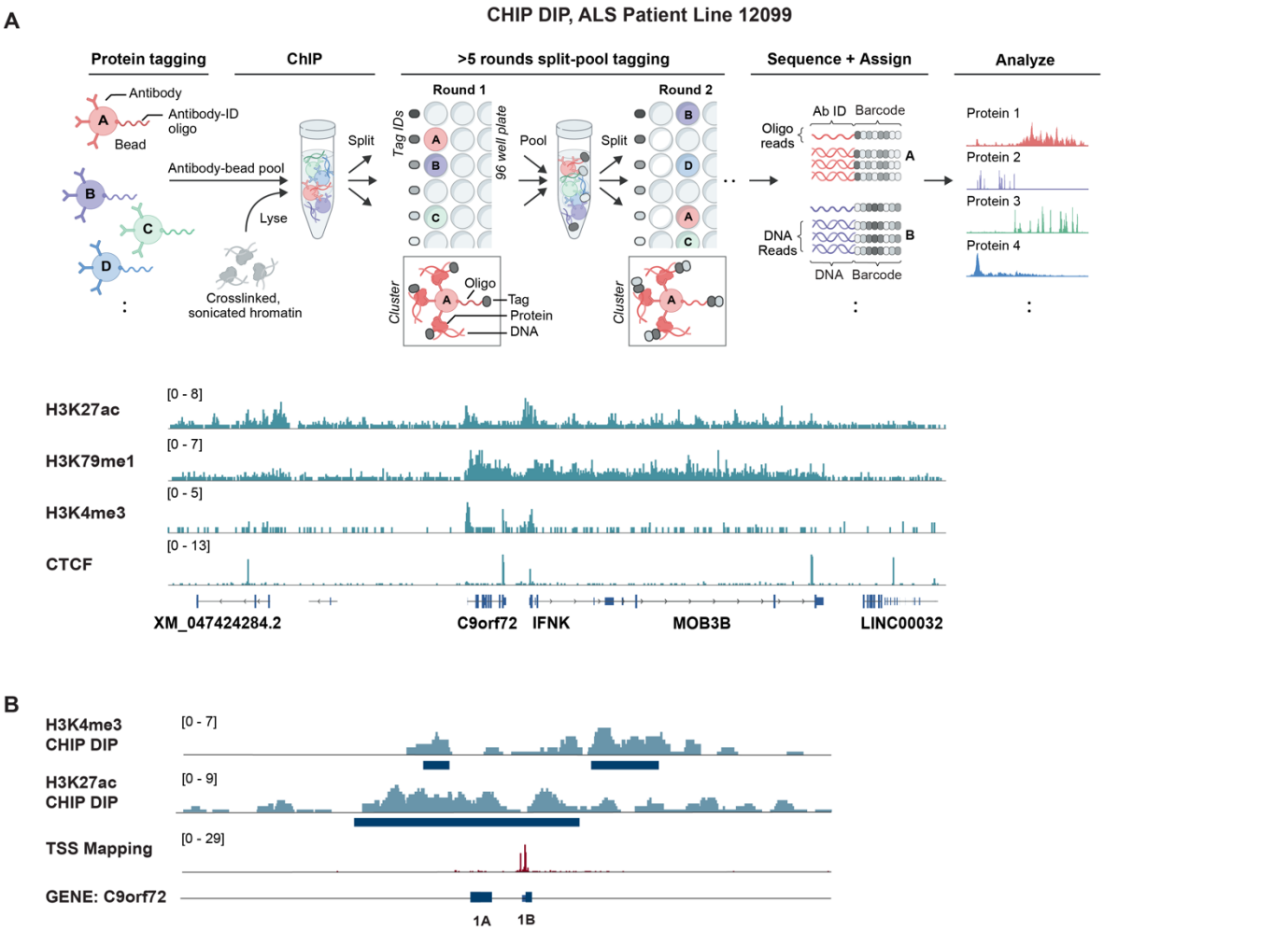
SUPPLEMENTARY FIGURE 3 (cont.)

S3C. Quantification of splice junction spanning reads (Exon1A, Exon1B, Exon2 of C9ORF72) across all C9-ALS Answer ALS patient cases (top) and 2 lines of in-house C9-ALS patient iPSN lines (bottom). Data shows very few reads spanning Exon1A-Exon2 even for spliced reads, highlighting low usage of Exon1A.

S3D. Splice-junction spanning reads from RNA sequencing (spanning Exon1A/Exon1B/Exon2 of C9ORF72, data obtained from GTEx expression portal 56), quantified in healthy postmortem brain tissue and split by brain region. Data shows no usage of Exon1A (even in the non-ALS context human brain).

S3E. RNA-sequencing tracks from whole-brain RNA sequencing (healthy patients) from GTex expression portal 56, separated by brain region. Tracks are shown zoomed in to the 5'end of the C9ORF72 gene. Red box shows undetectable read coverage over Exon1A across all brain regions (even in the non-ALS context human brain).

SUPPLEMENTARY FIGURE 4

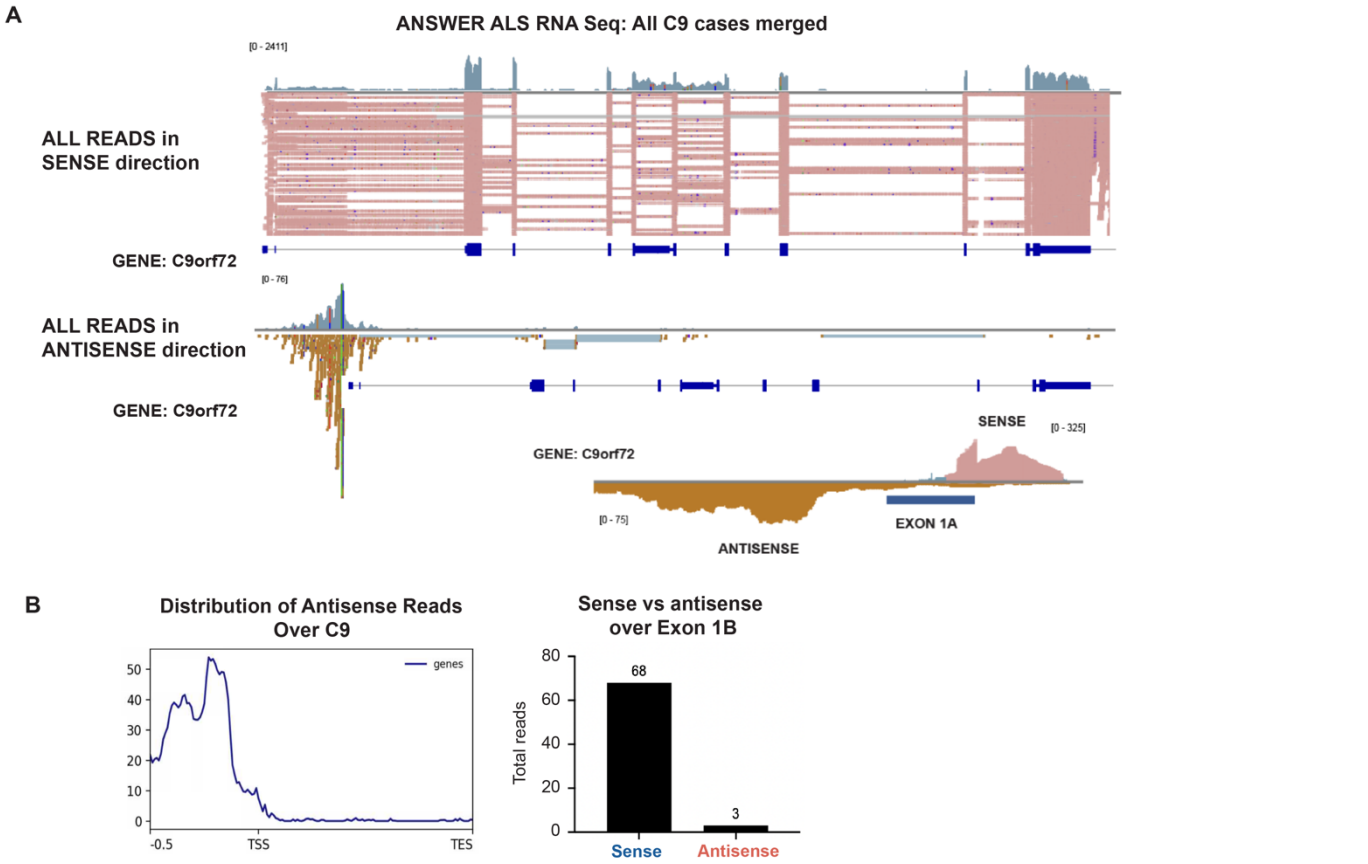


SUPP. FIG 4. ChIP-DIP controls demonstrate technical success of the method and analysis of bidirectional transcription of C9ORF72 gene highlights lack of Exon1B-initiating antisense reads.

S4A. (Top) Schematic of the ChIP-DIP method. (Bottom) ChIP-DIP localization profiles for selected protein targets for the genomic region around C9ORF72 (1) shows unique localization patterns per target and (2) recapitulates known binding patterns for these targets. These results demonstrate the technical success of the ChIP-DIP methodology performed in iPSNs.

S4B. ChIP-DIP localization profiles for H3K27ac and H3K4me1 zoomed in over 5' end of C9ORF72 gene shown alongside Terminator treated RNA sequencing data in the same in-house C9 patient line (12099) as used for ChIP-DIP. Peaks (called using HOMER) are shown (dark blue).

SUPPLEMENTARY FIGURE 5

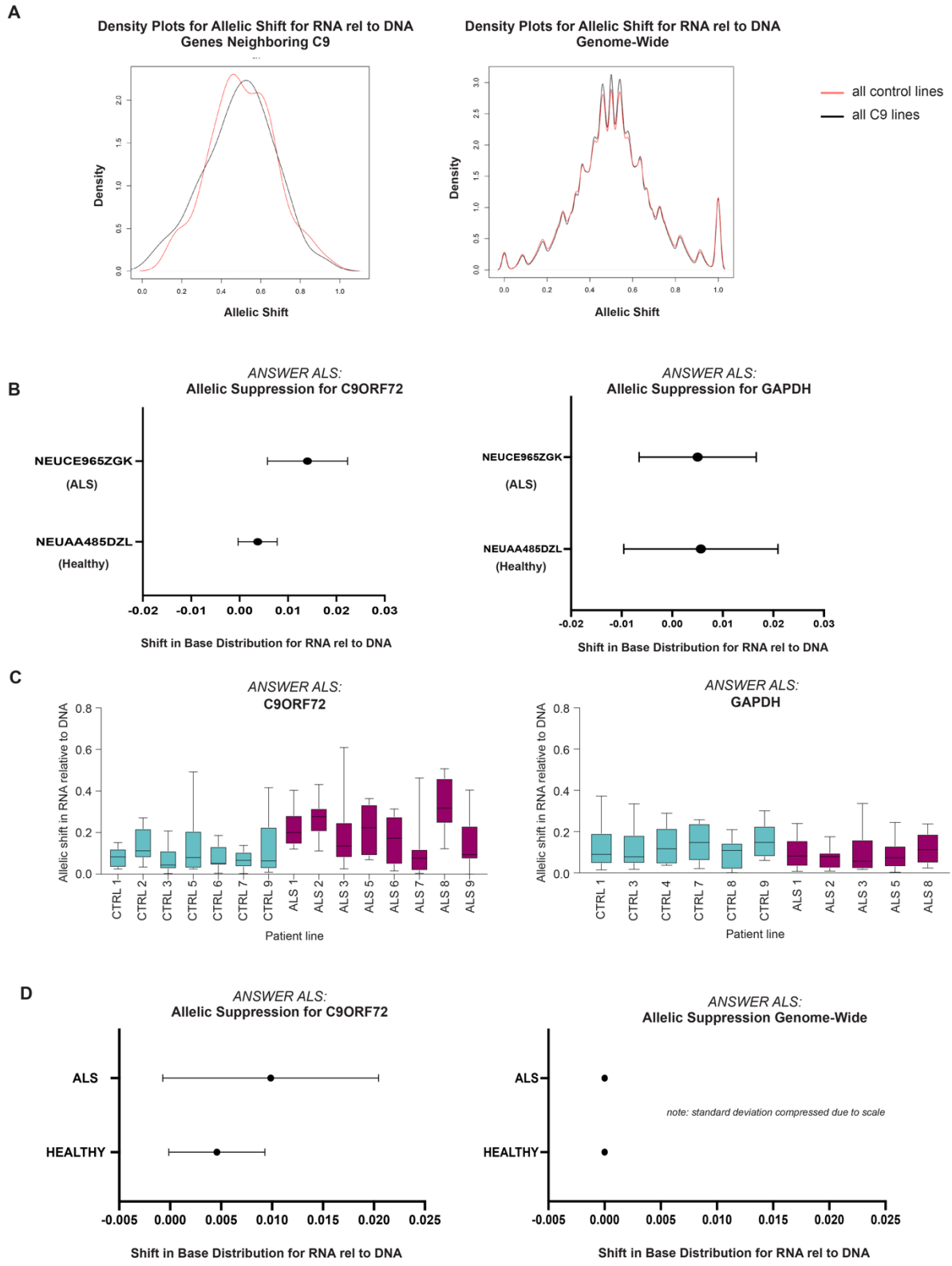


SUPP. FIG 5. Analysis of bidirectional transcription of C9ORF72 gene highlights lack of Exon1B-initiating antisense reads.

S5A. (Top) All reads over C9ORF72 gene for Answer ALS RNA sequencing (all C9-ALS cases merged, >50 repeats of G4C2). Coverage data is shown in blue. Sense reads (below coverage track) are shown in pink, antisense reads (below coverage track) are shown in yellow. (Bottom) Zoom-in to Exon1A shows bidirectional transcription (sense in pink, antisense in yellow), that is specific to the Exon1A region and not any other genomic region of the C9ORF72 gene.

S5B. (Left) Profile plot of antisense transcription over C9ORF72 gene shows antisense transcription specifically localized 5' to the TSS at Exon1A (and not located at Exon1B). (Right) Quantification of sense vs. antisense transcription at Exon1B (as quantified from fragment start bedgraphs, see Fig. 5B) demonstrates no detectable antisense transcription at Exon1B.

SUPPLEMENTARY FIGURE 6



SUPPLEMENTARY FIGURE 6 (cont.)

SUPP. FIG 6. Allelic shift of RNA relative to DNA is specific to C9 gene and not present for specific control genes or all genes genome-wide.

S6A. Density plots (all Answer ALS C9 lines vs healthy lines) showing allelic shift of RNA relative to DNA computed for SNPs over (1) genes neighboring C9 (left), and (2) all genes/ genome-wide (right). Controls (red), C9 cases (black)..

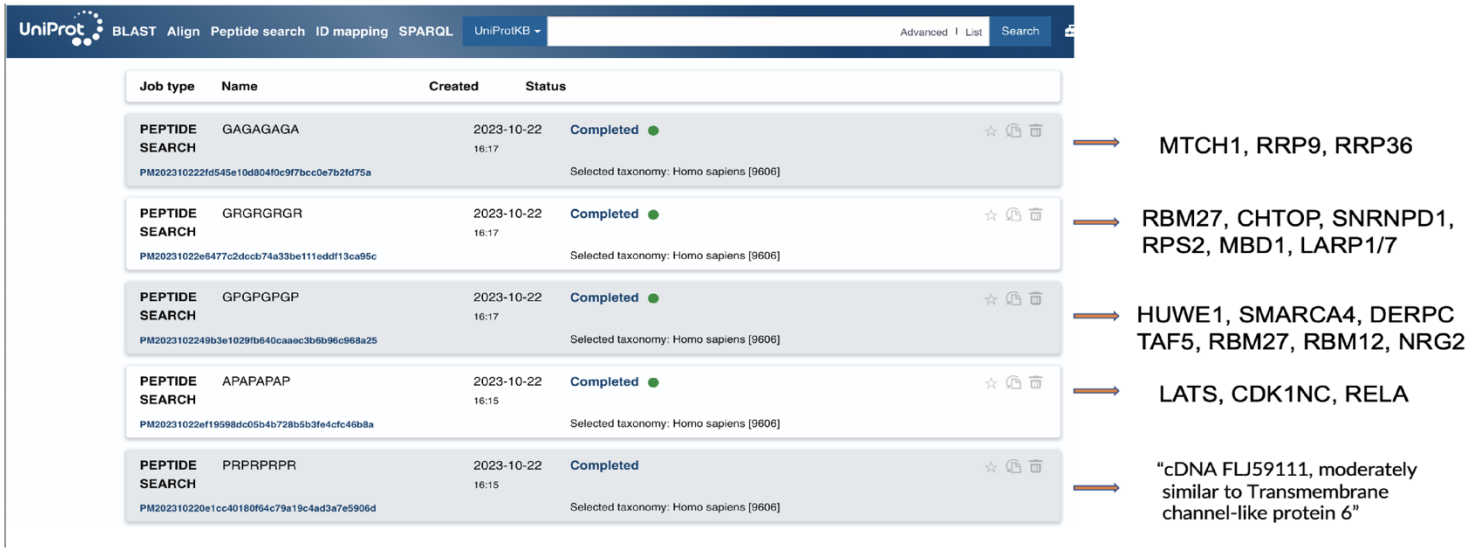
S6B. Shift in allelic distribution for 1 sample C9 lines vs 1 sample healthy line, for SNPs over C9 gene vs. GAPDH. Shift in mean for each is displayed alongside standard deviation. .

S6C.Allelic shift in RNA relative to DNA computed for all SNPs across all 10 C9-ALS lines and 10 healthy control lines are shown across the C9ORF72 gene (left) or GAPDH (right). Filtering criteria (lines with too few SNP positions, see Methods) explains why certain patient lines are not included in plot.

S6D.Shift in allelic distribution for all C9 lines vs healthy lines (pooled for each group), for SNPs over C9 gene vs. GAPDH. Shift in mean for each group displayed alongside standard deviation.

SUPPLEMENTARY FIGURE 7

A



(Note: not all protein matches listed)

SUPP. FIG 7. Sequence matches for DPR amino acid subsequences in other non-DPR proteins argues for critical re-evaluation of specificity of endogenous DPR expression.

S7A. Results of Uniprot peptide search for DPR amino acid strings of length 8 (eg. GRGRGRGR). Search results show matches to many other (non-DPR proteins) with this same amino acid string. Not all search results are shown in figure.

SUPPLEMENTARY TABLE 1

C9-ALS LINES	CONTROL LINES
CASE-NEUCE965ZGK	CTRL-NEUAA485DZL
CASE-NEUDE902GCT	CTRL-NEUCA748GF2
CASE-NEUDZ473EGM	CTRL-NEUFE306EFY
CASE-NEUEM720BUU	CTRL-NEUHZ716BZ2
CASE-NEUFV237VCZ	CTRL-NEUKW131XJ2
CASE-NEUGH995TFK	CTRL-NEULL933JXY
CASE-NEUJX990GR5	CTRL-NEUMA002VLD
CASE-NEUNL415AW	CTRL-NEUPW536ZKZ
CASE-NEUTN952DDG	CTRL-NEURV546WMW
CASE-NEUYY225MNZ	CTRL-NEUXP955XW7

(from Answer ALS)

SUPP. TABLE 1. Table of all C9-ALS and control lines from Answer ALS used for allelic suppression analysis.

SUPPLEMENTARY TABLE 2

Control and C9-ALS Patient iPSN Line Information. All patient lines contained 1000-4000 repeats.

NINDS/ Coriell Code	Sample name	Mutation	Disease	Age of Onset	Age at Sampling	Gender	iPSC Karyotype
ND10689	Line 5280	control	N/A*	N/A	72	F	normal
ND10689	Line 6769	C9ORF72	ALS/FTD	45	46	F	normal
ND10689	Line 10689	C9ORF72	ALS/FTD	49	51	F	normal
ND10689	Line 12099	C9ORF72	ALS/FTD	48	49	M	normal

**NA indicates Not Available in Answer ALS Metadata*

SUPP. TABLE 2. Table of C9-ALS lines used in this study [11] (referred to in the manuscript as in-house patient-derived iPSNs).

2.5 METHODS

MATERIALS AND METHODS

I. Plasmid and Reporter Cloning

The reporter construct expressing the G4C2 repeat stretch was a generous gift from the Trotti/Haeusler Labs at Thomas Jefferson University. This construct was used to generate an entry clone, which was then used in a Gateway cloning reaction with an mCherry-containing destination vector to produce a final expression clone that contained mCherry as a transfection marker. BamHI and XhoI restriction sites were used to generate a "repeats-removed" construct as a negative control. The length of the G4C2 repeat stretch was determined by restriction enzyme digestion of both the repeats-containing construct and the repeats-removed construct as control. The shared backbone of both constructs was sent out for Sanger sequencing by Primordium Labs and Laragen Inc. to verify sequence identity and create reference maps for genome alignments and other downstream analysis.

II. Reporter Transfection

HEK293T cells were cultured in media consisting of 1X DMEM media (Gibco), 1 mM MEM non-essential amino acids (Gibco), 1 mM Sodium Pyruvate (Gibco), 2 mM L-Glutamine (Gibco), 1X FBS (Seradigm). HEK293T cells were transiently transfected with the G4C2 repeats-containing plasmid using BioT transfection reagent (Bioland Scientific). mCherry fluorescence was monitored for 24-48 hours

post-transfection as a mark of transfection efficiency. After 48 hours, cells were scraped from culture dishes, washed 1x with PBS, and pelleted by centrifugation at 1000g at 4C for 5 min. Cell pellets were then flash frozen in liquid nitrogen for storage at -80 C.

III. iPSC Reprogramming

Human lymphocytes from ALS patients and healthy controls were obtained (NINDS Biorepository at the Coriell Institute for Medical Research/Loma Linda University Neurology Clinic) and reprogrammed into iPSCs using episomal plasmids [50-51]. The Adult Dermal Fibroblast Nucleofector Kit and Nucleofector 2b Device (Lonza) were used to introduce mammalian vectors expressing Oct4, Sox2, Klf4, L-Myc, Lin28, and a p53 shRNA into the lymphocytes according to the manufacturer's protocol. Cells were cultured on a MEF feeder layer until the appearance of iPSCs (> 26–30 days). Colonies were selected and expanded in mTESR1 medium on Matrigel. (Genotyping via qPCR was performed as in previous work[11] to quantify expression of episomal plasmid iPSC reprogramming constructs in iPSN lines.)

IV. Generation of iPSC-derived Neurons (iPSNs):

Lipofectamine Stem (Thermofisher) was used to transfect iPSCs with both the Super piggyBac transposase expression vector (SBI) and a dox-inducible h*NGN2*

expression cassette (Addgene 172115). mTESR1 medium was replaced 24 hours after transfection. At 48 hours after transfection, iPSCs were replated at sparse density using Accutase + 10μM ROCK-inhibitor (Ri, Selleckchem). Ri was removed 24 hours after seeding, and 1μg/mL puromycin (Caymen 13884) was added to the media for an initial selection. Puromycin concentrations were increased until a highly pure population of BFP expressing cells was visible, and these purified transgenic iPSCs were used for further experimentation. iPSCs expressing the h*NGN2* piggybac expression cassette were dissociated with Accutase + 10μM Ri to generate iPSN cultures. Then, dissociated iPSCs were seeded into Matrigel coated 6-well plates at ~150,000-200,000 cells as single cells, directly in induction medium (IM) containing DMEM+Glutamax (Thermofisher), NEAA (Gibco 100x), 1% penicillin streptomycin, N2 supplement (Gibco), and 10ng/ml each of BDNF(R&D) and NT-3 (Peprotech), and doxycycline (Caymen, 2μg/mL) to induce the expression of the h*NGN2* transgene. Fresh IM was added to cells 48 hours post-induction and an additional 1μg/mL of puromycin was added if non-transgenic (non-converting) cells were visible. To eliminate dividing cells, 40μM BrDu (Millipore Sigma) was added 72 hours post induction. Pure populations of early iPSNs were observed 5 days after induction, and cell media was then switched to neuronal maintenance medium (MM), containing neurobasal (thermofisher), N2 and B27 supplements (Gibco), 1% penicillin/streptomycin, glutamax (Thermofisher, 100x), and NEAA (Gibco 100x, BDNF and NT-3 (10ng/mL). Cell media was replaced every 72 hours.

Validation of G4C2 Repeat Length. As in previous work (see Supplement of Shi et al.[11]), repeat-primed PCR (RP-PCR) was used to quantify the C9ORF72 intronic repeat length (length of G4C2 repeats) in control and C9-ALS patient iPSC lines. Results were also cross-verified by using Southern blot to size the C9ORF72 repeat region in control and patient iPSC lines (note: only control line #6[11] and patient lines #1-3[11] were used in the current study). The presence of the expanded allele (red arrows) yields a higher molecular weight band (4-14 kb) in all 4 heterozygous patient lines (C9-ALS 1,2,3,8), compared to the wild-type allele which gives a single 2.4-kb band.

V. RNAtag-Seq, NGS Library Preparation, and RNA Sequencing

Our previously published method RNAtag-Seq[53] was used to generate a *single* RNA-seq library with all C9 iPSN and G4C2 repeats-containing transfected HEK-293 RNA samples; these samples were barcoded and pooled *prior to* library construction to minimize variability in downstream steps. Starting with 1 ug of purified total RNA for each sample, RNA was first fragmented at 91C for 2.5 minutes in FastAP Buffer (Thermo Fisher EF0652). Dephosphorylation and end repair were then performed using FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher EF0652) and T4 PNK repair (NEB M0201S, B0201S) with 1 uL of Turbo DNAse and 10X Turbo DNAse Binding Buffer (ThermoFisher AM2238) added at all steps to ensure degradation of genomic DNA. Samples were then

purified using the Zymo Clean and Concentrator kit with 1C columns and RNA was eluted in RNAse-free Ultrapure water. Samples were then barcoded by direct ligation of a unique adaptor to each individual RNA sample, enabling strand-specific sequencing of all samples. Samples were then pooled for cDNA synthesis using Maxima Reverse Transcriptase (ThermoFisher EP0742), 3' adaptor ligation, and amplification with Illumina indexing primers. Finally, the Jumpcode CRISPRclean human ribosomal RNA depletion kit (Jumpcode Genomics KIT1026) was used to deplete DNA derived from ribosomal RNA from the final library.

Sequencing was performed on an Illumina NextSeq 1000, 100 base pair paired-end flowcell. Sequencing reads were trimmed using TrimGalore to remove adaptor sequences. For iPSN samples, reads were aligned to the human hg38 genome using paired-end STAR aligner[73] and SAMtools[74] was used to deduplicate reads such that only uniquely mapped reads were retained for further analysis. Unmapped reads were generated as alignment output and analyzed for the presence of G4C2 repeat expansions. For the repeats-containing reporter construct, Bowtie2 (v2.3.5)[75] was used to build a custom genome. Reads were aligned to this genome using single-end Bowtie2 aligner, and unmapped reads were again generated and analyzed. For the alignment to C9ORF72 cDNA, Bowtie2 (v2.3.5)[75] was again used to build a custom genome that contained the sequences of Exon1A-initiating variants and Exon1B-initiating variant. Read coverage for Exon1A/Exon1B/Exon2/Exon3 were

quantified at their respective junctions to calculate the relative expression level of Exons.

For all RNA sequencing datasets, SAMtools was used to quantify read coverage over genomic elements and regions of the reporter construct. Integrated Genome Viewer (IGV) was used to visualize reads over the genome.

VI. Cell Crosslinking

iPSNs were differentiated to day 10, then crosslinked for ChIP-DIP with 3% formaldehyde (FA) -DSG as follows. Media was removed and cells were washed with 1X room temperature PBS. 10 mM of 2 mM DSG (ThermoFisher 20593) solution was added and cells were rocked gently at room temperature for 45 minutes. DSG solution was removed, cells were washed with 1X room temperature, and freshly made 3% formaldehyde (ThermoFisher 28906) in PBS was added to cells. Cells were then rocked for precisely 10 minutes at room temperature. 200 uL of 2.5M Glycine Stop Solution was added per 1 mL of media in the original culture dish to terminate the formaldehyde crosslinking. Cells were incubated for 5 minutes in stop solution, then incubated with 1X cold PBS for 1-2 minutes. Cells were washed with cold 1X PBS two additional times for 1-2 minutes each. Scraping Buffer (ice cold PBS + 0.5% BSA) was added with the cells at 4C (this step performed on ice). Cells were scraped from plate, transferred to a 15 mL tube, and both pipeted and vortexed to mix. Cells were then pelleted at 1000g at 4C for 5

minutes then resuspended in 1mL cold Scraping Buffer. Cells were centrifuged again at 2000g for 5 minutes at 4C, supernatant was removed, and pellets were flash frozen in liquid nitrogen for storage at -80C.

VII. RNA Isolation

Harvested iPSN cell pellets were lysed in 1 mL RLT + 10 uL BME and total RNA was isolated using the Qiagen RNeasy Mini Kit (Qiagen 74104). The total RNA samples were cleared of genomic DNA contamination using Turbo DNAse and 10X Turbo DNAse Binding Buffer (ThermoFisher AM2238) at 37C for 20 minutes. Samples were purified using the Zymo RNA Clean and Concentrator kit and eluted in Ultrapure RNAse-free water + RNAse inhibitor. Total RNA samples for all iPSN lines were stored at -80C in single-use aliquots until as needed for experimentation.

VIII. Enrichment of C9ORF72 mRNA

Probe Generation. For the Exon-Intron RAP RNA experiments, probes from Integrated DNA Technologies (IDT), Inc. were designed and ordered as 83-mer sequences against exon1-exon8 of the C9 gene including one probe against intron1A at the 5'end upstream of the G4C2 HRE. A biotinylated double-stranded splint adaptor was annealed in preparation for use as a capture handle on streptavidin beads.

For probe generation, 50 pmoles of the pool of RAP probes were treated with PNK enzyme and ATP for 30 minutes at room temperature. In parallel, the biotinylated splint adaptor was bound to Dynabeads MyOne Streptavidin C1 beads for 30 minutes in a 1M NaCl binding buffer. Then, the PNK-treated probes were flowed onto the streptavidin beads bound with biotinylated splint adaptor and ligated in 2X ISMM for 30 minutes at RT. The supernatant was discarded and the beads were washed 3X in PBS + 0.1% Tween. To elute the final biotinylated probe set, the streptavidin beads were heated in NLS Elution Buffer at 95C for 3 minutes, and the eluted probes were removed off bead as quickly as possible. Dynabeads MyOne Silane beads were used to purify the biotinylated probes from the elution buffer. Finally, the correct ligation of the probes to the biotinylated splint adaptor was verified on a 4% gel.

RNA Capture. For the RNA capture for both RAP experiments, 1 ug of purified total RNA from patient iPSNs was prepared in 50 mM HEPES pH 7.4. RNA and probes were individually heat denatured for 5 minutes at 65C. Then, RNA and probes were hybridized in 0.5X RLT (Qiagen 79216) /Detergent Buffer (insert recipe) for 1 hour at 45C. Meanwhile, Dynabeads MyOne Streptavidin C1 beads were washed 3x in 0.5X RLT/Detergent buffer and resuspended in 50 uLs of 0.5X RLT/Detergent buffer. Then, these resuspended streptavidin beads were bound to the hybridized RNA-probe mixture for 30 minutes at 45C. After 30 minutes of binding, beads were washed 3x with 0.5X RLT/Detergent Buffer at 30 seconds

each. For the elution from this first capture, the beads were resuspended in NLS Elution Buffer and heated at 90C for 5 minutes. The supernatant was quickly collected off beads, combined with 1X RLT, and heated at 45C for another hour for the second capture. Streptavidin beads were washed and prepared as for the first capture, then bound to the RNA and probes for an additional 30 minutes at 45C. Captured RNA was again eluted at 90C for 5 minutes. The RNA elution was then treated with Turbo DNAse (in 10X Turbo DNAse Buffer + RNAse inhibitor) at 37C for 15 minutes. The Zymo RNA Clean and Concentrator kit was used to then purify the RNA, with the final elution performed in Ultrapure RNAse-free water. The purified RNA in water was then used in the previously described RNA library preparation protocol, amplified with Illumina primers, and sequenced on an Illumina NextSeq.

IX. C9 Enrichment Analysis

Starting from raw fastq files, reads were first trimmed using TrimGalore to remove adaptor sequences, then aligned to the hg38 genome using paired-end STAR aligner. SAMtools was used to deduplicate reads as well as remove reads that represented excess hybridization probes from the RAP capture (identified as reads in the antisense direction with identical fragment length as the probes). SAMtools was also used to quantify read coverage over genomic regions. Integrated Genome Viewer (IGV) was used to visualize reads over the genome.

X. 5' Cap Enrichment

5 ugs of total iPSN RNA was fragmented to an average size distribution of 300-400 bases using the Ambion RNA Fragmentation Kit (Thermo Fisher, AM8740) at 70C for 2.5 minutes. RNA was phosphorylated using T4 PNK enzyme (NEB M0201S), PNK buffer (NEB B0201S), and 100 mM ATP (NEB P0756L) at 37C for 30 minutes. The sample was cleaned using the Zymo Clean and Concentrator kit with IC columns. The Terminator 5'-3' Exonuclease (Biosearch Technologies) was then incubated with the sample for 1 hour at 30C to digest RNA with 5′-monophosphate ends, preserving RNAs with 5′-triphosphate, 5′-cap. The sample was cleaned again using the Zymo Clean and Concentrator kit, then treated with FastAP Thermosensitive Alkaline Phosphatase (Thermo Fisher EF0652) and T4 PNK repair (NEB M0201S, B0201S), followed by library preparation: (1) RNA adaptor ligation, (2) cDNA synthesis, (3) 3' adaptor ligation, and (4) amplification with Illumina indexing primers. 1.2X SPRI beads (Bulldog Bio CNGS500) were used to clean the PCR product and elution was performed in 12 uL of Ultrapure RNAse-free water.

Sequencing was performed on the Element Aviti System on a 100 base pair paired-end flow cell. Sequencing reads were trimmed using TrimGalore to remove adaptor sequences and reads were aligned to the hg38 genome using paired-end STAR aligner. SAMtools was used to deduplicate reads such that reads that mapped uniquely were retained for further analysis. Starting with the aligned bam files,

second-read bedgraphs were generated to map the transcription start sites of every gene. Bedgraph scores at single-base resolution were summed over regions (eg. C9ORF72 Exon1A, Exon1B) to compute coverage. To identify the TSS of control genes and C9ORF72, Bedgraph scores were computed over the annotated TSS using the Bedtools intersect function and compared to the Bedgraph score over the entire gene. This analysis was performed for the 5'Cap treated RNA relative to input RNA sequencing performed in the same patient samples. The Deeptools[76] suite was used (specifically the computeMatrix and plotProfile functions) to compare genome-wide profile plots for total RNA-sequencing (input/control) compared to Terminator-treated total RNA sequencing (both experiments repeated three times and sequenced to similar depth).

XI. ***CHIP-DIP for Mapping DNA-Associated Proteins***

Lysis and Fragmentation. 1M iPSN cells for patient line #12099 were generated and crosslinked with 3% FA-DSG as previously described, then lysed first with 600 uLs of Lysis Buffer A (50 mM HEPES pH 7.4, 1 mM EDTA pH 8.0, 1 mM EGTA pH 8.0, 140 mM NaCl, 0.25% Triton-X, 0.5% NP-40, 10% Glycerol, 1X PIC) on ice for 10 minutes. Cells were then pelleted for 8 minutes at 850G, and the supernatant was removed. The pellet was then resuspended in 600 uLs of Lysis Buffer B (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 200 mM NaCl, 1X PIC) on ice for 10 minutes. Again, cells were pelleted for 8 minutes at 850G, and the supernatant was removed. The pellet was then resuspended in 550 uLs of

Lysis Buffer C (50 mM HEPES pH 7.4, 1.5 mM EDTA, 1.5 mM EGTA, 100 mM NaCl, 0.1% sodium deoxycholate, 0.5% NLS, 1X PIC). Next, chromatin was fragmented with sonication using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C for a total of 4 min at 4-5 W (pulses of 0.7 s on, followed by 3.3 s off). Size distribution of the fragmented DNA was verified by reverse crosslinking a 20 uL aliquot of sonicated lysate in ProK buffer (20 mM Tris pH 7.5, 100 mM NaCl, 10 mM EDTA, 10 mM EGTA, 0.5% Triton-X, 0.2% SDS) at 80C for 30 minutes. After reverse crosslinking, DNA was cleaned using the Zymo IC DNA Clean and Concentrator kit and quantified on the Tapestation High Sensitivity D1000. DNA fragments were quantified to be 200-1000bp with an average size of ~450 bp.

Preparation of Antibody-Coupled Beads. Antibody ID Oligos were designed and ordered from Integrated DNA Technologies, Inc. (IDT). Each antibody ID oligo contains a 5'phosphate group to enable ligation, a 3'biotin group to allow binding to streptavidin beads, a UMI (blue), a sticky end to ligate to subsequent ODD barcodes (red), and a unique antibody barcode sequence (green). An example Antibody oligo sequence is as follows: /5'Phos/TGACTTGNNNNNNNNTATTATGGTAGATCGGAAGAGCGTCGTG TACACAGAGTC/3Bio/. First, the biotinylated antibody ID oligos were coupled with purified streptavidin (BioLegend, 280302) to make a stock of 909 nM streptavidin conjugated oligo, then diluted 4X to make a final dilution plate of 227

nM. In parallel, Protein G beads were biotinylated as described in the ChIP-DIP protocol with 5 mM EZ-Link Sulfo-NHS-Biotin (Thermo, 21217) at room temperature for 30 minutes.

10 uLs of oligo-coupled Protein G beads were prepared for each antibody in the CHIP DIP experiment. As described in the original CHIP DIP protocol, biotinylated Protein G beads were first washed then aliquoted into a 96-well plate. Then, 14 μL from the from 227nM stock plate of streptavidin-coupled antibody ID oligos was added to each well.The streptavidin coupled antibody ID oligos and biotinylated Protein G beads were bound at room temperature for 30 minutes at 1200RPM. Beads were then washed twice in M2 buffer (20 mM Tris 7.5, 50 mM NaCl, 0.2% Triton X-100, 0.2% Na-Deoxycholate, 0.2% NP-40), and twice in PBST. Then, the number of oligos loaded per bead was quantified as a QC step before proceeding to immunoprecipitation. The “Terminal” tag from the split pool barcoding scheme was ligated onto a 20% fraction of the conjugated beads, and this ligated product was then PCR amplified for 10 cycles, and purified using 1X SPRI beads. The purified product was quantified using Tapestation D1000, and the concentration of the final library and the number of PCR cycles was used to quantify the pre-PCR oligo complexity. The pre-PCR oligo complexity was divided by the number of beads to obtain the bead loading ratio (oligos per bead).

Next, 2.5 ugs of each antibody was added to each well of the 96-well plate of oligo-coupled Protein G beads. The plate was incubated overnight at 4C, then washed twice with 1X PBST. Then, each well of beads was resuspended in 200 µL of 1x PBSt and 4mM biotin and 2.5ug Human IgG Fc and incubated at room temperature for 15 minutes to quench free Protein G or free streptavidin binding sites.

Pooled Immunoprecipitation, Split-Pool Barcoding, and Library Preparation. Next, all wells of oligo labeled and antibody coupled beads were washed 2x with PBST + 2 mM biotin. In preparation for the immunoprecipitation reaction, the volume from all wells were pooled together into one tube. In parallel, the fragmented lysate was diluted with PBSt +10mM biotin + 1x PIC + 2.5ug of human IgG Fc per 10 uL beads. The antibody coupled bead pool was then added to this diluted lysate and incubated on a HulaMixer at room temperature for 1 hour. Beads were washed 2x with IP Wash Buffer 1, 2X with IP Wash Buffer 2, 2X with M2 buffer (see Buffers).

Next, the NEB End Repair Module kit (E6050L; containing T4 DNA Polymerase and T4 PNK) was used to blunt end and phosphorylate the chromatin. The washed beads post-IP were incubated in NEBNext End Repair Enzyme cocktail with NEBNext End Repair Reaction Buffer, supplemented with 4mM biotin and 1 ug human IgG Fc per 10 uLs beads for 15 minutes at 20C. The reaction was quenched with PBS and 100mM EDTA, and beads were washed 2x with PBST. The beads

were next incubated in NEBNext dA-tailing Reaction Buffer and Klenow Fragment (exo-), supplemented with 4mM biotin + 1ug human IgG Fc per 10uL beads for 15 minutes at 37C. The reaction was again quenched with PBS and 100mM EDTA, and beads were washed 2x with PBST.

Split-pool barcoding was performed as described in previous publications with some modifications. First, a DPM barcode is ligated to all DNA molecules to provide a common sticky end for consequent rounds of barcode ligation. Then, beads were split-pool-tagged with 6 rounds with sets of “Odd,” “Even,” and “Terminal” tags. 6 rounds were chosen to ensure that all beads used in the experiment could be resolved, such that almost all barcode clusters (>95%) represented molecules belonging to unique, individual beads. All barcode ligation steps were supplemented with 5.4 uM Protein G and 2mM biotin and were performed at room temperature for 4 minutes. After the final round of barcoding, beads were resuspended in MyRNK buffer and aliquots of various sizes (0.05% to 2% of total beads) were prepared for library preparation. Each aliquot was then incubated with 8ul of Proteinase K (NEB P8107S) at 55C for 2 hours then and reverse crosslinked overnight at 65C. For library preparation, DNA from each reverse crosslinked aliquot was purified using Zymo IC Clean and Concentrator Kit. The cleaned DNA libraries were amplified for 12 cycles using Q5 Hot-Start Master Mix (NEB M0294L) and primers that added Illumina adaptor sequences. 1.2X SPRI beads (Bulldog Bio CNGS500) were used to clean the PCR product and

DNA High Sensitivity D1000 Tapestation was used to quantify library size distribution. Before sequencing, a 2% agarose gel was used to gel purify libraries to remove excess primer.

XII. CHIP DIP Analysis and Peak-Calling

Trimming and Alignment. Trim Galore! V0.6.2 was used to trim adaptor sequences, and trimming quality was then evaluated with FastQC v0.11.8. Cutadapt v3.4 was used to trimmed the RPM sequence from both the 5' and 3' end of reads. The Guttman lab's previously published Barcode ID v1.2.0 (https://github.com/GuttmanLab/sprite2.0-pipeline) was used to identify barcodes and assess ligation efficiency for each round of split-pool barcoding. DNA and oligo tag reads were separated to two output files (using DPM sequence to identify DNA reads) [68]. For DNA reads, Bowtie2 was first used to align to the hg38 human genome, and alignments with a mapq score greater to or equal to 20 were retained for downstream analysis steps. PCR duplicates were removed in the genomic alignment by identifying reads with identical start and stop positions in the genome. Similarly, the unique molecular identifier (UMI) on the oligo reads were used to remove duplicates.

Cluster Generation. Aligned DNA reads were then merged with oligo reads to generate cluster files as previously published (https://github.com/GuttmanLab/sprite2.0-pipeline), incorporating the filtering of barcode

strings that were not in the correct order. From the filtered cluster files, DNA reads were then split into separate files by oligo IDs that corresponded to each protein [68]. As each cluster in ChIP-DIP represents a unique bead, the distribution of oligo tags (each tag represents a unique protein) was measured to determine the metrics by which to split the cluster file. Specifically, a threshold was set that >80% of all tags in a cluster needed to correspond to a protein-oligo tag in order to designate that cluster as belonging to a given protein. Genomic alignment files for each protein were visualized in Integrated Genomics Viewer (IGV).

Bigwig Generation and Peak Calling. Protein-specific bigwig files were generated from the bam (alignment) files for each protein by using the 'bamCoverage' command from Deeptools[76] v3.1.3 with bin size of 10 base pairs. Integrated Genomics Viewer (IGV) was used to visualize the bigwig files genome-wide and over the C9 gene. The 'findPeaks' function from HOMER v3.11106 was used to call peaks on tag directories generated for target alignments files using '-region' with specific settings including '-size' (peak width) and '-minDist'(distance between adjacent peaks). Finally, peaks were called for H3K27ac and H3K4me1 using the parameters "-F 2 -P 0.001" as filtering thresholds.

XIII. Computational Pipeline for G4C2 Subset Reads

A threshold value of >50 repeats of G4C2 was set as the search parameter for patients with the C9ORF72 mutation on the Answer ALS data portal; at the time of

analysis, there was data available from 10 C9ORF72 mutation carriers with >50 repeats of G4C2. For these 10 lines, the relevant genomics (WGS) and transcriptomics (total RNA sequencing) patient datasets were extracted, including raw .fastq files, vcf files, and all alignment files. "G4C2 subset files" were generated by starting with the raw transcriptomics fastq files for each patient and extracting all reads from both Read1 and Read2 that contained the substring "GGGGCCGGGGCCGGGGCC" (G4C2x3). As there are known to be 3-20 G4C2 repeats in the wild-type/healthy case, a conservative threshold of 3 repeats of the G4C2 motif was set in the subset search. Next, a custom C9ORF72 genome was constructed that included the sequences of all 3 known transcript variants of C9 (reference sequences obtained from NCBI Refseq) with 3 repeats of the G4C2 motif added in intron1A at the same position as annotated in the hg38 genome. Then, these G4C2 subset fastq files from Read 1 and Read 2 were used to run single-end STAR alignments against the hg38 genome. Similarly, single-end Bowtie2 (v2.3.5) alignments were run against the custom C9 genome. When running both STAR and Bowtie2 alignments, unmapped reads were saved in a separate folder and analyzed to check for the presence of G4C2 expansion reads, in the possibility that these expansion reads were simply unmappable to the genome and therefore excluded in the alignment. The BLAT function of the UCSC Genome Browser was also used to search the genomic positions of all unmapped reads to check the gene and chromosome they belonged to, as well as to ensure that these reads did not belong to the C9ORF72 gene on chromosome 9. To confirm conclusions from the G4C2

subset analysis, it was noted that the Answer ALS data metadata for these patient lines reports G4C2 repeat lengths ranging from 133-993 as determined by Expansion Hunter (v2.5.5)[77], a tool that estimates repeat size by searching directly through the alignment (CRAM) file. This further confirmed that the G4C2 expansions in these patients are present in DNA sequencing after library preparation with PCR and at the end of the alignment pipeline.

In addition, because it has been reported that the G4C2 HRE may be bidirectionally transcribed, subset fastq files containing the (G2C4)x3 sequence substring were also generated from both the Read1 and Read2 raw fastq files for all patient lines. The same computational pipeline and alignment parameters was used for the (G2C4)x3 reads as for the (G4C2)x3 reads. First, these reads were aligned to both the hg38 genome and our custom C9ORF72 genome, with the alignment parameters set to save and output all unmapped reads. The unmapped reads were checked using the UCSC BLAT search for gene and chromosome, and they were visually inspected to ensure that they were not expansion reads that were entirely comprised of G4C2 repeats. Results for the (G2C4)x3 analysis were very similar to the (G4C2)x3 results.

XIV. ***GTEx Data Analysis***

Tissue expression data for C9ORF72 was obtained from the Genotype-Tissue Expression Portal [56], established by the Broad Institute of MIT and Harvard.

Specifically, RNA sequencing alignments, exon expression data, and splice junction read counts across distinct brain regions were obtained from GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2).

XV. ***Postmortem Data***

Previously published total RNA sequencing performed on C9 postmortem brain tissue was difficult to obtain, as preserving RNA quality and integrity is often challenging in tissue. As a result, there were limited datasets where the total RNA was sequenced to high-depth (~100M reads per sample) while retaining high quality reads with no visible evidence of RNA degradation after alignment to genome. Postmortem data was eventually obtained from "Divergent Single-Cell Transcriptome and Epigenome Alterations in ALS and FTD Patients with C9ORF2 Mutation," a joint effort from the labs of Dennis Dickson, Veronique Belzil, Eran Mukamel, and Stella Dracheva[57]. RNA sequencing datasets and H3K27ac datasets from the motor cortex were analyzed from two cell types also known to be affected in ALS: neurons and glia.

RNA-Sequencing. RNA sequencing data was analyzed from 5 C9 patients for the motor cortex neuron samples and from 3 C9 patients for the glia samples from the motor cortex. For all samples, the raw fastq sequencing files from the bulk RNA sequencing were aligned using paired-end STAR against the hg38 genome, with the same alignment parameters as for all other experiments. SAMtools was used to

deduplicate reads such that only uniquely mapped reads were retained for analysis, and unmapped reads were generated as alignment output and analyzed for the presence of G4C2 repeat expansions. Bam alignment files were analyzed individually for each line, then merged alignment tracks were made for each cell type: one merged track was generated for the motor cortex neuron samples and a separate merged track was generated for the motor cortex glia samples. SAMtools was also used to quantify read coverage over genomic elements and regions of the reporter construct. Integrated Genome Viewer (IGV) was used to visualize reads over the genome. All RNA-sequencing data from postmortem samples was compared to RNA sequencing data from in-house iPSN lines as well as Answer ALS data.

H3K27ac ChIP-Sequencing. H3K27ac ChIP sequencing data was analyzed from neuronal samples from the motor cortex of 2 C9 patients[57]. For both samples, the raw fastq sequencing files from the H3K27ac ChIP sequencing were aligned using paired-end STAR against the hg38 genome, with the same alignment parameters as for all other experiments. SAMtools was used to deduplicate reads such that only uniquely mapped reads were retained for analysis, and Bam alignment files were analyzed individually for each line. Integrated Genome Viewer (IGV) was used to visualize reads over the genome. Bigwigs were generated and peak-calling was performed as described in the ChIP-DIP section of Methods.

XVI. SNP Analysis of Allelic Suppression

To quantify allelic suppression in the C9 context compared to healthy, variant call format (vcf) files were analyzed from Answer ALS datasets, which contain SNP and structural variation calls at individual base positions. Specifically, vcf files were analyzed from the same 10 C9-ALS patient lines (with >50 repeats of G4C2) and 10 healthy control lines that were analyzed in prior sections of this work. Raw data was obtained by extracting the number of reads for each SNP variant at each position annotated in the vcf file for several genomic regions: (1) over the C9ORF72 gene, (2) specific control genes (eg. GAPDH), and (3) genome-wide or across *all* genes. Conceptually, if base 1 and base 2 are two possible SNPs at a given position in the vcf file for a particular patient, a "small allelic shift" would be defined as one where the distribution of base 1 and base 2 shifts minimally, comparing the base1/base2 ratio in the DNA (whole-genome sequencing) compared to that same ratio in the RNA sequencing. Similarly, a "large allelic shift" (consistent with the transcriptional suppression of one allele) is defined by a larger difference in the base1/base 2 ratio in the DNA compared to the RNA.

First, the raw data extracted from vcf files of all 20 patients was filtered to only include SNP positions that matched the following criteria: (1) the ratio of base 1: base 2 in the DNA (cram file) must be >0.4 and <0.6, (2) the total number of reads in the RNA (bam file) must be > 10. After filtering, if a particular patient line had <5 SNP positions total, that line was not included in downstream analysis (eg. Lines

1 and 4 over C9ORF72, Figure 7C). From the filtered data, allelic shifts were computed for the RNA and DNA across all patients for C9ORF72, GAPDH, and genome-wide. The unpaired t-test with Welch's correction was used to quantify statistical significance between C9 patient and healthy groups.

2.6 REFERENCES

1. Zhou, Z.-D., Jankovic, J., Ashizawa, T. & Tan, E.-K. Neurodegenerative diseases associated with non-coding CGG tandem repeat expansions. *Nat Rev Neurol* **18**, 145–157 (2022).

2. Depienne, C. & Mandel, J.-L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *The American Journal of Human Genetics* **108**, 764–785 (2021).

3. Paulson, H. Repeat expansion diseases. *Handb Clin Neurol* **147**, 105–123 (2018).

4. Ellerby, L. M. Repeat Expansion Disorders: Mechanisms and Therapeutics. *Neurotherapeutics* **16**, 924–927 (2019).

5. DeJesus-Hernandez, M. *et al.* Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. *Neuron* 72, 245–256 (2011).

6. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).

7. Xu, Z. *et al.* Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. *Proceedings of the National Academy of Sciences* 110, 7778–7783 (2013).

8. Lee, Y.-B. *et al.* Hexanucleotide Repeats in ALS/FTD Form Length-Dependent RNA Foci, Sequester RNA Binding Proteins, and Are Neurotoxic. *Cell Reports* 5, 1178–1186 (2013).

9. van Blitterswijk, M. *et al.* Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study. *Lancet Neurol* **12**, 978–988 (2013).

10. Majounie, E. *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).

9. Shi, Y. *et al.* Haploinsufficiency leads to neurodegeneration in C9ORF72 ALS/FTD human induced motor neurons. *Nat Med* **24**, 313–325 (2018).

10. Haeusler, A. R., Donnelly, C. J. & Rothstein, J. D. The expanding biology of the C9orf72 nucleotide repeat expansion in neurodegenerative disease. *Nat. Rev. Neurosci.* **17**, 383–395 (2016).

11. Barker, H. V., Niblock, M., Lee, Y.-B., Shaw, C. E. & Gallo, J.-M. RNA Misprocessing in C9orf72-Linked Neurodegeneration. *Front Cell Neurosci* **11**, 195 (2017).

14. Douglas, A. G. L. Non-coding RNA in C9orf72-related amyotrophic lateral sclerosis and frontotemporal dementia: A perfect storm of dysfunction. *Noncoding RNA Res* **3**, 178–187 (2018).

15. Balendra, R. & Isaacs, A. M. C9orf72-mediated ALS and FTD: multiple pathways to disease. *Nat Rev Neurol* **14**, 544–558 (2018).

16. Gendron, T. F. & Petrucelli, L. Disease Mechanisms of C9ORF72 Repeat Expansions. *Cold Spring Harb Perspect Med* **8**, (2018).

17. Frontiers | RNA Misprocessing in C9orf72-Linked Neurodegeneration. https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2017.00195/full.

18. Ghasemi, M. & Brown, R. H. Genetics of Amyotrophic Lateral Sclerosis. *Cold Spring Harb Perspect Med* **8**, a024125 (2018).

19. Zu, T. *et al.* RAN proteins and RNA foci from antisense transcripts in C9ORF72 ALS and frontotemporal dementia. *Proceedings of the National Academy of Sciences* **110**, E4968–E4977 (2013).

20. Wood, H. C9orf72 RNA foci—a therapeutic target for ALS and FTD? *Nat Rev Neurol* **9**, 659–659 (2013).

21. Lagier-Tourenne, C. *et al.* Targeted degradation of sense and antisense C9orf72 RNA foci as therapy for ALS and frontotemporal degeneration. *Proc. Natl. Acad. Sci. U.S.A.* **110**, E4530-4539 (2013).

22. Haeusler, A. R. *et al.* C9orf72 nucleotide repeat structures initiate molecular cascades of disease. *Nature* **507**, 195–200 (2014).

23. Donnelly, C. J. *et al.* RNA toxicity from the ALS/FTD C9ORF72 expansion is mitigated by antisense intervention. *Neuron* **80**, 415–428 (2013).

24. Martier, R. *et al.* Targeting RNA-Mediated Toxicity in C9orf72 ALS and/or FTD by RNAi-Based Gene Therapy. *Mol Ther Nucleic Acids* **16**, 26–37 (2019).

25. Cooper-Knock, J. *et al.* Sequestration of multiple RNA recognition motif-containing proteins by C9orf72 repeat expansions. *Brain* **137**, 2040–2051 (2014).

26. Zhang, K. *et al.* The C9orf72 repeat expansion disrupts nucleocytoplasmic transport. *Nature* **525**, 56–61 (2015).

27. Conlon, E. G. *et al.* The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife* **5**,.

28. Zhang, K., Grima, J. C., Rothstein, J. D. & Lloyd, T. E. Nucleocytoplasmic transport in C9orf72-mediated ALS/FTD. *Nucleus* **7**, 132–137 (2016).

29. Mori, K. *et al.* hnRNP A3 binds to GGGGCC repeats and is a constituent of p62-positive/TDP43-negative inclusions in the hippocampus of patients with C9orf72 mutations. *Acta Neuropathol.* **125**, 413–423 (2013).

30. Coyne, A. N. *et al.* G4C2 Repeat RNA Initiates a POM121-Mediated Reduction in Specific Nucleoporins in C9orf72 ALS/FTD. *Neuron* **107**, 1124-1140.e11 (2020).

31. G4C2 repeat RNA mediates the disassembly of the nuclear pore complex in C9orf72 ALS/FTD | bioRxiv. https://www.biorxiv.org/content/10.1101/2020.02.13.947721v1.

32. Freibaum, B. D. *et al.* GGGGCC repeat expansion in C9orf72 compromises nucleocytoplasmic transport. *Nature* **525**, 129–133 (2015).

33. Xue, Y. C. *et al.* Dysregulation of RNA-Binding Proteins in Amyotrophic Lateral Sclerosis. *Frontiers in Molecular Neuroscience* **13**, (2020).

34. Cooper-Knock, J. *et al.* C9ORF72 GGGGCC Expanded Repeats Produce Splicing Dysregulation which Correlates with Disease Severity in Amyotrophic Lateral Sclerosis. *PLoS ONE* **10**, e0127376 (2015).

35. Wang, X. *et al.* C9orf72 and triplet repeat disorder RNAs: G-quadruplex formation, binding to PRC2 and implications for disease mechanisms. *RNA* **25**, 935–947 (2019).

36. Fay, M. M., Anderson, P. J. & Ivanov, P. ALS/FTD-Associated C9ORF72 Repeat RNA Promotes Phase Transitions In Vitro and in Cells. *Cell Rep* **21**, 3573–3584 (2017).

37. Bampton, A., Gittings, L. M., Fratta, P., Lashley, T. & Gatt, A. The role of hnRNPs in frontotemporal dementia and amyotrophic lateral sclerosis. *Acta Neuropathol* **140**, 599–623 (2020).

38. Freibaum, B. D. & Taylor, J. P. The Role of Dipeptide Repeats in C9ORF72-Related ALS-FTD. *Front Mol Neurosci* **10**, 35 (2017).

39. Nonaka, T. *et al.* C9ORF72 dipeptide repeat poly-GA inclusions promote intracellular aggregation of phosphorylated TDP-43. *Hum. Mol. Genet.* **27**, 2658–2670 (2018).

40. Lee, K.-H. *et al.* C9orf72 dipeptide repeats impair the assembly, dynamics and function of membrane-less organelles. *Cell* **167**, 774-788.e17 (2016).

41. Cook, C. N. *et al.* C9orf72 poly(GR) aggregation induces TDP-43 proteinopathy. *Sci Transl Med* **12**, eabb3774 (2020).

42. Wen, X. *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).

43. The C9orf72 GGGGCC Repeat Is Translated into Aggregating Dipeptide-Repeat Proteins in FTLD/ALS | Science. https://www.science.org/doi/full/10.1126/science.1232927?versioned=true.

44. Mis, M. S. C. *et al.* Development of Therapeutics for C9ORF72 ALS/FTD-Related
Disorders. *Mol. Neurobiol.* **54**, 4466–4476 (2017).

45. Kim, G., Gautier, O., Tassoni-Tsuchida, E., Ma, X. R. & Gitler, A. D. ALS Genetics: Gains, Losses, and Implications for Future Therapies. *Neuron* **108**, 822–842 (2020).

46. Mejzini, R. *et al.* ALS Genetics, Mechanisms, and Therapeutics: Where Are We Now? *Front Neurosci* **13**, 1310 (2019).

47. Sattler, R. *et al.* Roadmap for C9ORF72 in Frontotemporal Dementia and Amyotrophic Lateral Sclerosis: Report on the C9ORF72 FTD/ALS Summit. *Neurol Ther* **12**, 1821–1843 (2023).

48. Cleary, E. M. *et al.* Improved PCR based methods for detecting C9orf72 hexanucleotide repeat expansions. *Mol Cell Probes* **30**, 218–224 (2016).

49. Ebbert, M. T. W. *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).

50. Okita, K. *et al.* A more efficient method to generate integration-free human iPS cells. *Nat Methods* **8**, 409–412 (2011).

51. Pfisterer, U. *et al.* Direct conversion of human fibroblasts to dopaminergic neurons. *Proceedings of the National Academy of Sciences* **108**, 10343–10348 (2011).

52. Answer ALS - Answer ALS. https://www.answerals.org/.

53. Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nat Methods* **12**, 323–325 (2015).

54. Engreitz, J. RNA Antisense Purification (RAP): Experimental Protocols.

55. Smeyers, J., Banchi, E.-G. & Latouche, M. C9ORF72: What It Is, What It Does, and Why It Matters. *Frontiers in Cellular Neuroscience* **15**, (2021).

56. GTEx Portal. https://www.gtexportal.org/home/gene/C9ORF72.

57. Li, J. *et al.* Divergent single cell transcriptome and epigenome alterations in ALS and FTD patients with C9orf72 mutation. *Nat Commun* **14**, 5714 (2023).

58. Jackson, J. L. *et al.* Elevated methylation levels, reduced expression levels, and frequent contractions in a clinical cohort of C9orf72 expansion carriers. *Mol Neurodegeneration* **15**, 7 (2020).

59. Xi, Z. *et al.* Hypermethylation of the CpG island near the G4C2 repeat in ALS with a C9orf72 expansion. *Am. J. Hum. Genet.* **92**, 981–989 (2013).

60. Belzil, V. V. *et al.* Reduced C9orf72 gene expression in c9FTD/ALS is caused by histone trimethylation, an epigenetic event detectable in blood. *Acta Neuropathol* **126**, 895–905 (2013).

61. Xi, Z. *et al.* The C9orf72 repeat expansion itself is methylated in ALS and FTLD patients. *Acta Neuropathol* **129**, 715–727 (2015).

62. On behalf of the BELNEU CONSORTIUM *et al.* The C9orf72 repeat size correlates with onset age of disease, DNA methylation and transcriptional downregulation of the promoter. *Mol Psychiatry* **21**, 1112–1124 (2016).

63. Liu, E. Y. *et al.* C9orf72 hypermethylation protects against repeat expansion-associated pathology in ALS/FTD. *Acta Neuropathol* **128**, 525–541 (2014).

64. Osman, S. Extended and exposed enhancer RNAs activate transcription. *Nat Struct Mol Biol* **29**, 503–503 (2022).

65. Yan, W. *et al.* Dynamic control of enhancer activity drives stage-specific gene expression during flower morphogenesis. *Nat Commun* **10**, 1705 (2019).

66. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**, 21931–21936 (2010).

67. Kristjánsdóttir, K., Dziubek, A., Kang, H. M. & Kwak, H. Population-scale study of eRNA transcription reveals bipartite functional enhancer architecture. *Nat Commun* **11**, 5963 (2020).

68. Perez, A. A., Goronzy, I. N., Blanco, M. R., Guo, J. K. & Guttman, M. ChIP-DIP: A multiplexed method for mapping hundreds of proteins to DNA uncovers diverse regulatory elements controlling gene expression. 2023.12.14.571730 Preprint at https://doi.org/10.1101/2023.12.14.571730 (2023).

69. Colak, D. *et al.* Promoter-Bound Trinucleotide Repeat mRNA Drives Epigenetic Silencing in Fragile X Syndrome. *Science* **343**, 1002–1005 (2014).

70. Park, C.-Y. *et al.* Reversion of FMR1 Methylation and Silencing by Editing the Triplet Repeats in Fragile X iPSC-Derived Neurons. *Cell Reports* **13**, 234–241 (2015).

71. Richter, J. D. & Zhao, X. The molecular biology of FMRP: new insights into fragile X syndrome. *Nat Rev Neurosci* **22**, 209–222 (2021).

72. Salomonsson, S. E. *et al.* Validated assays for the quantification of C9orf72 human pathology. *Sci Rep* **14**, 828 (2024).

73. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

74. Samtools. https://www.htslib.org/.

75. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

76. deepTools: tools for exploring deep sequencing data — deepTools 3.5.6 documentation. https://deeptools.readthedocs.io/en/develop/index.html.

77. Illumina/ExpansionHunter. Illumina (2024).

C h a p t e r 3

ONGOING WORK:

CONVERGENT MECHANISMS OF RNA DYSREGULATION IN ALS

3.1 INTRODUCTION

Our findings that the C9-repeat RNA itself is *not* causal to C9-ALS molecular pathogenesis has now led us to broaden our perspective and methodological approach in investigating RNA-mediated toxicity in ALS. Specifically, we have now begun to consider a broader mechanistic framework for how RNA-protein interactions may be disrupted across the numerous genetic subtypes of ALS, and whether there may be convergent molecular mechanisms disrupted that ultimately lead to neuronal death.

Recent years have seen a rapidly growing interest in understanding the role of RBPs in ALS pathogenesis[1-10]. RBPs are evolutionarily conserved and carry out myriad roles in cellular homeostasis by regulating various stages of RNA metabolism, including RNA stability, localization, splicing, transcription, and translation[11-12]. The central role that RBPs play in cellular survival have led to significant efforts to understand how their dysregulation contributes to disease biology[11-12]. One notable example in neurodegenerative disease is spinal muscular atrophy, where a mutation in the survival motor neuron 1 (SMA1) gene leads to a loss of function of the RBP SMN[13]. With expression levels especially high in the spinal cord, the SMN protein is a component of the SMN complex, which catalyzes small nuclear ribonucleoproteins (snRNPs) that comprise the spliceosome[13]. Therefore, the SMN protein is functionally critical for the effective splicing and processing of all neuronal genes, and it has been mechanistically

proven that its loss of function has a drastically deleterious effect on neuronal survival[12]. While SMA is an example where the loss of function of a single RBP is sufficient to lead to neurodegeneration, there are many other examples of dysregulated RBP function in other neurodegenerative diseases[14].

In a similar vein, ALS is increasingly considered to be a disease of RBP dysfunction and subsequent RNA misprocessing. Some of the strongest evidence in support of this mechanistic framework is that the majority of the most frequently mutated genes in ALS code for RBPs, many of which play wild-type functional roles in the same aberrant molecular processes that are already known to be affected in ALS, including DNA damage response/genome stability, RNA splicing, and nucleocytoplasmic transport (Fig. 1A-C) [15-37]. As a result, the therapeutic relevance of the dysregulation of RBPs has drawn significant recent attention with the development of clinical trials targeting RBPs or related aspects of RNA misprocessing[1,3,6].

However, despite the clear importance of RBP-RNA dysregulation in ALS molecular pathogenesis, the field lacks the fundamental knowledge of the RNA targets of a majority of these RBPs most frequently dysregulated in the disease context. Some of this lack of knowledge can be attributed to the technical challenges in probing RNA-protein interactions[38-41]: specifically, limitations in high-stringency, high-specificity, and high-throughput techniques to profile such

interactions with accuracy *in vivo*. As a result, the development and application of a method that is able to profile RNA-protein interactions on a genome-wide scale would be tremendously beneficial to our understanding of RNA-protein dysregulation in the ALS context. Given the limitations of existing RNA-protein profiling strategies such as eCLIP[38-41], we had previously developed our SPIDR[42-43] methodology to profile the RNA binding sites of dozens to hundreds of RBPs in a single experiment. We then further optimized the existing SPIDR protocol to also allow for the multiplexing of multiple cell lines in a single experiment, to enable the genome-wide profiling of RBP-RNA interactions across multiple ALS patient lines and healthy controls. The application of a multiplexed, genome-wide profiling method like SPIDR would not only enable a comprehensive understanding of known dysregulated RBPs, but would also allow for the discovery of novel RNA-protein interactions in an unbiased manner.

Consequently, we applied our SPIDR methodology in patient-derived fibroblasts to generate multiplexed RNA-RBP maps with the broader objective of uncovering convergent mechanisms of RNA dysregulation across the most common genetic subtypes of ALS: TDP-43, FUS, and C9orf72 (Fig. 2A). In doing so, we have begun to generate the first large-scale, genome-wide reference maps of RNA-protein interactions in ALS patient cells compared to healthy patients with a focus on RBPs involved in mRNA splicing and transport (Fig. 2B).

3.2 RESULTS

3.2.1 TECHNICAL VALIDATION OF SPIDR IN ALS PATIENT CELLS

We first assembled a panel of target RBPs, with a specific focus on mRNA splicing and transport (Fig. 2B). We then purchased antibodies for these RBP targets and performed a screening protocol using on-bead immunoprecipitation (from HEK-293 cell lysate) with a mass spectrometry readout (IP-MS, Fig. 2C). The IP-MS validation ensured that for the subsequent SPIDR experiments we were only using antibodies that effectively enriched for their target epitope relative to a negative control (beads only, no antibody). Using this strategy, we identified a set of "positive hit" antibodies by setting a threshold at 100-fold enrichment relative to the negative control.

Next, we first performed a SPIDR experiment for 2 FUS patient fibroblast lines (AW7 with FUSP525L mutation, AW9 with FUSR495X mutation) and 2 healthy patient control fibroblast lines (AW1, AW5). We performed a modified version of our previously published SPIDR protocol[42], multiplexed to include multiple cell lines in parallel in a single experiment. Specifically, we performed bead barcoding and antibody coupling as previously published, combined the barcoded bead-antibody complexes into a single pool, then split the pooled detection complexes across 4 separate immunoprecipitation reactions (4 cell lines, one IP for each line). We then proceeded through the RNA work-up with 4 separate samples, tagged each sample with a unique set of barcodes in the first round of

split-pool barcoding, then combined all samples for subsequent library preparation and sequencing.

In our sequencing data, we first recapitulated the RNA binding sites of known control proteins to ensure that the SPIDR was effective from a technical standpoint; for instance, we are able to see UPF1 localization at the 3'UTR of most genes, which is consistent with its known localization patterns to RNA and its functional role at a regulator of nonsense-mediated decay (Supp. Fig. 1). In addition, we are able to see localization patterns of known RBPs at new RNA targets, including neuronal genes. For instance, glutaminase (GLS) plays a critical role as a regulator of neuroinflammation and is also involved in the process of glutamatergic excitotoxicity, a phenomenon implicated in ALS neuronal death. In our FUS-healthy SPIDR experiment, we are able to identify specific binding sites of RBPs including MATR3, TAF15, and RBFOX2 at GLS (Fig. 3A). Furthermore, while RBFOX2 binding at the 5'end of NDEL1 is a known RBP-RNA binding interaction, we are also able to identify MATR3 localization to NDEL1 at the 5'end of the gene.

3.2.2 IDENTIFICATION OF NOVEL RNA BINDING PROTEINS

Furthermore, our SPIDR data thus far has also revealed RNA localization for proteins that were not previously known to be RNA-binding. One such example is NUP153, which is a nucleoporin that also plays a critical functional role in the

repair of DNA damage and checkpoint activation in the DNA damage response. We are able to identify specific NUP153 localization at numerous genes, including FUS and SNRNP70. As many of these target RNAs that feature NUP153 localization are involved in splicing, further functional exploration may reveal an interplay between dysregulation of the pathways of splicing, transport, and DNA damage repair, all three of which are phenotypes known to be affected in ALS pathogenesis.

3.2.3 DIFFERENTIAL RNA-RBP BINDING IN ALS VERSUS HEALTHY

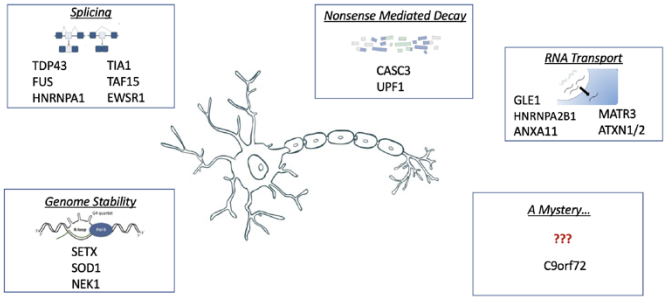
In our low-depth sequencing conducted thus far for the FUS-healthy SPIDR experiment, we are able to see a ~2-fold difference in FUS-U1 binding in the 2 FUS fibroblast lines compared to the 2 healthy lines. In addition to higher depth sequencing for the FUS-healthy SPIDR (currently underway, further analysis pending), we are also in the process of generating similar SPIDR datasets for the TDP43 and C9 fibroblast lines.

3.2.4 SPLICING AND SPLICEOSOMAL DYSREGULATION IN ALS

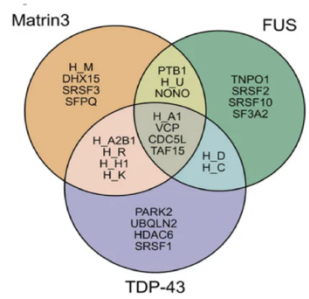
In preliminary results, we have begun to quantify differential binding patterns between FUS and healthy fibroblast lines for an enriched set of RBPs involved in splicing; more detailed investigation of these results is underway.

3.3 MAIN FIGURES

A



B

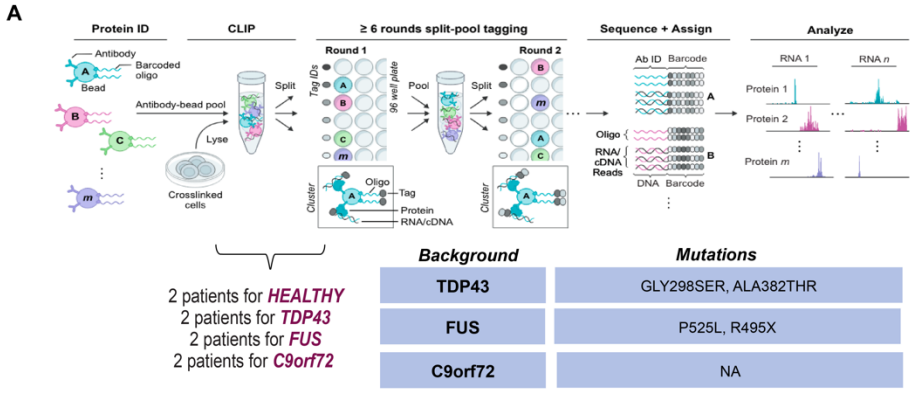


Duennwald et al., Frontiers in Molecular Biosciences, 2022

C

RNA Binding Protein	*Known Functional Role*
TDP43	mRNA splicing, DNA damage repair (R loop repair, NHEJ)
FUS	mRNA splicing (FET family of splicing regulators), mRNA transport, DNA damage repair (BER, R loop repair, NHEJ)
SOD1	DNA damage repair (but unknown if RNA-binding)
MATR3	mRNA transcript stability, mRNA transport, mRNA splicing, DNA damage repair
TIA1	mRNA splicing
TAF15	mRNA splicing (FET family of splicing regulators)
EWSR1	mRNA splicing (FET family of splicing regulators)
SETX	DNA damage repair (R loop repair)
HNRNPA1	mRNA splicing, DNA damage repair (R loop repair, NHEJ)
HNRNPA2B1	mRNA splicing, DNA damage repair (R loop repair, NHEJ)
ATXN2	mRNA splicing, DNA damage repair (R loop repair, NHEJ)

Fig 1: Characterization of G4C2-Expressing HEK System and RAP-MS.
(A) Schematic showing the G4C2 expansion site in the C9ORF72 gene. The mutation in DNA is thought to produce visible RNA foci, but the functional role of the expansion RNA in ALS pathogenesis remains unknown. (B) RNA FISH with (G4C2)x3 probes in HEK-transfected system with (-)Repeat construct and (+)Repeat construct. (C) RNA FISH with Oligo dT probes (yellow) and DAPI staining (blue) reveal a 28% increase in nuclear-retained total RNA in the (+)Repeat



B

RNA Splicing/Processing				RNA Transport		Proteins Linked To ALS	Translation	Genome Stability
ADAR1	FUS	ISY1	SF2/ASF	ALYREF	Nup133	ANXA11	4EBP1	CHD3
AQR	FUBP3	MATR3	SF3A3	XPO1	Nup153	C9orf72	EIF1A/2E	DDX3X
BUD13	HNRNPA1	PTBP1	SF3B1	GLE1	Nup160	DYNC1H1	EIF4G1	DDX6
CELF2	HNRNPA2B1	PTBP2	SF3B4	Nup42	Nup210	G3BP1	ILF3	PARP1
CELF4	HNRNPC	RBFOX1	SFPQ	Nup50	Nup214	FAM120C	PCBP1/2	SETX
CHTOP	HNRNPD	RBFOX2	SMN	Nup54	Nup358	FMRP	RPS2/6	NEK1
Coilin	HNRNPH	RBM15	STAU1	Nup60	TPR	FNDC3CA	LARP1/4	SRPK2
CPEB3	HNRNPK	RBM17	SYF2	Nup88	POM121	GSDMA		
EWSR1	HNRNPL	SAFB2	TAF15	Nup93	Rae1	SOD1		
ELAVL2	HNRNPM	SART3	TIA1	Nup98	Rangap	STXBP1		
ZCCHC11	YTHDC1	YTHDC2	TDP43	NXF1	TREX2	TNRC6C		

Nonsense-Mediated Decay	Vesicular Transport
CASC3	CHMP7
UPF1	RAB1A

C

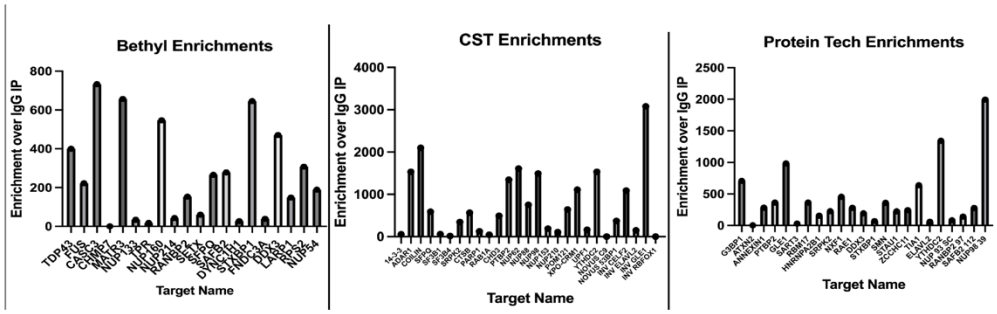


Fig 2: SPIDR Methodology, Experimental Plans, and Antibody Screening.
(A) Schematic showing SPIDR methodology. A modified "multiplexed" version of the SPIDR is performed, which includes a separate immunoprecipitation reaction for each unique patient line. Patient lines are individually tagged with a unique barcode during split pool barcoding, then combined for library preparation and sequencing.Patient lines will be used for FUS, TDP43, and C9ORF72 backgrounds with healthy patient lines as controls. (B) RBP targets for all SPIDR experiments, with a focus on splicing dysregulation and defects in RNA export. (C) Enrichments for antibodies screened for SPIDR, separated by manufacturer (Bethyl, Cell Signaling Technologies, Protein Tech). Antibodies that enriched >100-fold compared to bead-only controls were retained for downstream experiments.

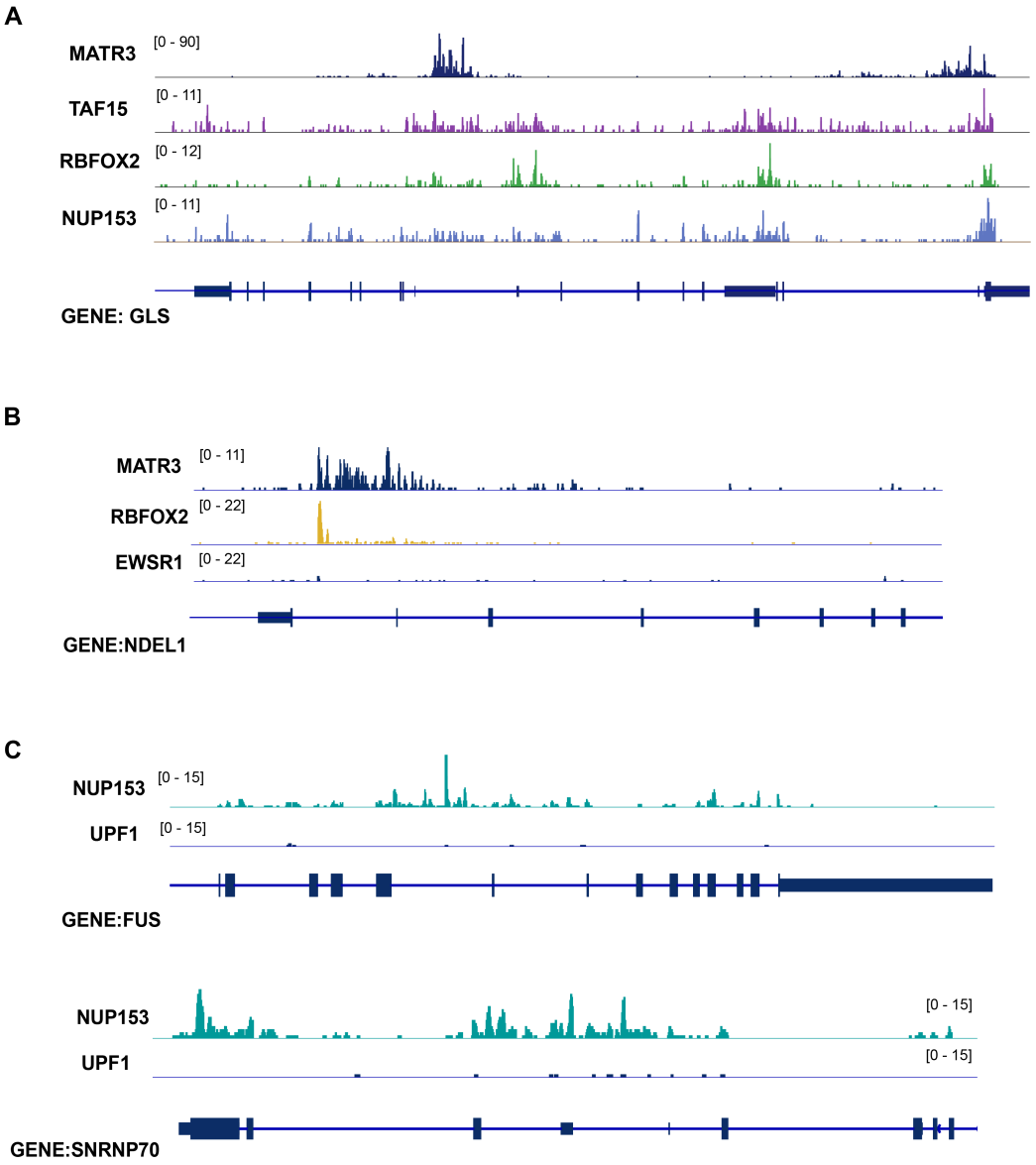


Fig 3: Preliminary results from SPIDR in FUS-ALS patient lines and healthy controls.
(A) RBP localization tracks (MATR3, TAF15, RBFOX2, NUP153) on GLS, a critical neuronal gene. (B) RBFOX2 and MATR3 localizaton on NDEL1. EWSR1 is shown as a control to show specificity of RBFOX2 and MATR3 binding sites. (C) NUP153 binding sites to RNA; NUP153 is not currently known to be a direct RNA binder. NUP153 binding sites shown on FUS and SNRNP70, shown alongisde UPF1 as a negative control.

3.4 METHODS

I. SPIDR for Genome-Wide Mapping of RBPs to RNA[42]

Lysis and Fragmentation. Cells were generated and crosslinked with 4K UV crosslinking as previously described. The HEK-293 SPIDR experiment was performed with 10M cells transfected with the (+)Repeat construct and 10M cells transfected with (-)Repeat construct. The iPSN SPIDR experiment was performed with 10M cells for patient line #6769 and 10M cells for its isogenic control.

For both experiments, each 10M cell pellet was lysed with 1 mL RIPA buffer (50mM HEPES pH 7.4, 100mM NaCl, 1% NP-40, 0.5% Na-Deoxycholate, 0.1% SDS) with 1X final concentration of Protease Inhibitor Cocktail (Sigma-#P8340-5mL), 5 uL of Ribolock RNase Inhibitor (Thermo Fisher, #EO0382)), 10 uL of Turbo DNAse ((Invitrogen, #AM2238), and 1X of Manganese/Calcium mix (2.5 mM MnCl2 , 0.5 mM CaCl2). Cells were incubated for 10 mins on ice in this lysis reaction. Next, cells were sonicated using a Branson needle-tip sonicator (3 mm diameter (1/8'' Doublestep), Branson Ultrasonics 101-148-063) at 4C for a total of 1.5 min at 4-5 W (pulses of 0.7 s on, followed by 3.3 s off). Cells were then incubated at 37C for 10 minutes to allow for DNase digestion, then the reaction was quenched with 0.25 M EDTA/EGTA mix for a final concentration of 10 mM EDTA/ EGTA. To fragment the RNA, a 1:500 dilution of RNAse If ((NEB, #M0243L) was added and lysate was incubated at 37C for 10 mins to fragment RNA to a final size distribution of 300-400 bps. The RNAse reaction

was quenched with 500 uL of ice-cold RIPA buffer supplemented with 5 uLs of Ribolock RNAse Inhibitor and 1X Protease Inhibitor Cocktail, followed by 3 minute incubation on ice for complete quenching. Centrifugation was performed at 15000g at 4C for 2 mins to clear the lysates. Supernatant was transferred to fresh tubes. Size distribution of the fragmented RNA was verified by cleaning the samples using the Zymo IC RNA Clean and Concentrator kit and quantification using the Tapestation High Sensitivity RNA kit. The final fragmented lysate was then stored on ice until the antibody-coupled beads were ready for immunoprecipitation.

Preparation of Antibody-Coupled Beads. The bead labeling strategy is adopted from the Guttman lab methodology ChIP DIP, which enables the multiplexed mapping of proteins to DNA (https://guttmanlab.caltech.edu/technologies/). First, antibody ID oligos were designed and ordered from Integrated DNA Technologies, Inc. (IDT). Each antibody ID oligo contains a 5'phosphate group to enable ligation, a 3'biotin group to allow binding to streptavidin beads, a UMI (blue), a sticky end to ligate to subsequent ODD barcodes (red), and a unique antibody barcode sequence (green). First, the biotinylated antibody ID oligos were coupled with purified streptavidin (BioLegend, 280302) to make a stock of 909 nM streptavidin conjugated oligo, then diluted 4X to make a final dilution plate of 227 nM. In parallel, Protein G beads were biotinylated as described in the

Guttman Lab ChIP-DIP protocol [REF] with 5 mM EZ-Link Sulfo-NHS-Biotin (Thermo, 21217) at room temperature for 30 minutes.

10 uLs of oligo-coupled Protein G beads were prepared for each capture antibody in the SPIDR experiment. As described in the original SPIDR protocol, biotinylated Protein G beads were first washed then aliquoted into a 96-well plate. Then, 14 μL from the from 227nM stock plate of streptavidin-coupled antibody ID oligos was added to each well. The streptavidin coupled antibody ID oligos and biotinylated Protein G beads were bound at room temperature for 30 minutes at 1200 RPM. Beads were then washed twice in M2 buffer (20 mM Tris 7.5, 50 mM NaCl, 0.2% Triton X-100, 0.2% Na-Deoxycholate, 0.2% NP-40), and twice in PBST. The number of oligos loaded per bead was quantified as a QC step before proceeding to immunoprecipitation. The “Terminal” tag from the split pool barcoding scheme was ligated onto a 20% fraction of the conjugated beads, and this ligated product was then PCR amplified for 10 cycles, and purified using 1X SPRI beads. The purified product was quantified using Tapestation D1000, and the concentration of the final library and the number of PCR cycles was used to quantify the pre-PCR oligo complexity. The pre-PCR oligo complexity was divided by the number of beads to obtain the bead loading ratio. Conditions were optimized such that 250-300 oligos/bead were loaded for each well of capture antibody.

Next, 2.5 ugs of each antibody was added to each well of the 96-well plate of oligo-coupled Protein G beads. The plate was incubated at RT for 30 mins, then washed twice with 1X PBST. Then, each well of beads was resuspended in 200 μL of 1x PBSt and 4mM biotin and incubated at room temperature for 15 minutes to quench free Protein G or free streptavidin binding sites.

Pooled Immunoprecipitation. Next, all wells of oligo labeled and antibody coupled beads were washed 2x with PBST + 2 mM biotin. In preparation for the immunoprecipitation reaction, the volume from all wells were pooled together into one tube. In parallel, the fragmented lysate was diluted with RIPA buffer such that the final volume of the reaction corresponded to 1 mL of RIPA buffer for every 100 uL Protein G beads. The antibody coupled bead pool was then split by volume and added to all tubes of diluted lysate. 1M biotin was added to a final concentration of 10mM biotin to quench any remaining free streptavidin-oligo. Each immunoprecipitation reaction (lysate + pooled antibody-oligos) were incubated on a HulaMixer at 4C overnight . The following day, beads were washed 2x with RIPA buffer, 2X with High Salt Wash Buffer (50 mM HEPES pH 7.4, 1 M NaCl, 1% NP-40, 0.5% Na-Deoxycholate, 0.1% SDS), and 2X with CLAP Tween buffer (50 mM HEPES pH 7.4, 0.1% Tween-20).

Tagging of RNA Molecules and Preparation for Barcoding. Beads were incubated at 37C for 10 mins with T4 Polynucleotide Kinase (NEB, #M0201L) to

modify the 3'ends of RNA to have 3'OH groups for subsequent ligation. Beads were washed 2x with High Salt Wash Buffer and 2x with CLAP Tween buffer after end repair of the RNA. Next, RNA was ligated with "RNA Phosphate Modified" (RNA) adaptor (Quinodoz et al., 2021[77]) and High Concentration T4 RNA Ligase I (NEB, M0437M) shaking on a Thermomixer for 1 hour 15 mins at 24C. Beads were then washed 3x with CLAP Tween buffer. RNA was converted to cDNA using Maxima RT 42C for 20 minutes with "RPM Bottom" as an RT primer, enabling ligation during split-pool barcoding by adding a 5'sticky end. Next, Exonuclease I (NEB, #M0293L) was used at 37C for 15 minutes to digest excess primer from the RT reaction.

Split Pool Barcoding Split-pool barcoding was performed as described in previous publications[42-43] with some modifications. Specifically, beads were split-pool-tagged with 6 rounds with sets of "Odd," "Even," and "Terminal" tags. 6 rounds were chosen to ensure that all beads used in the experiment could be resolved, such that almost all barcode clusters (>95%) represented molecules belonging to unique, individual beads. All barcode ligation steps were supplemented with 1:40 Ribolock RNAse Inhibitor and 2mM biotin and were performed at room temperature for 4 minutes. After the final round of barcoding, beads were resuspended in CLAP Tween buffer and aliquots (5% of total beads) were stored for library preparation.

Library Preparation. First, RNA in each aliquot was degraded by incubating with RNase cocktail (Invitrogen, #AM2286) and RNase H (NEB, #M0297L) for 20 minutes at 37C. “Splint ligation” (as described in Quinodoz et al 2021[77]) was used to attach double stranded oligonucleotides to the 3’ends of the resulting cDNA. 1X Instant Sticky End Master Mix (NEB #M0370) was used for the splint ligation reaction for 1 hour at 24°C at 1400 RPM on a ThermoMixer. Elution of the biotinylated oligo and barcoded cDNA were performed by boiling in NLS elution buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 2% N-lauroylsarcosine, 2.5 mM TCEP) for 6 minutes at 91C, at 1350 RPM shaking.

Then, biotinylated oligo was captured by eluting the eluate in 0.5X PBST, 5 mM Tris pH 8.0, 0.5 mM EDTA, 1M NaCl. This reaction was then bound to MyOne Streptavidin C1 Dynabeads (Invitrogen, #65001) for 30 minutes at room temperature. Beads were placed on a magnet and the cDNA-containing supernatant was stored in a fresh tube. 2X Q5 Hot Start Master Mix (NEB #M0494) was used to PCR amplify the biotinylated oligo on-bead, using indexed Illumina adaptor primers. In parallel, the cDNA was incubated with a “anti-RPM,” a biotinylated antisense ssDNA that removes empty insert products by hybridizing between the reverse transcription primer and the splint. After incubation with anti-RPM, the reaction was bound to MyOne Streptavidin C1 Dynabeads (Invitrogen, #65001) for 30 minutes at room temperature. The supernatant was purified using silane beads ((Invitrogen, #37002D) as described

in the manufacturer's protocol. 2X Q5 Hot Start Master Mix (NEB #M0494) was used to PCR amplify the cDNA, again using primers with indexed Illumina adaptor sequences.

For both the oligo and cDNA libraries, 1.2X SPRI beads (Bulldog Bio CNGS500) were used to clean the PCR product and DNA High Sensitivity D1000 Tapestation was used to quantify library size distribution. Before sequencing, a 2% agarose gel was used to gel purify libraries to remove excess primer.

Sequencing. Paired-end sequencing was performed using an Illumina NextSeq 2000 or the Element Biosciences AVITI system using reads lengths of 80x220 (Read1xRead2) nucleotides. Sequencing depth for each aliquot was calculated from the number of unique RNA molecules and number of barcoded beads in each aliquot. On average, cDNA was sequenced to 2X saturation and beads were sequenced to 5X saturation.

II. SPIDR Analysis

Trim Galore! V0.6.2 was used to trim adaptor sequences, and trimming quality was then evaluated with FastQC v0.11.8. Cutadapt v3.4 was used to trimmed the RPM sequence from both the 5' and 3' end of reads. The Guttman lab's previously published Barcode ID v1.2.0[43] (https://

github.com/GuttmanLab/sprite2.0-pipeline) was used to identify barcodes and assess ligation efficiency for each round of split-pool barcoding. RNA and oligo tag reads were separated to two output files (using RPM sequence to identify RNA reads). For RNA reads, Bowtie2 was first used to align to a reference genome and contains the sequences for repetitive RNAs (eg. rRNAs, snRNAs, tRNAs, 45S pre-rRNAs, snoRNAs). For the remaining unaligned reads, STAR aligner was then used to align to the hg38 human genome. PCR duplicates were removed in the genomic alignment by identifying reads with identical start and stop positions in the genome. Similarly, the unique molecular identifier (UMI) on the oligo reads were used to remove duplicates.

Aligned RNA reads were then merged with oligo reads to generate cluster files as previously published (https:// github.com/GuttmanLab/sprite2.0-pipeline) [43], incorporate filtering of barcode strings that were not in the correct order. From the filtered cluster files, RNA reads were then split into separate files by oligo IDs that corresponded to each protein. As each cluster in SPIDR represents a unique bead, the distribution of oligo tags (each tag represents a unique protein) was measured to determine the metrics by which to split the cluster file. Specifically, we set a threshold that >80% of all tags in a cluster needed to correspond to a protein-oligo tag in order to designate that cluster as belonging to a given protein. We then visualized the genomic alignment files for each protein in Integrated Genomics Viewer (IGV).

3.5 REFERENCES

1. Xue, Y. C. *et al.* Dysregulation of RNA-Binding Proteins in Amyotrophic Lateral Sclerosis. *Frontiers in Molecular Neuroscience* **13**, (2020).

2. Frontiers | Matrin3: Disorder and ALS Pathogenesis. https://www.frontiersin.org/journals/molecular-biosciences/articles/10.3389/fmolb.2021.794646/full.

3. Frontiers | RNA Misprocessing in C9orf72-Linked Neurodegeneration. https://www.frontiersin.org/journals/cellular-neuroscience/articles/10.3389/fncel.2017.00195/full.

4. Yamazaki, T. *et al.* FUS-SMN Protein Interactions Link the Motor Neuron Diseases ALS and SMA. *Cell Reports* **2**, 799–806 (2012).

5. Kapeli, K. *et al.* Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. *Nat Commun* **7**, 12143 (2016).

6. Svetoni, F., Frisone, P. & Paronetto, M. P. Role of FET proteins in neurodegenerative disorders. *RNA Biol* **13**, 1089–1102 (2016).

7. Bampton, A., Gittings, L. M., Fratta, P., Lashley, T. & Gatt, A. The role of hnRNPs in frontotemporal dementia and amyotrophic lateral sclerosis. *Acta Neuropathol* **140**, 599–623 (2020).

8. The role of FUS gene variants in neurodegenerative diseases | Nature Reviews Neurology. https://www.nature.com/articles/nrneurol.2014.78.

9. Mackenzie, I. R. A. & Rademakers, R. The role of TDP-43 in amyotrophic lateral sclerosis and frontotemporal dementia. *Curr Opin Neurol* **21**, 693–700 (2008).

10. Mackenzie, I. R. A. & Neumann, M. FET proteins in frontotemporal dementia and amyotrophic lateral sclerosis. *Brain Res* **1462**, 40–43 (2012).

11. Guo, J. K. & Guttman, M. Regulatory non-coding RNAs: everything is possible, but what is important? *Nat Methods* **19**, 1156–1159 (2022).

12. Kelaini, S., Chan, C., Cornelius, V. A. & Margariti, A. RNA-Binding Proteins Hold Key Roles in Function, Dysfunction, and Disease. *Biology* **10**, 366 (2021).

1.

13. Bowerman, M. *et al.* Therapeutic strategies for spinal muscular atrophy: SMN and beyond. *Dis Model Mech* **10**, 943–954 (2017).

14. Conlon, E. G. & Manley, J. L. RNA-binding proteins in neurodegeneration: mechanisms in aggregate. *Genes & Development* **31**, 1509 (2017).

15. Jutzi, D. *et al.* Aberrant interaction of FUS with the U1 snRNA provides a molecular mechanism of FUS induced amyotrophic lateral sclerosis. *Nat Commun* **11**, 6341 (2020)

16. Zhou, Y., Liu, S., Liu, G., Öztürk, A. & Hicks, G. G. ALS-Associated FUS Mutations Result in Compromised FUS Alternative Splicing and Autoregulation. *PLOS Genetics* **9**, e1003895 (2013).

17. Sharma, A. *et al.* ALS-associated mutant FUS induces selective motor neuron degeneration through toxic gain of function. *Nat Commun* **7**, 10465 (2016).

18. Sun, S. *et al.* ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nat Commun* **6**, 6171 (2015).

19. Arnold, E. S. *et al.* ALS-linked TDP-43 mutations produce aberrant RNA splicing and adult-onset motor neuron disease without aggregation or loss of nuclear TDP-43. *Proceedings of the National Academy of Sciences* **110**, E736–E745 (2013).

20. Scekic-Zahirovic, J. *et al.* Cytoplasmic FUS triggers early behavioral alterations linked to cortical neuronal hyperactivity and inhibitory synaptic defects. *Nat Commun* **12**, 3028 (2021).

21. Wang, H., Kodavati, M., Britz, G. W. & Hegde, M. L. DNA Damage and Repair Deficiency in ALS/FTD-Associated Neurodegeneration: From Molecular Mechanisms to Therapeutic Implication. *Frontiers in Molecular Neuroscience* **14**, (2021).

22. Konopka, A. & Atkin, J. D. DNA Damage, Defective DNA Repair, and Neurodegeneration in Amyotrophic Lateral Sclerosis. *Frontiers in Aging Neuroscience* **14**, (2022).

23. Higelin, J. *et al.* FUS Mislocalization and Vulnerability to DNA Damage in ALS Patients Derived hiPSCs and Aging Motoneurons. *Front. Cell. Neurosci.* **10**, (2016).

24. Mori, K. *et al.* hnRNP A3 binds to GGGGCC repeats and is a constituent of p62-positive/TDP43-negative inclusions in the hippocampus of patients with C9orf72 mutations. *Acta Neuropathol.* **125**, 413–423 (2013).

25. Kannan, A., Cuartas, J., Gangwani, P., Branzei, D. & Gangwani, L. Mutation in senataxin alters the mechanism of R-loop resolution in amyotrophic lateral sclerosis 4. *Brain* **145**, 3072–3094 (2022).

26. Kim, H. J. *et al.* Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1 cause multisystem proteinopathy and ALS. *Nature* **495**, 467–473 (2013).

27. Bennett, C. L. *et al.* Senataxin mutations elicit motor neuron degeneration phenotypes and yield TDP-43 mislocalization in ALS4 mice and human patients. *Acta Neuropathol* **136**, 425–443 (2018).

28. Richard, P. *et al.* SETX (senataxin), the helicase mutated in AOA2 and ALS4, functions in autophagy regulation. *Autophagy* **17**, 1889–1906.

29. Tsuiji, H. *et al.* Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. *EMBO Molecular Medicine* **5**, 221–234 (2013).

30. Bright, F., Chan, G., van Hummel, A., Ittner, L. M. & Ke, Y. D. TDP-43 and Inflammation: Implications for Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. *Int J Mol Sci* **22**, 7781 (2021).

31. Brown, A.-L. *et al.* TDP-43 loss and ALS-risk SNPs drive mis-splicing and depletion of UNC13A. *Nature* **603**, 131–137 (2022).

32. Giannini, M. *et al.* TDP-43 mutations link Amyotrophic Lateral Sclerosis with R-loop homeostasis and R loop-mediated DNA damage. *PLOS Genetics* **16**, e1009260 (2020).

33. Izumikawa, K. *et al.* TDP-43 regulates site-specific 2'-O-methylation of U1 and U2 snRNAs via controlling the Cajal body localization of a subset of C/D scaRNAs. *Nucleic Acids Res* **47**, 2487–2505 (2019).

34. Conlon, E. G. *et al.* The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. *eLife* **5**, e17820 (2016).

35. Jo, M. *et al.* The role of TDP-43 propagation in neurodegenerative diseases: integrating insights from clinical and experimental studies. *Exp Mol Med* **52**, 1652–1662 (2020).

36. Tyzack, G. E. *et al.* Widespread FUS mislocalization is a molecular hallmark of amyotrophic lateral sclerosis. *Brain* **142**, 2572–2580 (2019).

37. Wang, Q., Conlon, E. G., Manley, J. L. & Rio, D. C. Widespread intron retention impairs protein homeostasis in C9orf72 ALS brains. *Genome Res.* **30**, 1705–1715 (2020).

1.

38. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP) | Nature Methods. https://www.nature.com/articles/nmeth.3810.

1.

39. Zhang, C. & Darnell, R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotechnol* **29**, 607–614 (2011).

1.

40. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA-protein interactions. *Nat Methods* **16**, 225–234 (2019).

1.

41. Lorenz, D. A. *et al.* Multiplexed transcriptome discovery of RNA-binding protein binding sites by antibody-barcode eCLIP. *Nat Methods* **20**, 65–69 (2023).

42. Wolin, E. *et al.* SPIDR: a highly multiplexed method for mapping RNA-protein interactions uncovers a potential mechanism for selective translational suppression upon cellular stress. *bioRxiv* 2023.06.05.543769 (2023) doi:10.1101/2023.06.05.543769.

43. Quinodoz, S. A. *et al.* SPRITE: a genome-wide method for mapping higher-order 3D interactions in the nucleus using combinatorial split-and-pool barcoding. *Nat Protoc* 17, 36–75 (2022)