UNIVERSITY OF CALIFORNIA SAN DIEGO

Engineering RNA Base-editing Tools and Uncovering Post-transcriptional Regulatory Roles of RNA-binding Proteins in Neurodegeneration

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biomedical Sciences

by

Ryan Jared Marina

Committee in charge:

Professor Eugene Yeo, Chair
Professor Susan Ackerman
Professor Don Cleveland
Professor Jens Lykke-Andersen
Professor Amy Pasquinelli

2022

The Dissertation of Ryan Jared Marina is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2022

DEDICATION

This work is dedicated to my family, who made me all that I am today; to my friends, who gave me the encouragement to persevere; to my teachers, who provided me with curiosity; and to my scientific colleagues, who continue to motivate and inspire me. Thank you all truly.

EPIGRAPH

New knowledge is the most valuable commodity on earth.
The more truth we have to work with, the richer we become.

- Kurt Vonnegut

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTAL FILES

RM_Table_1.xlsx

RM_Table_2.xlsx

RM_Table_3.xlsx

ACKNOWLEDGEMENTS

I would like to personally thank Dr. Gene Yeo for serving as my thesis advisor and scientific mentor throughout graduate school. Your continuous support, optimism, and overall trust in my abilities has been immensely important for my development as an scientist. Additionally, I would also like to acknowledge my thesis committee members for their intellectual contributions and academic mentorship, and career advice.

I would like to all of my friends and colleagues that I was fortunate enough to work with in the Yeo lab. Although there are too many to truly thank individually, I wanted to name a few specifically for your significant contributions in one way or another: Anthony Vu, Florian Krach, Emily Wheeler, Thai Nguyen, Archana Shankar, Kevin Dong, En-Ching Luo, Alex Chaim, Kris Brannan, Brian Yee, Julia Nussbacher, Ron Batra, Fred Tan, Phuong Le, Evan Boyle, Steven Blue, Megan Huang, and Mark Perelis. You (and everyone else) have all collectively helped me through the bulk of my PhD and have given me a tremendous amount of support when I needed it most. You are also some of the most brilliant people I have ever had the pleasure of working with.

I would also like to thank the scientific mentors who sent me on this journey, including my high school biology teacher, Ms. Heather Moran; Dr. Mike King and Dr. Anuj Kumar at the University of Michigan; and Dr. Shalini Oberdoerffer at the NIH for taking me under her wing as a postbac and encouraging me to take on anything and everything. I would also like to thank Dr. Bruce Hamilton and the UCSD Genetics Training Program for academic and career mentorship, as well as 8AM coffee and bagels.

To my friends, both in California and Ohio, who kept me sane with your constant support and words of encouragement. You helped me feel proud of my work and kept me going, and for that I will always be grateful.

Most importantly, I would like to thank my family: my mom Sandra, my dad Richard, and my sister Alexis. You stood by me throughout my entire academic journey and kept me going every step of the way. You taught me all that I know about working hard, taking responsibility,

and bettering myself. I could not be here without you, and love you very much.

Chapter 1, in full, is a reprint of the material as it appears in Cell Reports 2020. Marina RJ, Brannan KW, Dong KD, Yee BA, Yeo GW, Cell Press, 2020. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, is adapted from material as it appears in Nature Methods 2021. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, Yeo GW. Nature Press, 2021. The dissertation author was an investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication. Marina RJ, Becker LA, Nguyen TB, Gitler AD, Yeo GW. The dissertation author was the primary investigator and author of this material.

VITA

2013 Bachelor of Science, University of Michigan

2022 Doctor of Philosophy, University of California, San Diego

PUBLICATIONS

Marina RJ, Becker LA, Nguyen TB, Gitler AD, Yeo GW. "Transcriptome-wide analysis of ataxin-2 depletion confirms efficacy of therapeutic target potential in model of TDP-43 proteinopathy" *In Preparation.*

Morelli KH, Wu Q, Gosztyla ML, Liu H, Zhang C, Chen J, **Marina RJ**, Lee K, Jones KL, Duan W, Yeo GW. "RNA-Targeting CRISPR/Cas13d System Eliminates Disease-Related Phenotypes in Pre-clinical Models of Huntington's Disease", *Nature Neuroscience, In Review*

Brannan KW, Chaim IA, **Marina RJ**, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, Yeo GW. "Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes" *Nature Methods*. 2021 May; 18(5): 507–519.

Marina RJ*, Brannan KW*, Dong KD, Yee BA, Yeo GW. "Evaluation of engineered CRISPR-Cas-mediated systems for site-specific RNA editing" *Cell Reports*. 2020 Nov 3;33(5):108350.

Batra R, Nelles DA, Pirie E, Blue SM, **Marina RJ**, Wang H, Chaim IA, Thomas JD, Zhang N, Nguyen V, Aigner S, Markmiller S, Xia G, Corbett KD, Swanson MS, Yeo GW. "Elimination of toxic microsatellite repeat expansion RNA by RNA-targeting Cas9" *Cell*. 2017 Aug 24;170(5):899-912.

Diao Y, Fang R, Li B, Meng Z, Yu J, Qiu Y, Lin KC, Huang H, Liu T, **Marina RJ**, Jung I, Shen Y, Guan KL, Ren B. "A tiling deletion based genetic screen for cis-regulatory element identification in mammalian cells" *Nature Methods*. 2017 Jun; 14(6): 629–635.

Marina RJ, Oberdoerffer S. "Epigenomics meets splicing through the TETs and CTCF" *Cell Cycle*. 2016 Jun 2;15(11):1397-9.

Marina RJ*, Sturgill D*, Bailly MA*, Thenoz M*, Varma G, Prigge MF, Nanan KK, Shukla S, Haque N, Oberdoerffer S. "TET-catalyzed oxidation of intragenic 5-methylcytosine regulates CTCF-dependent alternative splicing" *The EMBO Journal*. 2016 Feb 1; 35(3): 335–355.

Marina RJ, Fu XD. "Diabetic Insult–Induced Redistribution of MicroRNA in Spatially Organized Mitochondria in Cardiac Muscle" *Circulation Cardiovascular Genetics* 2015 Dec; 8(6): 747–748.

ABSTRACT OF THE DISSERTATION

Engineering RNA Base-editing Tools and Uncovering Post-transcriptional Regulatory Roles of RNA-binding Proteins in Neurodegeneration

by

Ryan Jared Marina

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2022

Professor Eugene Yeo, Chair

RNA-binding proteins (RBPs) play integral roles in mediating all cellular functions through RNA processing. Given the importance of post-transcriptional gene regulation in maintaining cellular function, and the role that RNA processing dysregulation plays in disease causation, it is of critical importance to understand the underlying biology of RNA-RBP interactions. At the same time, engineered RBP paradigms offer tremendous potential to increase our ability to both further probe and manipulate the transcriptome. In this dissertation, we aim to explore both natural and synthetic RBP systems and the role they play on cellular function. We begin by describing molecular tools developed to perform site-directed RNA editing using

CRISPR/Cas technologies and natural adenosine-deaminase enzymes. We demonstrate efficacy of our new platform, as well as perform a systematic evaluation of similar editing platforms in both on- and off-target efficiency. We then describe a novel and simplified sequencing-based platform (STAMP) that demonstrates how RNA editing enzymes can be leveraged to characterize global RNA-RBP interactions both at bulk and single-cell levels. Next, to explore the regulatory roles of a specific RBP in disease modification, we profile the functional landscape of ataxin-2 (ATXN2) in spinal cord harvested from a mouse model of amyotrophic lateral sclerosis (ALS) using RNA sequencing and enhanced CLIP. In doing so, we uncover hundreds of differentially expressed transcripts of moderate fold-change affected by ataxin-2 knockout, the majority of which occur in a dose-dependent manner and can be explained by direct ataxin-2 binding. We also discover a subset of genes whose signatures can be reversed in the context of ALS with ataxin-2 deletion, corresponding to suppression of neuroinflammatory pathways and promotion of neuronal processes, supporting the notion that depletion of ataxin-2 levels may be protective against progression of ALS pathogenesis. Lastly, we describe efforts to further characterize regulatory roles of ataxin-2 in disease susceptibility using newly engineered human stem cell models.

Chapter 1

Evaluation of engineered CRISPR/Cas-mediated systems for site-specific RNA editing

1.1 Abstract

Site-directed RNA editing approaches offer great potential to correct genetic mutations in somatic cells while avoiding permanent off-target genomic edits. Nuclease-dead RNA-targeting CRISPR/Cas systems recruit functional effectors to RNA molecules in a programmable fashion. Here, we demonstrate an *S. pyogenes* Cas9-ADAR2 fusion system that utilizes a 3' modified guide RNA (gRNA) to enable adenosine-to-inosine (A-to-I) editing of specific bases on reporter and endogenously expressed mRNAs. Due to the sufficient nature of the 3' gRNA extension sequence, we observe that Cas9 gRNA spacer sequences are dispensable for directed RNA editing, revealing that Cas9 can act as an RNA aptamer-binding protein. We demonstrate that Cas9-based A-to-I editing is comparable in on-target efficiency and off-target specificity with Cas13 RNA editing versions. This study provides the first systematic benchmarking of RNA-targeting CRISPR/Cas designs for reversible nucleotide-level conversion at the transcriptome level.

1.2 Introduction

Adenosine-to-inosine (A-to-I) RNA editing is an essential process that occurs naturally across approximately 2.5 million sites of the human transcriptome where RNA sequence is chemically altered relative to that of the genome (Tan et al. (2017). In A-to-I editing, the adenosine deaminases acting on RNA (ADAR) family of enzymes catalyze deamination of adenosine to form the base analog inosine (I), which is recognized as guanosine (G) by the splicing and translational machinery (Nishikura, 2010). ADAR proteins are a highly conserved family of proteins, containing a single deaminase domain (DD) as well as one or more double-stranded RNA binding domains (Phelps et al., 2015). The flexible nature of ADAR proteins, as well as the modular properties of their deaminase domains, has made targeted RNA editing possible through either recruitment of endogenously expressed ADAR proteins to specific sites (Merkle et al., 2019, Qu et al., 2019) or fusion of these deaminase domains to programmable protein modules (Montiel-Gonzalez et al., 2019). Both remain desirable strategies for achieving site-directed RNA editing *in vivo*.

The bacterial adaptive immune CRISPR/Cas9 system from *Streptococcus pyogenes* (SpCas9) has proven to be a powerful tool for manipulating eukaryotic genomes (Adli, 2018, Hsu et al., 2014). Most CRISPR–Cas9 genome editing applications utilize single guide RNAs (gRNAs) in complex with Cas9 nucleases to induce double-stranded DNA breaks at genomic target loci specified by a gRNA spacer sequence (Jinek et al., 2012, Mali et al., 2013). Such breaks can either be used to destroy gene function through disruption of coding sequences, or facilitate specific gene edits through donor template-driven homology-directed repair (Ran et al., 2013). These methods of genomic manipulation are irreversible and potentially damaging to cells (Kosicki et al., 2018), and remain largely impractical for post-mitotic cell types that lack reliable DNA break-repair mechanisms (Cox et al., 2015, Ran et al., 2013). Therefore, it has been of great interest to develop universally programmable methods to reversibly alter genetic information by targeting RNA instead.

Programmable targeting of cellular RNA using RNA-targeting nuclease-dead SpCas9 (RCas9) was demonstrated in mammalian cells (Batra et al., 2017, Nelles et al., 2016) and *in vivo* to reverse disease-relevant phenotypes in a mouse model of myotonic dystrophy (Batra et al., 2020). This technology utilizes the capacity of Cas9 to bind specific RNAs when complexed with gRNAs that target nucleotide sequences that are absent of protospacer adjacent motifs (PAM). Subsequent studies by other groups have demonstrated RNA-targeting capabilities for other Cas9 systems including CjCas9 (Dugar et al., 2018), NmCas9 (Rousseau et al., 2018), and SaCas9 (Strutt et al., 2018). All of these RNA-targeting Cas9 systems maintain RNA-binding capacity when DNA nuclease activity is inactivated, opening up the exciting possibility that RCas9 fusions to enzymatic RNA editing modules would allow programmable RNA editing at the single base level, analogous to Cas9-deaminase fusions acting on target DNA (Gaudelli et al., 2017, Kim et al., 2017b). Recent studies have demonstrated the natural and exclusive RNA-targeting capabilities of the Cas13 family of proteins including Cas13a, Cas13b, and Cas13d in vitro and in cells (Abudayyeh et al., 2017, Cox et al., 2017, Konermann et al., 2018). Similar to the RCas9 system, nuclease-dead versions of these Cas13 systems have been repurposed for effector module recruitment, including fluorescent proteins for dynamic RNA imaging (Yang et al., 2019), splicing factors for regulation of alternative splicing and expression of specific protein isoforms (Konermann et al., 2018), and ADAR RNA deaminase domains to dCas13b to direct A-to-I, as well as C-to-U, RNA editing (Cox et al., 2017, Abudayyeh et al., 2019).

CRISPR/Cas systems offer promise in the area of transcriptional editing, as the elements of these systems are genetically encodable and amenable to therapeutic viral delivery strategies (Ran et al., 2015, Konermann et al., 2018, Kim et al., 2017a). To date dCas13b represents the only demonstrated example of an RNA-targeting CRISPR/Cas system capable of site-directed RNA base conversion, although the relative efficiency and specificity of this RNA editing approach compared to other CRISPR/Cas systems have not been systematically characterized (Vogel et al., 2018, Vogel and Stafforst, 2019). Moreover, it is still unclear if these or other potential Cas-directed RNA editing systems perform in a fully Cas-dependent manner (Qu et al., 2019). Here

we expand the Cas-mediated RNA editing toolbox by introducing an RCas9-ADAR2 platform capable of editing both reporter and endogenously expressed cellular transcripts in live cells. In characterizing this system, we also discovered a unique Cas9-dependent RNA-aptamer binding mechanism that is independent of a gRNA spacer sequence previously thought to be necessary for both DNA and RNA binding. To assess relative editing efficiencies and specificities of these Cas-based platforms, we assembled and tested several combinations of nuclease-dead Cas9 and Cas13-ADARDD fusions for a side-by-side comparison of RNA editing potential. Finally, we performed RNA sequencing for a subset of these platforms with respect to a shared mRNA target to assess transcriptome-wide off-target biases for each of these systems. Our results demonstrate the comparable efficacy of an RCas9-mediated RNA editing platform and define the parameters and limitations of currently available CRISPR/Cas-based RNA editing tools.

1.3 Results

1.3.1 RNA-targeting Cas9 fused to ADAR2DD supports A-to-I editing in live cells

To evaluate if Cas9 can mediate RNA editing, we fused the deaminase domain of human ADAR2 (ADAR2DD) to the N-terminus of catalytically inactive *S. pyogenes Cas9* (dCas9), separated by an XTEN peptide linker (Figure 1.1A). To achieve site-specific A-to-I editing, we modified the gRNA to contain an additional 3' terminal sequence designed to mimic the double-stranded (dsRNA) substrate for ADAR2DD when base-paired to the target RNA, presenting a mismatched bulged cytidine base opposite of the targeted adenosine (Figure 1.1A, B). A-C mispairings have been shown to facilitate ADAR editing in previously described Cas13 or aptamer-based systems (Wong et al., 2001, Cox et al., 2017, Vogel et al., 2018). These components together comprise an RCas9-ADAR2DD system that can be delivered to mammalian cells. To determine levels of Cas9-independent background editing, we also created a matched human ADAR2DD-only control (Figure 1.1A).

To assess the efficacy of this RNA editing system, we generated stable Flp-In T-REx 293 cell lines containing single genomic copies of either an RCas9-ADAR2DD fusion or ADAR2DD-only control under an inducible Tet-On promoter. We then evaluated editing efficiency of these lines utilizing a previously established (Hanswillemenke et al., 2015) gain-of-function EGFP reporter containing a premature amber termination codon (UAG, W58X) which we cloned downstream of an mCherry cassette separated by a self-cleaving P2A peptide sequence (Figure 1.1A). Upon successful editing (UAG to UGG), EGFP signal is restored and detectable in cells (Figure 1.6A) (Hanswillemenke et al., 2015). Co-expression of RCas9-ADAR2DD with a targeting gRNA containing a complementary 3' extension sequence led to successful EGFP editing at both mRNA and protein levels (Figure 1.1C), whereas a gRNA with a reverse complementary, non-targeting (NT) extension sequence or ADAR2DD-alone produced no detectable editing (Figure 1.6B, C). At the protein level, we observed comparable expression between RCas9-ADAR2DD and its hyperactive E488Q variant (Figure 1.6D). ADAR2DD-only controls exhibited protein expression (likely due to their shorter lengths) higher than Cas9-fusion proteins, and yet the editing efficiencies by them as measured by percentage of EGFP cells were nonexistent (Figure 1.1D), further supporting that our on-target editing was RCas9-driven. Together, these results demonstrate that this directed editing system is dependent on both the specificity of the guide 3' extension sequence and the Cas9 module. Surprisingly, an alternately designed gRNA with a 5' extension sequence did not elicit this context-specific response, as the ADAR2DD control condition also resulted in editing with the gRNA (Figure 1.6E). Attempts to design a gRNA with a spacer sequence that could also serve as an editing substrate failed to achieve successful EGFP restoration as well (Figure 1.6F, G). We also evaluated a hyperactive E488Q ADAR2DD variant, which increased the editing level in the W58X EGFP reporter restoration at both RNA and protein levels, as expected (Montiel-Gonzalez et al., 2016, Cox et al., 2017) (Figure 1.1C). The effects of this hyperactive mutant led to a 2.5-fold increase in EGFP signal as measured by fluorescence-activated cell sorting (FACS) analysis (Figure 1.1D). Expression of only the ADAR2DD(E488Q) domain alone, in the presence of either targeting

or non-targeting gRNAs, was not successful in restoring EGFP signal (Figure 1.1C, D), further highlighting the specificity of this interaction. Previous work showed that RNA-targeting using SpCas9 was augmented using a synthetic antisense oligonucleotide called a PAMmer, which carries a short DNA motif (the protospacer adjacent motif or PAM of the form 5'-NGG-3') (Nelles et al., 2016, Strutt et al., 2018). Similar to Batra et al. (Batra et al., 2017), we observed that the PAMmer is dispensable for efficient on-target editing of the reporter RNA (Figure 1.1E).

1.3.2 RCas9 gRNA spacer sequence is dispensable for RNA-targeting

For initial design, we arbitrarily chose a spacer sequence targeting a region approximately 50nt away from the targeted adenosine residue. To evaluate the optimal distance between where the spacer sequence recruits Cas9 on the target RNA and where the extension sequence directs ADAR2DD for editing, we tiled the mRNA transcript using spacer sequences targeting sites up to 80nt away from the extension sequence (Figure 1.2A). Surprisingly, the nucleotide distance between these gRNA sections did not show a detectable influence on target reporter editing efficiency, and although certain sites showed some target biases ¬(e.g. 40nt, 60nt), modulation of distance within a window of 5 to 80 bases between spacer and extension did not affect EGFP restoration (Figure 1.2B). Furthermore, removal of the spacer sequence from the gRNA (Figure 1.7A) conferred editing efficiency comparable to that of spacer-containing guides (Figure 1.2B). This minimal 'no spacer' (NS) gRNA was unaffected by the presence of a PAMmer sequence (Figure 1.7B) and failed to elicit background editing in the presence of the deaminating domain alone (Figure 1.7C). This result suggests that in the absence of a gRNA spacer sequence, Cas9's interaction with the scaffold of the gRNA molecule is equivalent to an aptamer-binding effector fusion. Seeing that both target specificity and editing efficiency are driven by the 3' RNA extension motif while still behaving in a Cas9-dependent manner, this led us to opine an 'aptamer-binding model' in which RNA-RNA hybridization mediates RNA-protein interactions through dCas9-scaffold assembly (Figure 1.2C).

We next assessed the capacity for RCas9-ADAR2DD to modify endogenously expressed

transcripts in cells. For our initial analysis, we chose to focus on a target site within the 3'UTR of *ACTB* as well as a coding variant within the *CYFIP2* gene, a known ADAR2 target transcript normally expressed but not edited in HEK293 cells. As we observed with reporter mRNA, we were able to achieve successful on-target editing within cellular targets *ACTB* and *CYFIP2* using both spacer-containing and spacer lacking gRNAs in 293XT cells through transient transfection (Figure 1.2D, E), which we quantified using next-generation sequencing (NGS). Editing for both of these targets were detected when both Cas9 and targeting gRNA were present, but gRNA alone or a non-targeting gRNA (NS-NT gRNA) failed to elicit detectible A-to-I editing. Moreover, ADAR2DD(E488Q) variants improve on-target editing signal by 2-3 fold over wildtype without dramatically increasing editing for adjacent adenosine residues. In this instance, NS gRNA performed as well if not slightly better than spacer-containing gRNAs for *ACTB* and *CYFIP2* target sites. Though these editing rates were reported from transiently transfected cell lines, we were also able to verify successful editing in our stable Flp-In 293 cell lines (Figure 1.7D, E). This targeted editing is particularly noteworthy in the case of *CYFIP2*, as this process is normally mediated through exon-intron complementarity to allow the natural dsRNA substrate formation (Licht et al., 2016). We expanded this targeting mechanism to other target genes including *GAPDH* and *GUSB*, which are not naturally edited by endogenous ADAR proteins (Figure 1.7F, G). We conclude that RCas9-ADAR2DD is capable of directing specific A-to-I edits on both reporter and cellular transcripts using a 3' modified gRNA, highlighting a unique spacer-independent mechanism that can be used for RNA-targeting applications.

1.3.3 Comparison of RNA-targeting CRISPR platforms reveals parameters influencing on-target specificity

Studies have discovered other RNA-targeting CRISPR/Cas systems, some of which have been adapted as orthogonal A-to-I RNA editing platforms (Cox et al., 2017). However, these CRISPR-based platforms have not yet been compared to determine criteria that would make for an ideal RNA base editor. To investigate the site-directed RNA editing potentials for these systems,

we cloned a representative panel of Cas-ADAR2DD(E488Q) fusions, in both ADAR2DD N-terminal (N-TERM ADAR2DD) and C-terminal (C-TERM ADAR2DD) orientations in the presence of either a nuclear localization (NLS) or export signal NES), and co-transfected them into HEK293XT cells with respective W58X EGFP-targeting gRNA reporter constructs (Figure 1.3A, B). This diverse panel included SpCas9, S. aureus Cas9 (SaCas9) (Strutt et al., 2018), *Leptotrichia wadei* Cas13a (LwaCas13a) (Abudayyeh et al., 2017), *Prevotella sp. P5-125* Cas13b (PspCas13b) (Cox et al., 2017), and *Ruminococcus flavefaciens XPD3002 Cas13d* (RfxCas13d) (Konermann et al., 2018). For gRNA designs, Cas9 systems were outfitted with NS gRNAs with targeting or non-targeting 3' extensions, while guides for Cas13 fusions were designed with the "bulge-containing" C-mismatch within the spacer sequence that determines the RNA targeting site. We decided to use these orientations for Cas13 modules (spacer-with-mismatch) based upon gRNA designs that were deemed to be optimal by the REPAIR and RESCUE studies (Cox et al., 2017, Abudayyeh et al., 2019) (Fig 3B).

Except for LwaCas13a, we found that editing efficiency as determined by EGFP correction was comparable across all nuclear-localized systems tested in the N-TERM ADAR2DD-orientation (N-terminal ADAR2DD; Figure 1.3C). Importantly we found that C-TERM ADAR2 DD fusion orientations (C-terminal ADAR2DD), although resulting in the highest editing rate, also yielded high background reporter editing with non-targeting gRNAs for all Cas9 and Cas13 fusions, sometimes even exceeding targeting gRNA editing, leading us to conclude that certain orientations may be subject to non-specific background editing, at least when domains are connected through an XTEN linker. Of the N-TERM ADAR2DD fusions which reported $>30\%$ editing, the background-subtracted editing rate relative to non-targeting gRNA controls was comparable between dSpCas9 and dSaCas9, while dPspCas13b seemed to have the highest signal for N-terminal ADAR2DD fusions among Cas13 proteins (Figure 1.8A). Addition of an NES to all fusions yielded similar results, in some cases raising editing rates by a few percentage points (SaCas9, which had the highest overall signal) while others experienced no change or a minor decrease in editing efficiency (SpCas9 and RfxCas13d, Figure 1.3D, Figure 1.8B). Seeing

that editing rates were not appreciably different between nuclear and cytoplasmic localization, choosing between the two options will likely depend on the target and context, bearing in mind recent reports that off-target editing might be more prominent with cytoplasmic localization (Vallecillo-Viejo et al., 2018). Additionally, single-stranded RNA molecules of a certain length may be sufficient to trigger endogenous ADAR activity to result in editing though RNA-RNA hybridization kinetics alone (Qu et al., 2019). While dCas13 and dCas9-driven fusions had comparable editing efficiency with respect to the EGFP reporter, expression of ADAR2DD(E488Q) alone was able to confer noticeable editing signal for both Cas13b and Cas13d gRNAs (Figure 1.8C). We believe that this observation might be related to length of RNA-RNA hybridization region, as gRNAs with target hybridization lengths of 30 nucleotides or longer had increased background editing potential (Figure 1.8D, E) or local off-target edits (Figure 1.8F, H). Regardless of mechanism, this result is consistent with Vogel et. al who showed that 50nt-REPAIR RNA editing is in part Cas-independent and conferred by expression of guide/ADAR2DD combination alone (Vogel et al., 2018).

We next tested the performance of these Cas-based systems on endogenously expressed transcripts in live cells. To assess editing efficiency at a per-read level, we performed targeted-amplicon-specific NGS on target sites of endogenously expressed transcripts *ACTB* and *CYFIP2* following co-transfection of N-terminal ADAR2-Cas fusions and respective gRNAs into 293XT cells. For this panel, we focused on characterizing SpCas9 (simply Cas9 going forward) alongside the two most successful Cas13 modules in our hands: Cas13b and Cas13d. Similar to what we had seen through EGFP restoration, SpCas9 in both spacer-dependent and independent contexts performed targeted A-to-I editing at a similar efficiency to both Cas13b and Cas13d (Figure 1.3E, F). Additionally, sequence edits called were seemingly confined to the targeted adenosine residues, even with the introduction of the hyperactive E488Q mutation to the SpCas9-ADAR2DD fusion, though in the case of *CYFIP2* there appeared to be instances of adjacent base editing in all targeting gRNA contexts (Figure 1.3F). This off-target editing may highlight a recurrent tradeoff between base specificity for editing efficiency that should be considered when

employing RNA editing technologies.

1.3.4 Transcriptome-wide RNA-Seq uncovers consequences of RNA-targeting CRISPR editing platforms

Next, RNA-seq was used to systematically assess off-target consequences from expression of Cas-ADAR2(E488Q)DD systems in HEK293XT cells. Cas9, Cas13b, and Cas13d fusions in the N-TERM ADAR2DD-NLS orientation were introduced with either a *CYFIP2*-targeting or scrambled NT gRNA control. After 48 hours, RNA-seq libraries were prepared and sequenced to 42 million reads on average. Aligned reads were downsampled prior to variant calling to allow for comparisons unaffected by variable sequencing depth. We processed aligned reads using the SAILOR algorithm (Deffit et al., 2017) specifically designed to identify high confidence A-to-I editing events from transcriptome-wide sequencing data (Figure 1.4A). SAILOR assigns a confidence score for A-to-G mismatches that differ from the genome using a beta distribution factoring in both site coverage and editing percent following removal of annotated hg19 SNPs which may be falsely identified as hits (Washburn et al., 2014). For our analyses, only events with a confidence score exceeding 90% detected in multiple replicates were considered as consensus off-target events. This pipeline identified several thousand A-to-I editing events with transient expression of Cas-ADAR2(E488Q)DD editing systems, ranging from as few as 3,336 with Cas13b with a *CYFIP2*-targeting gRNA to as many as 5,545 events with Cas9 co-expressed with a no-spacer (NS) *CYFIP2*-targeting gRNA (Figure 1.4B, Supplementary Table 3). Though the number and percentage of editing events seemed to vary between Cas conditions, all unsurprisingly were elevated over an untransfected control condition (900 A-to-I edits). We also confirmed that the mRNA expression of each of the respective fusion proteins were similar by the Transcripts Per Million (TPM) metric (Figure 1.9A). Although the distribution of editing percent was similar between all conditions, Cas-ADAR2(E488Q)DD fusions introduced more "subtly" edited events, whereas basally detected A-to-I edit events in untransfected cells were mostly above 10% (Figure 1.4B). Many of these events were "novel" in nature, meaning that

only a minority were found to overlap previously identified ADAR targets or naturally occurring editing events found in untransfected HEK293XT cells (Figure 1.9B, C). This result illustrates that off-target events detected occur largely *de novo* and are not simply an exacerbation of cellular RNA editing events.

With respect to on-target editing, successful *CYFIP2* editing was only detected in the presence of a targeting gRNA for each of the Cas constructs tested, again highlighting the gRNA-dependent nature of Cas-directed RNA editing (Figure 1.9D). In directly comparing on-target efficiencies of Cas-constructs we noticed slightly higher on-target editing rates for Cas9 and Cas13d versus Cas13b (Figure 1.4C), however this effect is counterbalanced by the observation that these fusions tended to produce a higher number of global editing events as well. Although a subset of editing sites overlapped between Cas proteins, most sites seemed unique to experimental condition, indicating that there were few "hot spots" and that the majority of off-target edits introduced were randomly deposited across the transcriptome as a secondary consequence of Cas-ADAR2(E488Q)DD overexpression (Figure 1.4D). However, sequence preference was similar transcriptome-wide among all conditions, with edits predominantly being sequestered into (U/A/C)AG trinucleotide sequences, previously described as being favorable edit motifs for ADAR family proteins (Figure 1.4E, Figure 1.9E) (Kim et al., 2004). We also observed an inverse relationship between read coverage and editing percentage per event, where extent of off-target events appeared to decrease with increasing coverage (Figure 1.9F). Although this observation is not a direct commentary on the severity of downstream transcriptional consequences, it does insinuate that highly edited events tend to be expressed at a relatively low level in relationship to the remainder of the transcriptome.

In characterizing the nature of these of off-target edits, we noticed that that majority of high confidence edits detected were in the 3'UTR (44-49%) or CDS (22-38%) of annotated genes (Figure 1.5A, Figure 1.10A). These edits are potentially consequential, with up to 70% of variant effects in CDS corresponding to a predicted missense mutation in at least one protein coding isoform (Figure 1.10B). This general phenomenon was observed across all editing platforms

regardless of gRNA, indicating that off-target effects are largely a product of non-specific effects of ADAR2(E488Q)DD overexpression. It is worth noting that the majority (52.3–69.7%) of global A-to-I edits tend to be low (less than 20% editing) for all the RNA editors tested (Figure 1.5B). Although off-target rates will inevitably decline with the discovery of improved base editors (Cox et al., 2017, Abudayyeh et al., 2019), it is important to note that all present Cas platforms tested behaved similarly transcriptome-wide and thus are seemingly prone to the same benefits and limitations with respect to targeted RNA editing.

1.4 Discussion

Programmable transcriptome manipulation offers the ability to transiently alter genetic information without conferring permanent changes to the genome. Site-directed RNA editing tools are currently under rapid development to allow for high precision single-nucleotide editing capability. Due to the well-characterized biological function and modular structure of ADAR family proteins, most of these technologies direct ADAR deaminase domains to a target adenosine for A-to-I conversion using a guide RNA bearing a mismatched cytidine base. There have been two major strategies in designing such base-editing systems: 1) recruitment of endogenous ADAR proteins using synthetic oligonucleotides (Qu et al., 2019, Wettengel et al., 2017, Woolf et al., 1995) or 2) exogenous expression of deaminase domains fused to effector proteins or oligonucleotides (Azad et al., 2017, Cox et al., 2017, Montiel-Gonzalez et al., 2016, Stafforst and Schneider, 2012, Vogel et al., 2018, Katrekar et al., 2019). Engineered ADAR fusions have an advantage in that they can be encoded genetically and are not dependent on the requirement for host cells to express ADAR proteins. Previously established ADAR fusion systems utilize the aptamer-binding Lambda N (λNN) and MS2 coat proteins, which bind BoxB and MS2 RNA hairpins adjacent to ADAR directing sequences within guide RNAs (Azad et al., 2017, Montiel-Gonzalez et al., 2016, Katrekar et al., 2019), to perform target-specific RNA editing in mammalian cells. Even more recently, RNA-targeting Cas13b proteins have been showcased as

ADAR effector fusion systems capable of directing both A-to-I (REPAIR) and C-to-U (RESCUE) edits on target transcripts (Abudayyeh et al., 2019, Cox et al., 2017). Cas13b represents just one of many RNA-targeting Cas proteins with potential utility as programmable RNA editors, including the other characterized Cas13 family members Cas13a and Cas13d (Abudayyeh et al., 2016, Konermann et al., 2018), as well as dual DNA-RNA targeting Cas9 (RCas9) proteins including SaCas9 and SpCas9 (Batra et al., 2017, Nelles et al., 2016, Strutt et al., 2018).

In this work, we demonstrate an additional function for the RCas9 system by fusing catalytically inactivated SpCas9 to the deaminase domain of human ADAR2 (ADAR2DD) to perform programmable A-to-I RNA editing on both reporter and cellular mRNA targets. RCas9-ADAR2DD editing is Cas9 and gRNA-dependent, as overexpression of ADAR2DD in the presence of a targeting gRNA alone is insufficient to trigger RNA editing (Figure 1.1C-D). While previous reports have stressed the significance of a spacer sequence and complementary PAMmer for RCas9 RNA localization and binding (Nelles et al., 2016, O'Connell et al., 2014, Liu et al., 2019), we find that RCas9-ADAR2DD is not reliant on a synthetic PAMmer oligonucleotide (Figure 1.1E) or even spacer sequences (Figure 1.2A, B), which simplifies delivery and expands targeting capacity. We find that RCas9-ADAR2DD is only reliant upon 21nt extension sequences located on the 3' terminus of the gRNA. We hypothesize that RCas9-ADAR2DD editing is enabled through hybridization of this extension with the target RNA and recruitment of the ADAR2DD fusion to the forced dsRNA substrate via RCas9 (Figure 1.2C). This targeting approach is mechanistically reminiscent to the recently described CIRTS system, which relies only on Watson-Crick-Franklin base pairing through a gRNA to recruit functional RNP complexes to specific sites on RNA (Rauch et al., 2019). Placing the A-C mismatch within the spacer region of the Cas9 gRNA could not induce successful targeting and RNA editing (Figure 1.6F, G) as has been observed with class 2 type VI RNA-guided RNA-targeting CRISPR–Cas systems (Abudayyeh et al., 2019, Cox et al., 2017), and hints at differences among various Cas systems in terms of RNA-targeting mechanism.

The aptamer-binding model of catalytically inactive Cas9 to scaffold-containing gRNA

for specific RNA-targeting could potentially be a useful tool for other applications requiring targeted protein tethering to specific RNA molecules bearing the gRNA scaffold sequence, analogous to the well-characterized MS2 coat protein and lambda N – BoxB systems (Keryer-Bibens et al., 2008). Seeing that the affinity for Cas9 binding to tracrRNA is potentially tighter than the binding affinity of these other coat protein systems, with Cas9 being shown to bind gRNA hairpins in the picomolar range (Wright et al., 2015), it will be of great interest to determine the search mechanism utilized by Cas9-effector fusions targeted to RNA with 3' gRNA extensions and whether this minimal design represents a universal Cas9-RNA targeting rule. Furthermore, this alternative 'spacerless' gRNA strategy expands the scope of RNA sites that can be targeted using RCas9, as this gRNA is no longer restricted to non-PAM adjacent target sequences, a constraint in designing target sites for RCas9 to prevent binding to genomic DNA. Notably, SaCas9 was also capable of RNA editing with a no-spacer (NS) gRNA configuration (Figure 1.3C, D), reinforcing the proposed aptamer-binding model across bacterial species and illustrating adaptability of this strategy to other Cas9 proteins.

We also evaluated the ability of RCas9-ADAR2DD fusions to edit endogenously expressed mRNAs in living cells. Like previously published RNA technologies, we show programmable and specific A-to-I editing within the 3'UTRs of housekeeping genes not normally edited by endogenous ADAR (Figure 1.2D), and within the coding sequence of the ADAR target mRNA *CYFIP2* (Figure 1.2E), which is normally edited through co-transcriptional exon-intron pairing. Since this editing platform can perform specific RNA editing on a variety of transcripts across multiple genic regions, it can readily be used for biologically relevant applications to study effects of individual RNA editing events and to reverse G-to-A mutations associated with disease cause or risk.

Seeing that the RCas9 system functions by repurposing a naturally DNA-targeting protein to recognize cognate RNA sequences, it is conceivable that that either natural RNA-targeting class 2 type VI Cas13 proteins would direct RNA editing with higher efficiency or specificity than Cas9 family members or that Cas9 and Cas13 family members behave differently from

one another in directed editing applications. To test this, we performed comparative analyses of available CRISPR-based RNA editing tools in various structural orientations and cellular localizations. Here we found comparable on-target EGFP-reporter and cellular mRNA editing rates between Cas9 and several Cas13 ADAR2DD fusions (Figure 1.3C-F). We observe that targeted RNA editing is context-specific, as alteration of orientation, and to a much lesser degree subcellular localization, could alter both on-target and non-specific RNA editing (Figure 1.3C, D). We also see some degree of background editing for Cas13b and Cas13d gRNAs, perhaps through recruitment of endogenous or overexpressed ADAR proteins (Figure 1.8C). This background editing may be caused by a number of factors including hybridization length as well as direct-repeat (DR) length or complexity, as gRNAs that have smaller DRs (Cas13b: 36nt, Cas13d: 30nt) tend to have higher backgrounds than those with longer scaffolds (SaCas9: 76nt, SpCas9: 85nt), though this mechanism is unclear. Background editing may also be due to hybridization of the duplex forming portion of these "bulge-containing" guides, as there appears to be some length-dependent increases in both on-target and background editing rates in the case of Cas13b gRNAs (Figure 1.8D-I). Together we were able to observe comparable on-target and off-target activity of our Cas9 RNA editors despite their secondary purpose as RNA-binding proteins.

In addition to evaluating local, on-target editing rate with Cas-ADAR2DD systems using targeted-amplicon NGS, we characterized global off-target capacity of each of these systems using poly(A)+ RNA-seq. Similar to our observation of comparable on-target efficiency between these Cas systems, we identified a widespread number of off-target edits for each of the systems we tested (Cas13b, Cas13d, Cas9). Although we were able to identify our targeted editing event (*CYFIP2*) via sequencing (Figure 1.9C), all conditions generated several thousand potential off-target A-to-I editing events, the highest numbers being detected in Cas13d (5,407) and Cas9 (5,545) (Figure 1.4B). The precise number of edits was variable among Cas protein and gRNA and was largely independent of gRNA design, as non-targeting gRNAs produced a comparable number of edits compared to their targeting counterparts (Figure 1.4B). These edits were largely random and artificial in nature, as most sites did not overlap with known editing

loci or previously characterized Alu sites, which comprise >95% of endogenous A-to-I editing modifications (Levanon et al., 2004). Instead, most detected editing sites were deposited in CDS and 3'UTR of genes (Figure 1.5A, Figure 1.9F), which could spell downstream consequences for protein coding capacity and stability of target transcripts (Figure 1.9G). We find that most of the most of the global editing events skewed low based upon percent editing (Figure 1.5B), and that events with highest sequencing coverage tended to have the lowest editing rate (Figure 1.9E). Overall, both the number of editing events and degree of off-target editing must be considered limitations in these cases of post-transcriptional genome manipulation, though may still be a viable alternative to DNA editing so long as RNA editing tools are deployed in a transient manner.

Due to their encodable and highly programmable nature, as well as demonstrated RNA-targeting capacity, CRISPR/Cas technologies remain tremendously promising for targeted RNA base editing. In this work we expand and benchmark the list of orthogonal and available RNA editing CRISPR technologies. This study furthers our understanding of Cas-based technologies and helps to set the stage for further optimization. So far, rational directed evolution-based approaches have been employed to expand the sequence targeting toolbox of both DNA and RNA base editors (Abudayyeh et al., 2019, Thuronyi et al., 2019). We believe that a combination of both these engineering approaches, as well as a firm understanding of the sequence-based targeting 'rules' of RNA-guided RNA targeted CRISPR/Cas system, will be necessary for further development of these powerful platforms in order to accomplish highly efficient and specific therapeutics for disease-relevant mutations.

1.5 Materials and Methods

1.5.1 Plasmid construction

For dCas9-ADAR2DD mammalian expression constructs, dCas9-2XNLS was amplified from pCDNA3.1- dCas9-2xNLS-EGFP (Addgene plasmid #74710) and fused to the human

ADAR2 deaminase using Gibson assembly into a pcDNA(-)3.1 (Invitrogen) backbone, which had been digested using FastDigest EcoRI (Thermo Scientific). Domains were designed to be separated by an XTEN linker peptide, the sequence of which was included in amplification primers. The ADAR2DD-XTEN control plasmid was generated by eliminating the dCas9-2XNLS domain through inverse PCR using the primers ADAR2_CD_Inverse_F and ADAR2_CD_Inverse_R (primer sequences located in Supplementary Table 1), followed by ligation with T4 DNA Ligase (New England Biolabs). Inverse PCR and ligation was also used to perform site-directed mutagenesis to generate the E488Q variants using the primers E488Q_Mut_F and E488Q_Mut_R. These fusions were then amplified and cloned into the pcDNA5/FRT/TO (Invitrogen) backbone using HindIII and NotI restriction sites for stable line generation. For comparison studies, sequences for an HA affinity tag in frame with either a 2XNLS or HIV1 NES localization sequence was cloned into the pcDNA(-)3.1 multiple cloning site using EcoRI and HindIII restriction sites. These constructs were then digested with EcoRI and used as backbones into which both AXC and CXA fusions were integrated via Gibson assembly. For each 3' extension Cas9 guide RNA constructs, the gRNA scaffold sequence was amplified from pBluescriptSKII+ U6-gRNA(F+E) empty (Addgene plasmid #74707) with forward primer containing the desired spacer sequence with and reverse primer containing the 3' extension sequence on the 5' tail of the oligonucleotide (spacer and extension sequences located in Supplementary Table 2), as well as complementary sequences for Gibson assembly (5'- AAAGGACGAAACACC-3' for forward primer, 5'- CCCGGGCTGCAGGAAAAAAA-3' for reverse primer). These products were then Gibson assembled into the pBluescriptSKII+ U6-gRNA(F+E) backbone which had been inversely amplified using gRNA_backbone_F and gRNA_backbone_R primers. Other Cas-construct gRNAs were assembled in this fashion using primers to amplify SaCas9 gRNA (IDT gBlock of 3F gRNA sequence described in Chen et al. 2016), LwaCas13a gRNA (a gift from Feng Zhang; Addgene plasmid # 91906), PspCas13b gRNA (a gift from Feng Zhang; Addgene plasmid #103854), and RfxCas13d gRNA (a gift from Patrick Hsu; Addgene plasmid # 10905). For the dual fluorescence reporter, the W58X EGFP coding sequence was PCR

amplified from the pCDNA3.1-EGFP(W58X) construct (a gift from Thorsten Stafforst's lab, University of Tübingen), and fused to P2A-mCherry sequence using Gibson assembly downstream of an Ef1a constitutive promoter. This sequence was then inserted into the pBluescriptSKII+ U6-gRNA(F+E) backbone using the SmaI restriction site and Gibson assembly to allow for simultaneous gRNA and reporter delivery to cells.

1.5.2 Human cell culture conditions and maintenance

The Flp-In T-REx 293 cells (Invitrogen) and lenti-X HEK293T cells (HEK293XT, Takara Bio) are derived from transformed female human embryonic kidney tissue. Cells were maintained in DMEM (4.5 g/L D-glucose) supplemented with 10% FBS (Gibco) at 37°C with 5% CO2. Cells were periodically passaged once at 70-90% confluency by dissociating with TrypLE Express Enzyme (Gibco) at a ratio of 1:10. Additionally, Flp-In-293 cells were maintained under 5 μg/ml Blasticidin S (Gibco) to ensure presence of an integrated FRT site. These lines were purchased directly from manufacturers.

1.5.3 Generation of Flp-In 293 cell lines

Generation of the Tet-inducible dCas9-ADAR2DD lines was performed according to the manufacturer's instructions. Open reading frames were cloned into the pcDNA5/FRT/TO backbone as previously described. Once reaching 70-90% confluency on a 10 cm^2 dish, Flp-In T-REx cells were transfected with 9μg of plasmid and pcDNA5 plasmid and 1μg of pOG44 Flp-Recombinase Expression Vector (Invitrogen) using polyethylenimine (PEI, Sigma Aldrich) in 250 ul Opti-MEM (Gibco). Cells were passaged with TrypLE to 25% confluency 48 hours after transfection and incubated for an additional 2-3 hours at 37o C with 5% CO2. Cells were then selected with 200 ug/ml Hygromycin B (Gibco) for 3-4 days until individual clones could be picked, expanded, and genotyped.

1.5.4 Western Blot

Cells were lysed in 1% NP-40 lysis buffer (20 mM Tris HCl pH 8, 137 mM NaCl, 1% Nonidet P-40, 2 mM EDTA) and quantified with the Pierce BCA Protein Assay Kit (Thermo Fisher). After quantification, 10 μg total protein lysate was loaded on a 4-12% NuPAGE Bis-Tris gel (Thermo Fisher) and transferred onto a PVDF membrane. Membranes were blocked and probed with primary and secondary antibodies in 5% milk and imaged with Pierce ECL Western Blotting Substrate (Thermo Fisher). For immunoblotting, the following antibodies were used: anti-HA.11 Epitope Tag Antibody (Biolegend, # MMS-101R), anti-GAPDH (Abcam, # ab9485), Goat anti-Mouse IgG (H+L) Secondary Antibody, HRP (# 32430).

1.5.5 Transient transfection of human cell lines for EGFP restoration

For comparison experiments, cells were seeded in a 24-well plate 24 hours prior to transfection. At approximately 70-80% confluency, cells were transiently transfected with 500 ng gRNA/EGFP reporter plasmid for stable lines, or 300 ng gRNA/EGFP reporter and 200 ng Cas9-ADAR2 fusions for HEK293XT cells using Lipofectamine 3000 Reagent (Invitrogen) per well, according to manufacturer's instructions. For endogenously expressed targets, 500 ng gRNA plasmids were transfected per well. For experiments with PAMmers, 0.5 pmol was transfected with Lipofectamine RNAiMAX (Invitrogen) immediately after DNA transfection. For experiments involving Flp-In T-REx lines, fusion protein expression was induced by adding doxycycline (Sigma Aldrich) at a final concentration of 1 ug/ml to cell media. Experiments unless otherwise noted were performed in triplicate. After 48 hours, cells were dissociated with TrypLE and harvested for both RNA and FACS analysis. Data was plotted and statistics were calculated using Prism 6 software.

1.5.6 RNA editing of W58X EGFP reporter using FACS

Following 48 hours post transfection, cells were rinsed once with 1X DPBS (Corning) and dissociated with 400 ul TrypLE for 5-10 minutes at 37o C. Cells were then resuspended in

FACS Buffer (10% FBS in 1X DPBS), strained over a 35 um nylon mesh, collected in round-bottom Falcon tubes and subjected to fluorescence-activated cell sorting (FACS) analysis using a BD LSRFortessa flow cytometer. After excluding cellular debris and doublets, transfected cells were first gated on mCherry-positive signal. The overall fraction of EGFP-positive cells in this mCherry+ population as calculated as a readout for successful RNA editing. FACS data was analyzed and plotted using the FlowJo software package.

For comparison experiments, HEK293XT cells were seeded on 48-well plates 24 hours prior to transfection. When cells were 70-80% confluent, they were co-transfected with 300 ng respective gRNA/EGFP reporter and an equimolar amount of Cas-ADAR2DD expression plasmid (for reference, SpCas9-ADAR2 2XNLS fusion: 0.007 pmol = 50ng) in triplicate. After 48 hours, cells were dissociated with TrypLE, resuspended in FACS buffer as previously describes, and transferred to flat-bottom 96-well plates. Samples were then processed on High Throughput Sampler (HTS) mode on a BD LSRFortessa X-20 instrument. Gating and subsequent analyses was performed as described above, where approximately 25,000 live cell events were counted for each sample.

1.5.7 Fluorescence visualization of live cells

After 48-hours following transfection and dCas9-ADAR2DD protein expression, live cells were subject to fluorescence imaging at 10-20X using a Zeiss Axio Vert.A1 fluorescence microscope equipped with an X-Cite 120Q illumination system. Images were captured at 20 ms for brightfield and mCherry, and 180 ms for EGFP fluorescence at 1X gain. Images were processed with compatible ZEN software, then exported and adjusted for brightness/contrast by matching Maximum/Minimum values across samples for fluorescent images in analysis software FIJI (Schindein et al. 2012).

1.5.8 Extraction of RNA and RT-PCR

RNA isolations were carried out by resuspending cells 500 ul TRIzol reagent (Invitrogen) and extracting using the Direct-zol RNA Miniprep kit (Zymo Research) according to manufacturer's instructions. Samples were eluted in 25-50 ul nuclease-free H2O and concentrations were measured using a NanoDrop 2000 spectrophotometer (Thermo Scientific). For each sample, 500 ng to 1 ug of total RNA was reverse transcribed to cDNA with Superscript III Reverse Transcriptase (Invitrogen) using oligo(dT) and random hexamers according to manufacturer's instructions. From this cDNA, 1 ul was taken and amplified using flanking primers for 35 cycles. Amplified products were extracted using QIAquick PCR Purification Kit (Qiagen) and subjected to Sanger sequencing. When quantifying Sanger traces, editing efficiencies were calculated using FIJI software using the following metric: (height of 'G' peak) / ((height of 'G' peak) + (height of 'A' peak)) at each target site.

1.5.9 Targeted RNA editing analysis for endogenous transcripts

For targeted amplicon next-generation sequencing (NGS), HEK293XT cells were first seeded onto 24-well plate 24 hours prior to transfection. When cells reached 70-90% confluency, cells were co-transfected with 100 ng Cas-ADAR2DD plasmid and 400 ng of respective gRNA plasmid using Lipofectamine 3000. After 48 hours, RNA was extracted using TRIzol reagent and 10 ng total RNA was reverse transcribed using Protoscript II Reverse Transcriptase (New England Biolabs) with target-specific reverse primers according to manufacturer's instructions. 4 ul cDNA was amplified for 15 cycles using target-specific primers containing partial Illumina adapters using NEBNext Ultra II Q5 polymerase (New England Biolabs). PCR products were purified using AMPure XP beads (Beckman Coulter) and 2ul of the first reaction was subjected to a second round of PCR for 15 cycles to affix Illumina-compatible indices (based upon TruSeq RNA adapters #1-14) using NEBNext. Reactions were purified again using AMPure XP beads, size-checked and quantified on the Agilent 2200 Tapestation system using D1000 ScreenTape

reagents, pooled equally and diluted to 2nM final concentration, spiked-into high complexity RNA-seq libraries at 1% of input material and read out on a 75-nt single-end run with an Illumina HiSeq4000 instrument.

Successfully sequenced libraries were first subject to quality control using FastQC, then aligned using Burroughs-Wheeler Aligner (BWA) (Version: 0.7.15-r1140)(Li and Durbin, 2009) using the command 'bwa mem -t 8 (index.db) (sample.fq)'. Aligned BAMs were sorted using Samtools (Li et al., 2009). Editing percentage per base position was calculated at each annotated adenosine (A) residue by calculating:

$$\frac{(fraction\ of\ reads\ called\ 'G')}{((fractionofreadscalled'G'))/(((fraction\ of\ reads\ called\ 'G')+(fraction\ of\ reads\ called\ 'A')))}$$

1.5.10 Transcriptome-wide RNA sequencing

RNA editing analysis was performed in transiently transfected HEK293XT cells. Cells were first seeded onto a 24-well plate format, 24 hours prior to transfection. Upon reaching 70-90% confluency, cells were co-transfected with 100 ng Cas-ADAR2DD plasmid along with 400 ng of either a scrambled, non-targeting or CYFIP2-targeting gRNA plasmid using Lipofectamine 3000. After 48 hours incubation, RNA was extracted using TRIzol reagent and Direct-zol RNA purification reagents. Following elution, 100 ng of total RNA was quantified and used as input for poly(A)+ library preparation using TruSeq Stranded mRNA preps (Illumina) according to manufacturer's instructions. All libraries were sequenced on an Illumina HiSeq 4000 platform (100bp, paired-end setting) with the exception of the untransfected control HEK293XT samples, which were sequenced on an Illumina NovaSeq 6000 platform (100bp, single-end setting) during a separate experimental run.

1.5.11 RNA-seq analysis

RNA-seq reads were adapter-trimmed using Cutadapt (version 1.14) (Martin, 2011) and aligned to the hg19 (GRCh37) genome using STAR (version 2.5.2b) (Dobin et al., 2013)using default options, and subsequently sorted with samtools (version 1.5). Transcript-per-million

(TPM) normalized counts were calculated following raw counts quantitation using featureCounts (version 1.5.0) (Liao et al., 2014). For quantification of Cas-ADAR2 fusions, raw reads were mapped to the hg19 reference genome with Cas-ADAR2 sequences added. For editing analysis, to accommodate sensitivity towards differing sequencing depths in variant calling, aligned reads were then randomly downsampled to 16-35 M mapped read pairs (32-70 M total mapped reads) using samtools view -bs to achieve desired sequencing depths for paired-end libraries. Aligned reads were then subjected to variant calling using the SAILOR (version 1.0.4) software program (Deffit et al. 2017) using default parameters. Candidate A-G variants (or T-C for negative strand) identified were furthermore filtered for read coverage (minimum 5 reads per site) and as well as naturally occurring variants using hg19 Common SNPs (147). These candidate editing sites were assigned a confidence score using a previously described Bayesian model (Bahn et al., 2012; Li et al., 2008) factoring in both read coverage and percent edited reads. Individual sites with less a 90% confidence score were not considered for downstream analysis. Consensus sites were for each experimental condition were determined as such by being called in both replicates. For each of these sites, per-site editing rate was recalculated by combining read coverage for both replicates and computing:

$$\frac{(fraction\ of\ reads\ called\ 'G')}{((fractionofreadscalled'G'))/(((fraction\ of\ reads\ called\ 'G')+(fraction\ of\ reads\ called\ 'A')))}$$

for positive strand and:

$$\frac{(fraction\ of\ reads\ called\ 'C')}{((fractionofreadscalled'G'))/(((fraction\ of\ reads\ called\ 'C')+(fraction\ of\ reads\ called\ 'T')))}$$

for negative strand. Gene region annotations were gathered from human hg19 gencode release 19 (GRCh37.p13). Intersecting sites were determined using bedtools/pybedtools software libraries (Quinlan and Hall, 2010, Dale et al., 2011), and sequence logos were generated using Weblogo 3 (Crooks et al., 2004). Protein coding consequences were predicted using Snpeff (Cingolani et al., 2012).

1.6 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Cell Reports 2020. Marina RJ, Brannan KW, Dong KD, Yee BA, Yeo GW, Cell Press, 2020. The dissertation author was the primary investigator and author of this paper.

We acknowledge members of the Yeo. lab, particularly Dr. David Nelles and Dr. Frederick Tan for comments and assistance in initial construct design. We thank Dr. Thorsten Stafforst of the University of Tübingen, Germany for the gift of the W58X EGFP reporter and Dr. Patrick Hsu of the University of California, Berkeley for the gift of the RfxCas13d expressing plasmid used in cloning. This work was partially supported by grants from the NIH (R01HG004659, R01EY029166, and R01NS103172) to G.W.Y. R.J.M. was supported in part by an institutional award to the UCSD Genetics Training Program from the National Institute for General Medical Sciences (T32GM008666) and by a NIH/NINDS Ruth L. Kirschstein National Research Service Award (F31NS111859). K.W.B is a University of California President's Postdoctoral Fellow and is supported by a NIH/NINDS Career Transition Award (K22NS112678). This publication includes data generated at the UC San Diego IGM Genomics Center utilizing an Illumina NovaSeq 6000 that was purchased with funding from a NIH SIG grant (S10OD026929).

1.7 Figures

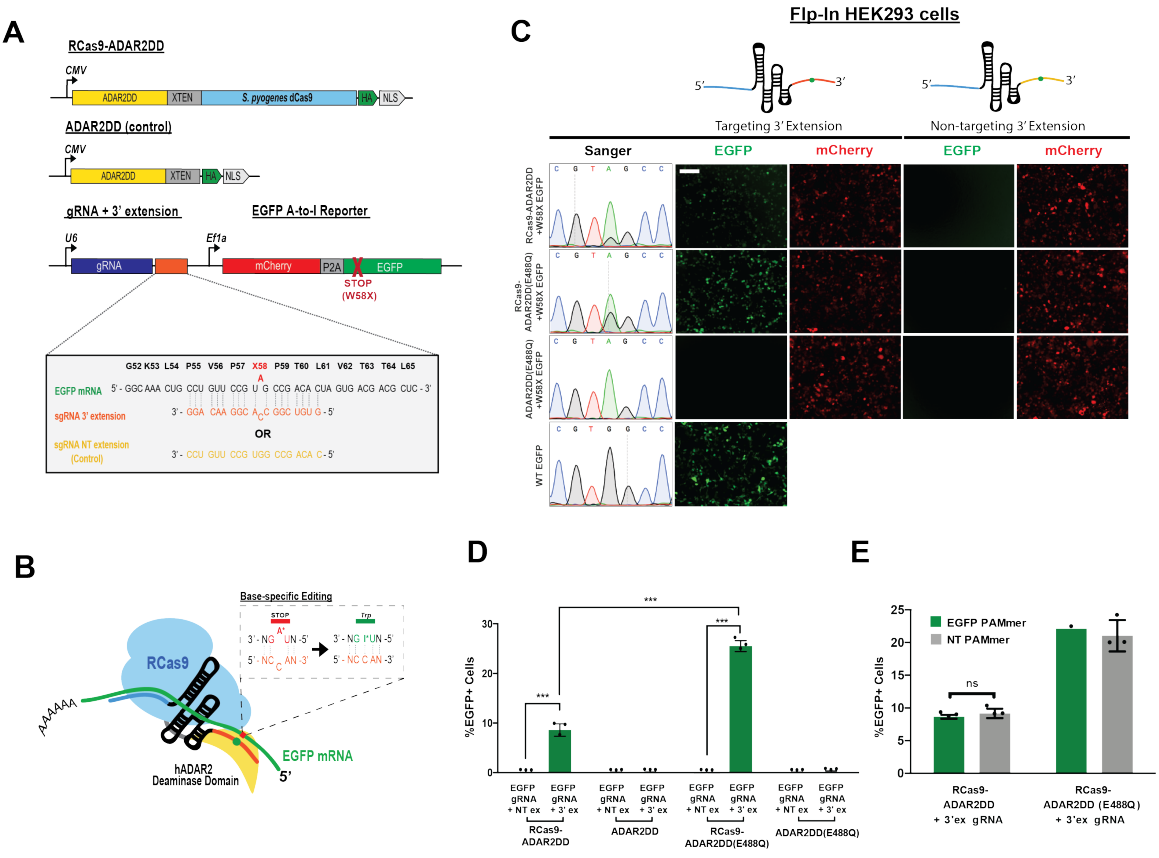


Figure 1.1. Editing cellular transcripts using RCas9-ADAR2DD. A, Schematic of RCas9-ADAR2 deaminase domain fusion constructs and modified guide RNA (gRNA) constructs fused to an W58X EGFP reporter. B, Schematic design of RCas9-directed RNA editing strategy using RCas9-ADAR2DD fusion proteins. A 3' extension sequence forms a double-stranded RNA (dsRNA) substrate with the target RNA and forms a bulged residue to allow for directed adenosine ('A') base editing. C, Sanger electropherograms and representative fluorescent images for EGFP+ cells measuring successful RNA editing of the W58X reporter in RCas9-ADAR2DD, hyperactive mutant (RCas9-ADAR2DD(E488Q)), or ADAR2DD(E488Q) control expressing Flp-In T-REx 293 lines along with either a targeting or a reverse complement non-targeting (NT) 3' extension (3'ex) sequence attached to the gRNA following doxycycline induction. As a transfection control, cells were also imaged for mCherry+ signal. Scale bar = 500 μm. D, Quantification of EGFP+ cells for Flp-In 293 cells following transfection with targeting or non-targeting 3' extension gRNAs using FACS. EGFP+ percentages were calculated as a percentage of mCherry+ cells. E, Cells were transfected with a targeting 3' ex gRNA in the presence of either an EGFP-targeting or scrambled, non-targeting (NT) PAMmer sequence in both RCas9-ADAR2DD and RCas9-ADAR2DD(E488Q) expressing Flp-In 293 lines and quantified using FACS. Data are mean values ± s.d. with n= 1 - 3; unpaired two-tailed Student's t-test, * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$, n.s. = not significant

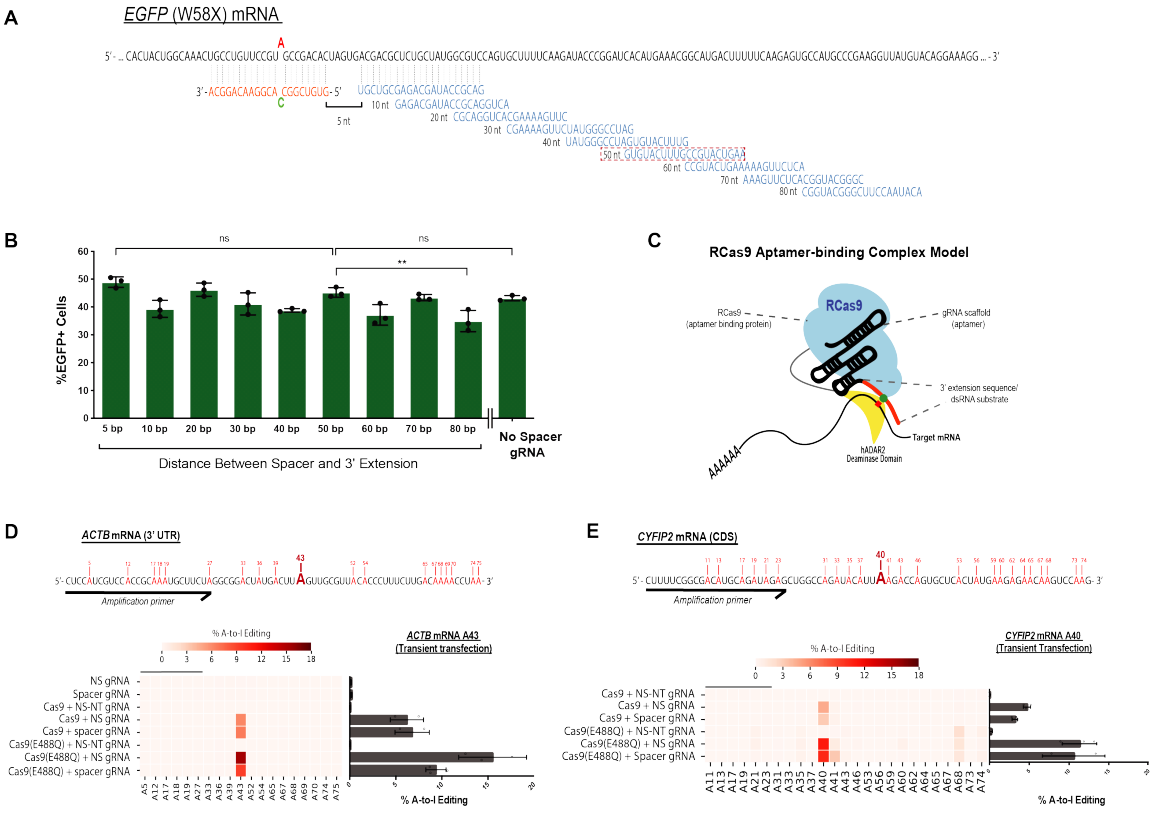


Figure 1.2. Spacer sequence of the modified gRNA is dispensable for RCas9-mediated RNA editing. A, Schematic of EGFP target mRNA and tiling gRNA designs. 3' extension sequence (orange) remains constant while the spacer sequence (blue) is variable in tiling EGFP mRNA reporter (black). The red box notes the original spacer sequence used in previous experiments B, FACS quantification of EGFP+ cells for tiling gRNAs transfected into RCas9-ADAR2DD(E488Q) expressing Flp-In T-REx 293 cells. Number of nucleotides (nt) refers to spatial distance between spacer sequence target and edit site on EGFP mRNA. No Spacer (NS) gRNA contains a targeting 3' extension sequence with no spacer complementary spacer sequence. C, Schematic for proposed 'aptamer-binding complex' model which would feasibly allow for spacer-independent RNA targeting and editing. D, E, Heatmaps depicting targeted amplicon next-generation sequencing (NGS) of ACTB and CYFIP2 target and adjacent adenosine residues. Bar plots (right) summarizing A-to-I editing efficiency of target adenosine for both ACTB (A43) and CYFIP2 (A40). Data are mean values ± s.d. with n= 2 - 3; unpaired two-tailed Student's t-test, * $P < 0.05$; * $P < 0.01$; *** $P < 0.001$; n.s. = not significant.

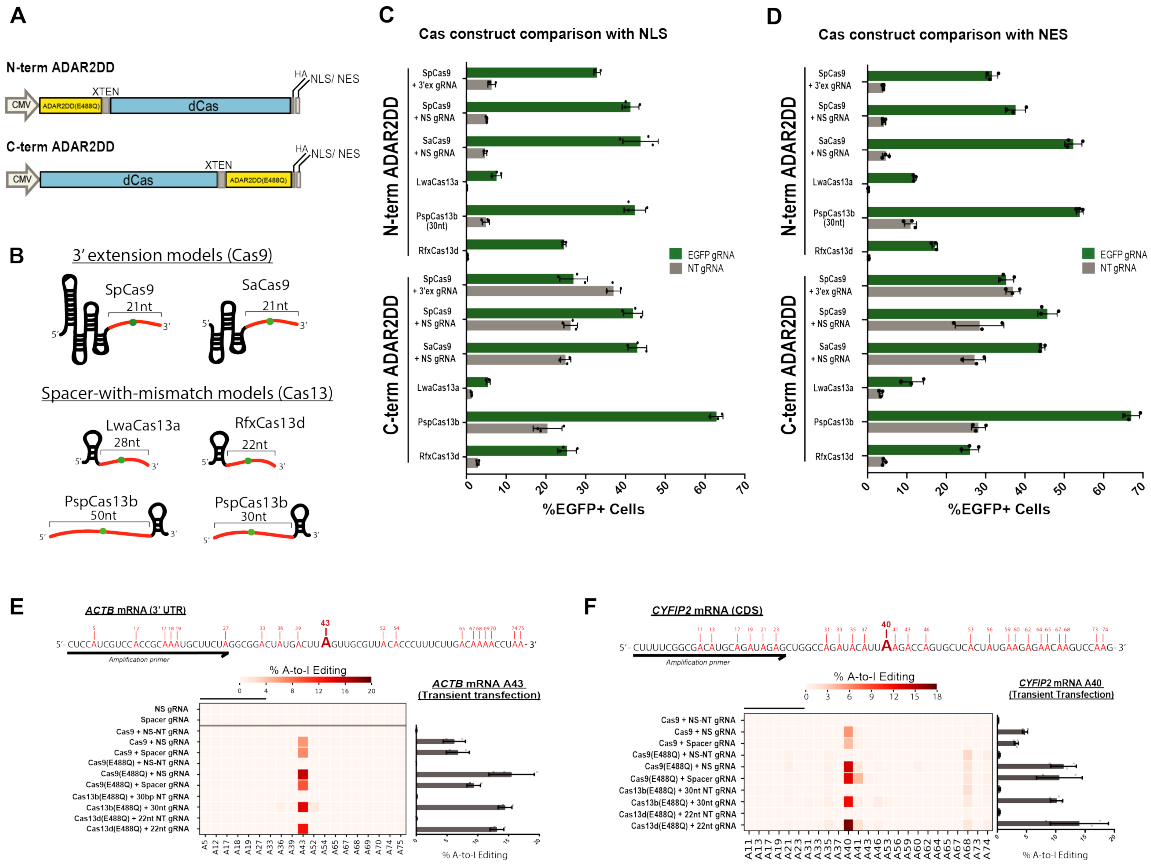


Figure 1.3. Comparative analysis of editing efficiencies across Cas-ADAR2DD platforms. A, Schematic depicting protein fusion orientations for editing comparisons. N-terminal ADAR2DD fusions were noted as being in the 'N-TERM ADAR2DD' orientation, while C-terminal ADAR2 fusions were noted as being in the 'C-TERM ADAR2DD' orientation. Domains were separated by an 'XTEN' peptide linker, and subcellular localization was determined by a 2X SV40 nuclear localization signal (NLS) or HIV1 cytoplasmic localization signal (NES). B, Visual representations of gRNAs being tested. The respective RNA-hybridization sequences (orange) are marked with the mismatched base (green). High throughput EGFP+ FACS analysis was performed for both orientations on NLS C, and NES D, containing Cas-ADAR2DD fusions through co-transfection of HEK293XTs after 48 hours. Samples were measured in both orientations using specific EGFP-targeting or scrambled NT gRNAs for each Cas species. In all conditions, n= 3. E, F, Targeted amplicon NGS of ACTB (E) and CYFIP2 (F) target regions transiently transfected with SpCas9 (RCas9), Cas13b, and Cas13d along with targeting and non-targeting gRNAs. RNA was collected and libraries were prepared 48 hours after transfection. Heatmaps (left) depict relative A-to-I editing ratios for both target and adjacent adenosine residues, while bar plots (right) represent A-to-I editing efficiency of target adenosine for both *ACTB* (A43) and *CYFIP2* (A40) in replicate, with n=2-3. Data represented are mean values ± s.d.

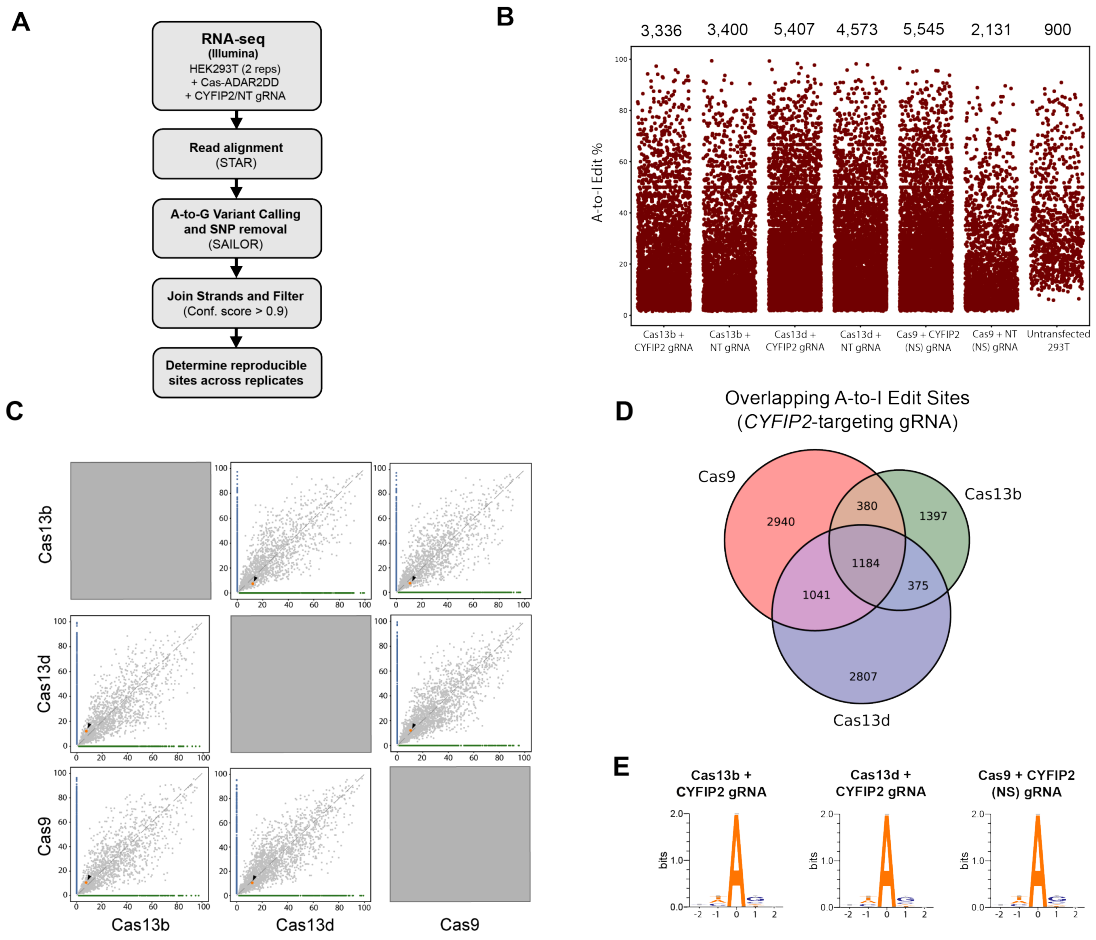


Figure 1.4. Global RNA-seq reveals transcriptome-wide off target consequences across Cas-ADAR2DD(E488Q) platforms. A, Outline of RNA-seq processing pipeline to identify global A-to-I edits. B, Strip plots illustrating number of global A-to-I identified in each experimental condition, as well as percent edited per site (y-axis). Also represented are the total consensus editing events identified from untransfected 293T cells. Editing rate calculation for experimental RNA-seq samples consists of n=2 individual replicates together, while untransfected control data was generated with n=3 replicates. C. Pairwise scatterplots illustrating relative editing efficiencies between Cas proteins. The target edit site (*CYFIP2* CDS) is colored orange and indicated with an arrow. Green and blue events symbolize those detected exclusively in sample #1 (x-axis) or sample #2 (y-axis). D, Three-way Venn diagram illustrating number of reproducible edits between Cas conditions. The majority of off-target events appeared to be unique in the cases of Cas9 and Cas13d, which contained the highest number of global edit events. E, Sequence logos of edited adenosine residues identified from RNA-seq data in the presence of *CYFIP2*-targeting gRNA and Cas13b, Cas13d, or Cas9.

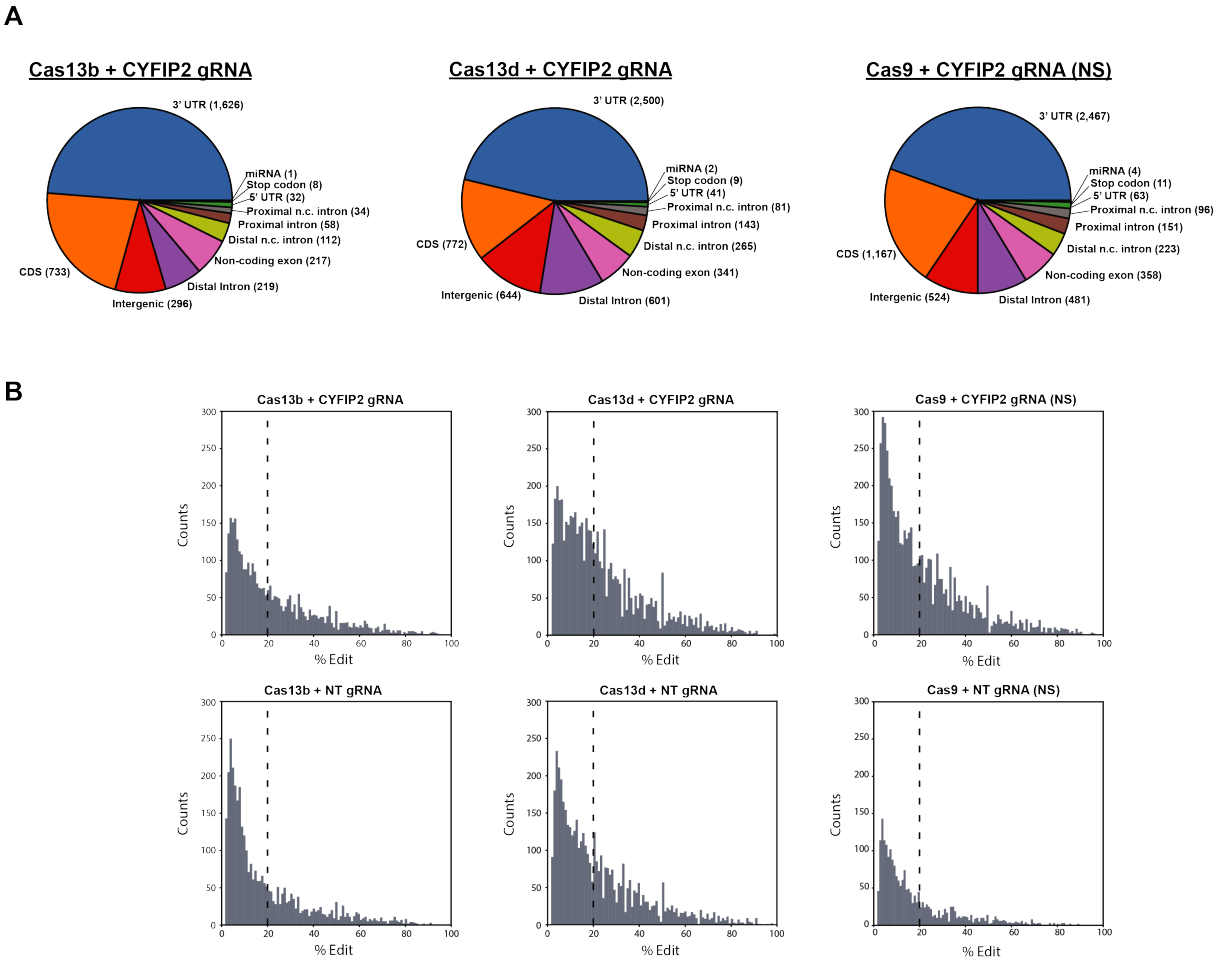
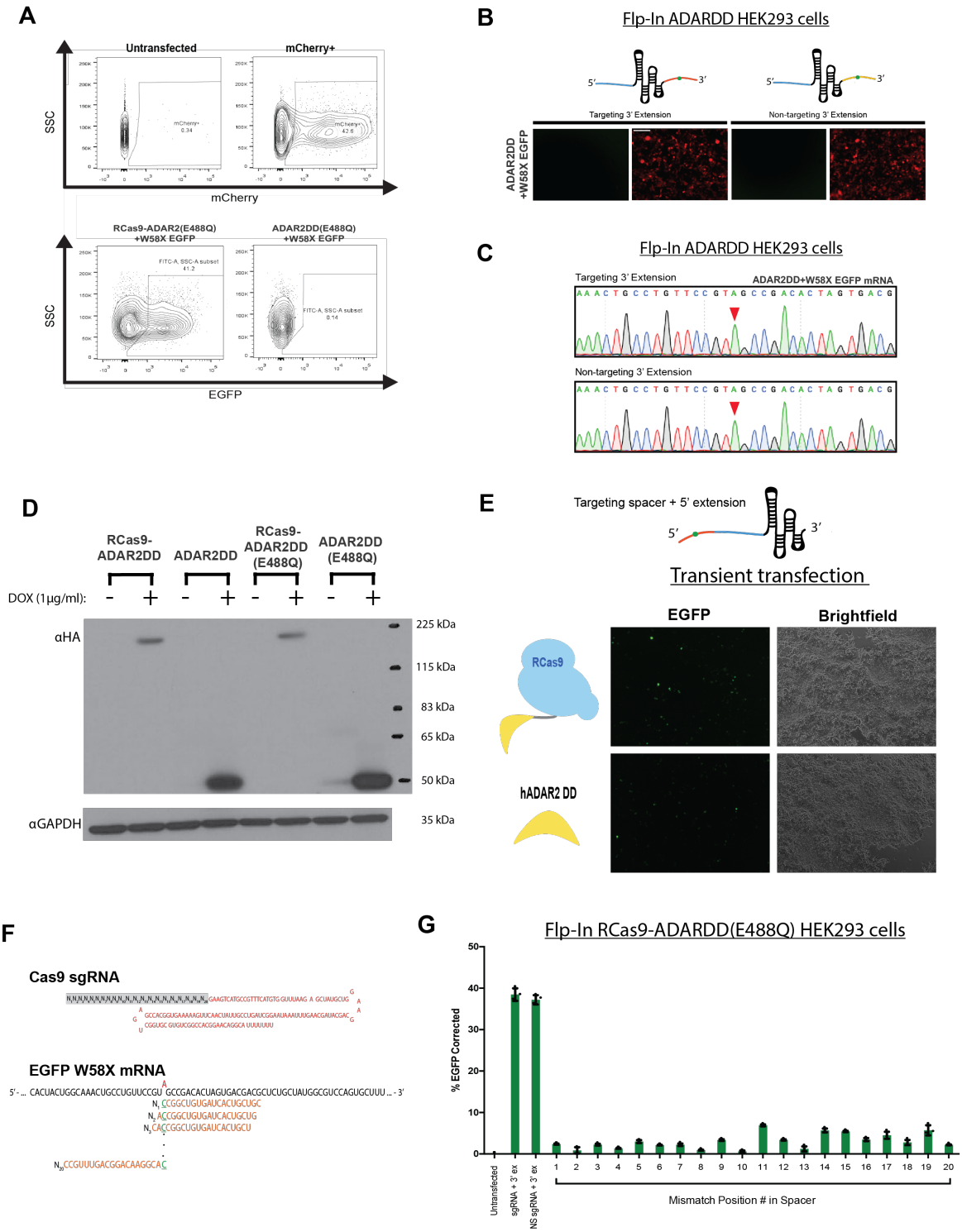


Figure 1.5. Nature of global RNA edits introduced through Cas-ADAR2DD(E488Q) expression. A, Regional distribution of total edits per Cas condition with *CYFIP2*-targeting gRNA. The vast majority of edits identified compiled in gene bodies, particularly in 3'UTR and CDS for all Cas proteins. The number of events per annotated region are located in parentheses. B, Histograms illustrating distribution of identified sites by editing rate. The vertical lines depicted are meant to signify the fraction of sites edited less than 20% in each condition.

Figure 1.6. Optimization of components necessary for specific RCas9-ADAR2DD editing. A, Representative FACS plots summarizing gating and event counting strategy for reporter editing quantification. Transfected cells were first counted by gating for mCherry+ cells (top). Of this subpopulation, an EGFP+ population could further be determined (bottom). A conversion percentage was calculated by taking the ratio of EGFP+ cells over the total number of mCherry+ cells. B, ADAR2DD expressing T-REx Flp-In 293 cells fail to restore EGFP signal in the context of both a targeting and non-targeting (NT) 3' extension modified gRNA. Scale bar = 500 μm. C, Sanger traces showing an inability to edit target or directly adjacent adenosine residues using ADAR2DD control for both targeting and non-targeting 3' extension gRNAs. D, Western blot of Flp-In 293 stable cell lines expressing RCas9-ADAR2DD and ADAR2DD fusion proteins. E, Presence of background EGFP signal (left) initiated through a 5' extension modified targeting gRNA. Both RCas9-ADAR2DD and ADAR2DD control were able restore EGFP signal, indicating that RNA editing occurs in a Cas9-independent fashion. Brightfield images (right) show that the relative confluency of cells imaged was similar. F, Schematic of SpCas9 gRNA design, with each 'N' residue representing a base where the mismatched cytosine could be placed (top). EGFP reporter mRNA was tiled with several spacer-as-substrate gRNAs (bottom). Nn refers to the base position of the mismatched cytosine residue (green) within the spacer sequence. G, EGFP+ FACS data for tiled EGFP gRNAs. Untransfected cells were used as negative controls, while gRNAs with 3' extension with and without spacer sequence (gRNA + 3' ex and NS gRNA + 3' ex, respectively) were used as positive controls. All data represented are mean values ± s.d with n= 3.


A
Untransfected
mCherry+
SSC
mCherry
RCas9-ADAR2(E488Q)
+W58X EGFP
ADAR2DD(E488Q)
+W58X EGFP
EGFP
B
Flp-In ADARDD HEK293 cells
Targeting 3' Extension
Non-targeting 3' Extension
ADAR2DD
+W58X EGFP
C
Flp-In ADARDD HEK293 cells
Targeting 3' Extension
ADAR2DD+W58X EGFP mRNA
Non-targeting 3' Extension
D
RCas9-
ADAR2DD
ADAR2DD
RCas9-
ADAR2DD
(E488Q)
ADAR2DD
(E488Q)
DOX (1µg/ml):
αHA
αGAPDH
225 kDa
115 kDa
83 kDa
65 kDa
50 kDa
35 kDa
E
Targeting spacer + 5' extension
Transient transfection
EGFP
Brightfield
RCas9
hADAR2 DD
F
Cas9 sgRNA
EGFP W58X mRNA
G
Flp-In RCas9-ADARDD(E488Q) HEK293 cells
% EGFP Corrected
Untransfected
sgRNA + 3' ex
NS sgRNA + 3' ex
Mismatch Position # in Spacer

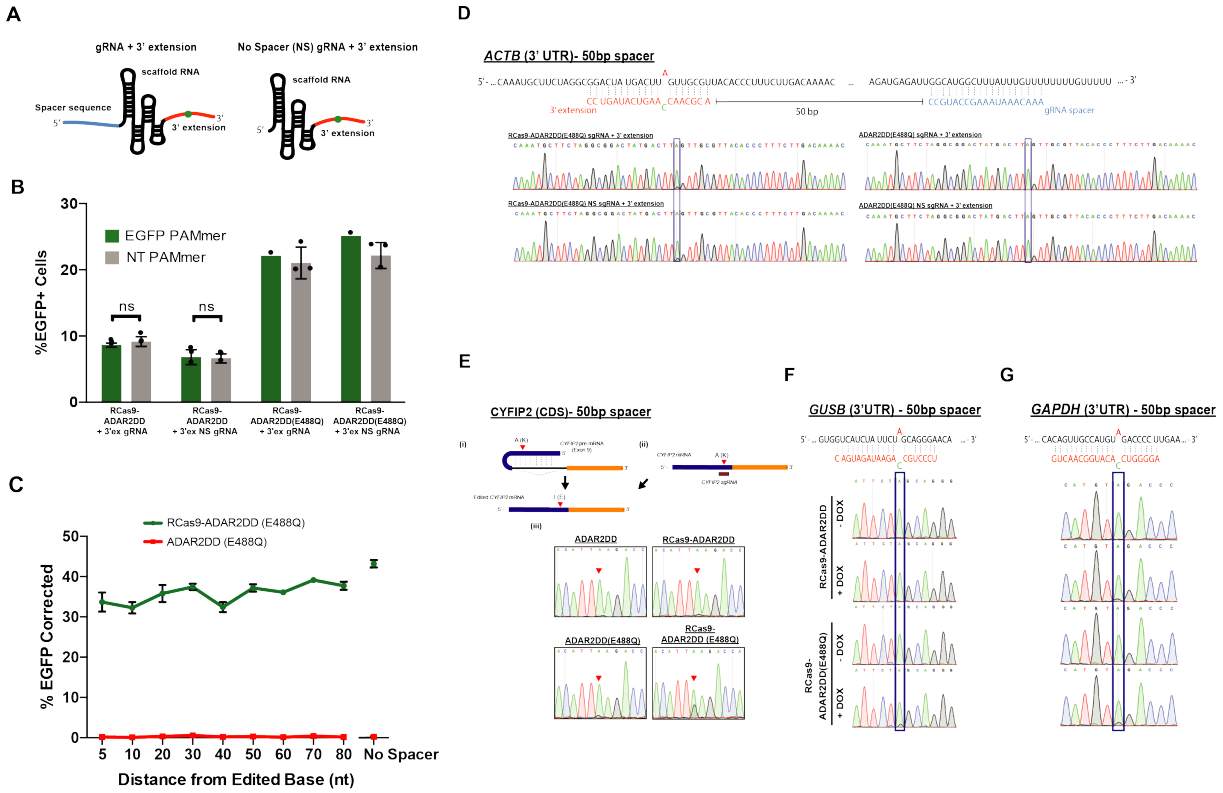
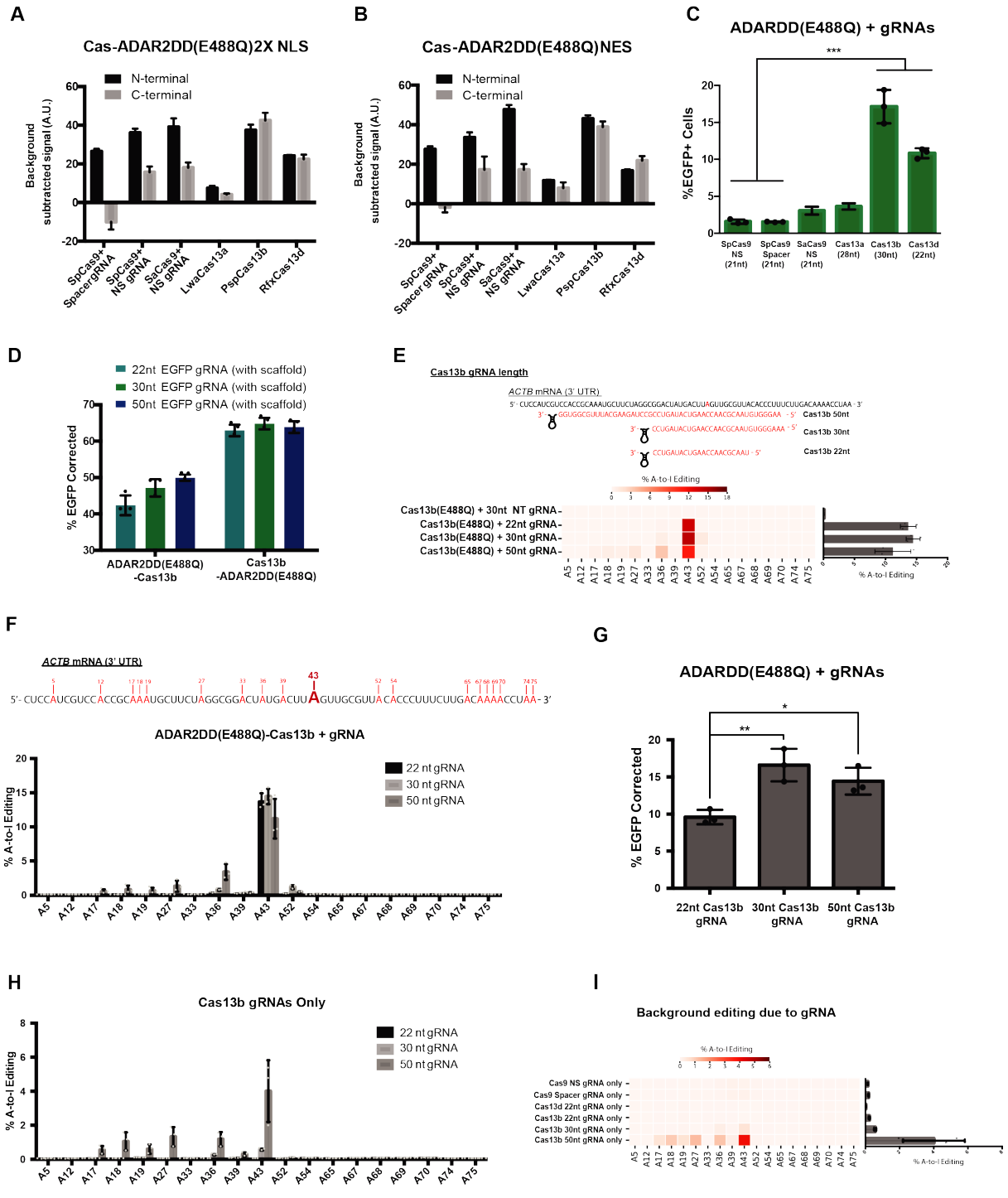
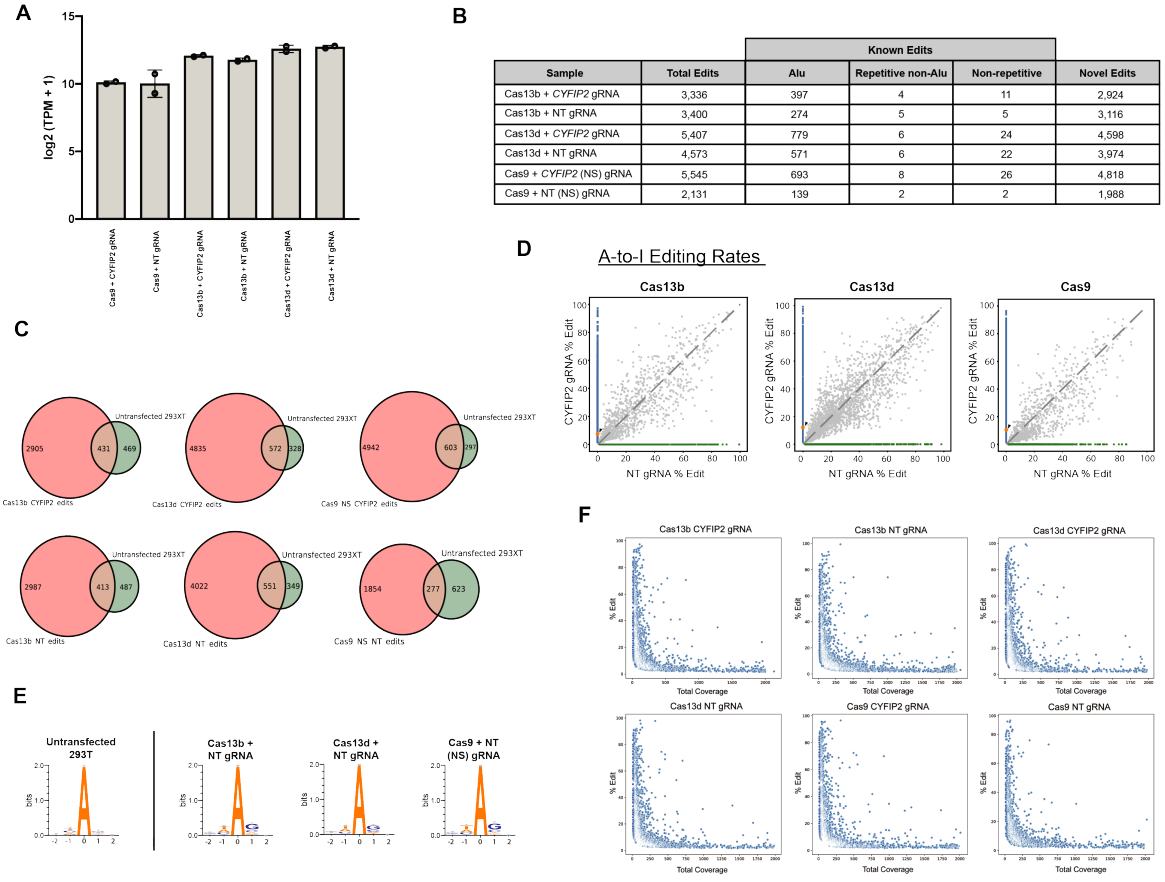


Figure 1.7. Targeted editing of cellular transcripts reveals Cas9-dependent and spacer-independent nature of RCas9-ADAR2DD. A, Schematic for proposed 'aptamer-binding complex' model which would feasibly allow for spacer-independent RNA targeting and editing. B, EGFP+ FACS data quantifying editing efficiency of spacer containing and NS gRNAs in the presence of EGFP-targeting or NT PAMmer for RCas9-ADAR2DD and RCas9-ADAR2DD(E488Q). C, FACS quantification of EGFP+ cells for tiling gRNAs transfected into RCas9-ADAR2DD(E488Q) and ADAR2DD(E488Q) expressing Flp-In T-REx 293 cells. None of the guides tested, result in noticeable EGFP editing in ADAR2DD(E488Q) negative control conditions. Samples numbers are n= 3 for RCas9-ADAR2DD(E488Q) and n=1 for ADAR2DD(E488Q). D-G, representative Sanger sequencing traces of endogenously expressed mRNAs at targeted adenosine residues in the 3'UTR for genes *ACTB* (d), CDS for *CYFIP2* (g), and 3'UTR for *GUSB* (f) and *GAPDH* (g) in stable, T-REx 293 cells. For *ACTB*, schematics representing both spacer and 3' extension target sequences are displayed. For *CYFIP2*, cartoons illustrating co-transcriptional and post-transcriptional recognition of edit target by ADAR2 and RCas9-ADAR2DD, respectively are depicted as well. Data are mean values ± s.d. with n= 1 - 3; unpaired two-tailed Student's t-test, * $P < 0.05$; * $P < 0.01$; *** $P < 0.001$; n.s. = not significant.

Figure 1.8. Comparative on-target editing of Cas-ADAR2DD system orientations. A, Background subtracted EGFP signal for each of the transiently transfected Cas-ADAR2DD constructs bearing a 2X nuclear localization signal (NLS) tested in both N-TERM ADAR2DD and C-TERM ADAR2DD orientations. B, Background subtracted EGFP signal for each of the transiently transfected Cas-ADAR2DD constructs bearing a HIV1 nuclear export signal (NES) tested in both N-TERM ADAR2DD and C-TERM ADAR2DD orientations. All values were calculated by subtracting the mean percentage value for NT gRNA conditions from that of the EGFP-targeting gRNA (% EGFPtargeting - % EGFPNT) and are represented in arbitrary units (A.U.) ± propagation of error values. C, Relative percentage of EGFP+ cells for targeting gRNAs co-transfected transiently with ADAR2DD(E488Q) control. D, Co-transfection of EGFP-targeting, scaffold-containing gRNAs with increasing sequence length (22nt, 30nt, 50nt) with ADAR2DD (E488Q)-Cas13b and Cas13b-ADAR2DD(E488Q). E, Targeting schematic (top) and relative A-to-I editing efficiencies of target and adjacent adenosine residues in *ACTB* with increasing Cas13b gRNA length, quantified by NGS. F, RNA A-to-I editing rates observed at both target (A43) and adjacent adenosine sites in ACTB target region following co-transfection of 22nt, 30nt, and 50nt gRNAs with ADAR2DD(E488Q)-Cas13b. G, background EGFP+ fluorescence introduced through co-expressing of EGFP-targeting gRNAs with increasing sequence length with ADAR2DD(E488Q) control plasmid. H, background editing observed adenosine sites in ACTB amplicon following transfection of 22nt, 30nt, and 50nt Cas13b gRNAs alone. I, background editing rates observed at both target and adjacent sites in ACTB region following transfection of targeting SpCas9, Cas13b, or Cas13d gRNAs alone. Data represented are mean values ± s.d. with n= 3; unpaired two-tailed Student's t-test, * $P < 0.05$; * $P < 0.01$; *** $P < 0.001$; n.s. = not significant.


A
Cas-ADAR2DD(E488Q)2X NLS
N-terminal
C-terminal
Background subtracted signal (A.U.)
SpCas9+ SpacergRNA
SpCas9+ NS gRNA
SaCas9+ NS gRNA
LwaCas13a
PspCas13b
RfxCas13d
B
Cas-ADAR2DD(E488Q)NES
C
ADARDD(E488Q) + gRNAs
%EGFP+ Cells
SpCas9 NS (21nt)
SpCas9 Spacer (21nt)
SaCas9 NS (21nt)
Cas13a (28nt)
Cas13b (30nt)
Cas13d (22nt)
D
22nt EGFP gRNA (with scaffold)
30nt EGFP gRNA (with scaffold)
50nt EGFP gRNA (with scaffold)
% EGFP Corrected
ADAR2DD(E488Q) -Cas13b
Cas13b -ADAR2DD(E488Q)
E
Cas13b gRNA length
ACTB mRNA (3' UTR)
Cas13b 50nt
Cas13b 30nt
Cas13b 22nt
% A-to-I Editing
Cas13b(E488Q) + 30nt NT gRNA
Cas13b(E488Q) + 22nt gRNA
Cas13b(E488Q) + 30nt gRNA
Cas13b(E488Q) + 50nt gRNA
F
ACTB mRNA (3' UTR)
ADAR2DD(E488Q)-Cas13b + gRNA
22 nt gRNA
30 nt gRNA
50 nt gRNA
% A-to-I Editing
G
ADARDD(E488Q) + gRNAs
% EGFP Corrected
22nt Cas13b gRNA
30nt Cas13b gRNA
50nt Cas13b gRNA
H
Cas13b gRNAs Only
I
Background editing due to gRNA
Cas9 NS gRNA only
Cas9 Spacer gRNA only
Cas13d 22nt gRNA only
Cas13b 22nt gRNA only
Cas13b 30nt gRNA only
Cas13b 50nt gRNA only



		Known Edits			
Sample	**Total Edits**	**Alu**	**Repetitive non-Alu**	**Non-repetitive**	**Novel Edits**
Cas13b + *CYFIP2* gRNA	3,336	397	4	11	2,924
Cas13b + NT gRNA	3,400	274	5	5	3,116
Cas13d + *CYFIP2* gRNA	5,407	779	6	24	4,598
Cas13d + NT gRNA	4,573	571	6	22	3,974
Cas9 + *CYFIP2* (NS) gRNA	5,545	693	8	26	4,818
Cas9 + NT (NS) gRNA	2,131	139	2	2	1,988

Figure 1.9. Nature of transcriptome-wide consequences of transient Cas-ADAR2DD(E488Q) expression. A, Summary of Cas-ADAR2 normalized read counts (TPM) mapped in each sequenced sample. Data represented are mean values ± s.d. with n= 2 B. Table summarizing off-target sites produced by Cas13b-, Cas13d-, and Cas9-ADAR2DD enzymes. Known edit sites previously found in human RNA-seq data were determined using the RADAR RNA editing database, while novel events were characterized as such if not previously reported. C, Venn diagrams comparing shared edits identified between untransfected 293T cells and each Cas protein with either *CYFIP2*-targeting or NT gRNA. D, Pairwise scatterplots illustrating relative per-site editing rates of sites discovered for each Cas protein between CYFIP2-targeting and a scrambled NT gRNA. The target edit site (CYFIP2 CDS) is colored orange and indicated with an arrow. E, Sequence logos of edited adenosine residues identified in either untransfected 293T RNA-seq data (left) or with a scrambled, non-targeting (NT) gRNA and Cas13b, Cas13d, or Cas9. F, Depiction of RNA editing rate per-site (y-axis) versus total read coverage per-site (x-axis).

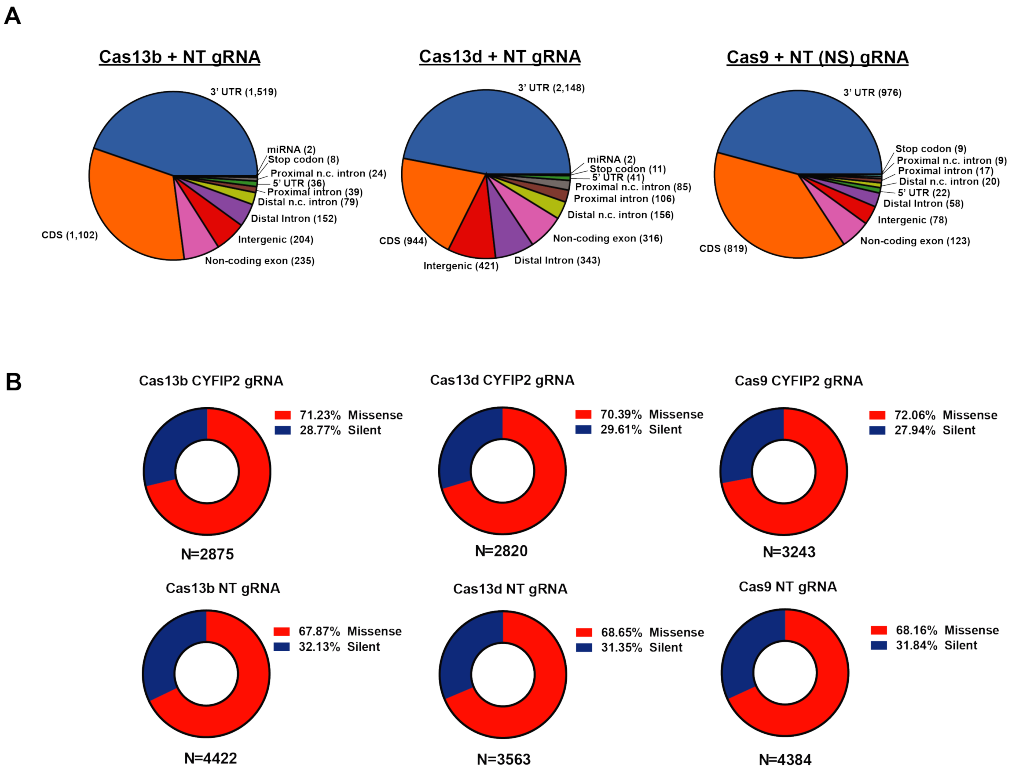


Figure 1.10. Coding sequence consequences of RNA edits with Cas-ADAR2DD(E488Q). A, Regional distribution of total edits per Cas condition with a non-targeting gRNA. As seen with the targeting gRNA, the majority of off-target edits here are localized to 3'UTR and CDS or genes. B, All potential functional consequences of editing events falling within annotated coding sequences of genes (CDS). Total synonymous (silent) and non-synonymous (missense) consequences were predicted and reported using the SnpEff variant predictor with GRCh37 (hg19) ENSEMBL transcript annotations.

References

Abudayyeh, O. O., Gootenberg, J. S., Essletzbichler, P., Han, S., Joung, J., Belanto, J. J., Verdine, V., Cox, D. B. T., Kellner, M. J., Regev, A., Lander, E. S., Voytas, D. F., Ting, A. Y. & Zhang, F. 2017. RNA targeting with CRISPR-Cas13. Nature, 550, 280-284.

Abudayyeh, O. O., Gootenberg, J. S., Franklin, B., Koob, J., Kellner, M. J., Ladha, A., Joung, J., Kirchgatterer, P., Cox, D. B. T. & Zhang, F. 2019. A cytosine deaminase for programmable single-base RNA editing. Science, 365, 382-386.

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B., Shmakov, S., Makarova, K. S., Semenova, E., Minakhin, L., Severinov, K., Regev, A., Lander, E. S., Koonin, E. V. & Zhang, F. 2016. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science, 353, aaf5573.

Adli, M. 2018. The CRISPR tool kit for genome editing and beyond. Nat Commun, 9, 1911.

Azad, M. T. A., Bhakta, S. & Tsukahara, T. 2017. Site-directed RNA editing by adenosine deaminase acting on RNA for correction of the genetic code in gene therapy. Gene Ther, 24, 779-786.

Batra, R., Nelles, D. A., Pirie, E., Blue, S. M., Marina, R. J., Wang, H., Chaim, I. A., Thomas, J. D., Zhang, N., Nguyen, V., Aigner, S., Markmiller, S., Xia, G., Corbett, K. D., Swanson, M. S. & Yeo, G. W. 2017. Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. Cell, 170, 899-912 e10.

Batra, R., Nelles, D. A., Roth, D. M., Krach, F., Nutter, C. A., Tadokoro, T., Thomas, J. D., Sznajder, L. J., Blue, S. M., Gutierrez, H. L., Liu, P., Aigner, S., Platoshyn, O., Miyanohara, A., Marsala, M., Swanson, M. S. & Yeo, G. W. 2020. The sustained expression of Cas9 targeting toxic RNAs reverses disease phenotypes in mouse models of myotonic dystrophy type 1. Nat Biomed Eng.

Cingolani, P., Platts, A., Wang Le, L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X. & Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin), 6, 80-92.

Cox, D. B., Platt, R. J. & Zhang, F. 2015. Therapeutic genome editing: prospects and challenges. Nat Med, 21, 121-31. Cox, D. B. T., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J. & Zhang, F. 2017. RNA editing with CRISPR-Cas13. Science, 358, 1019-1027.

Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. 2004. WebLogo: a sequence logo generator. Genome Res, 14, 1188-90.

Dale, R. K., Pedersen, B. S. & Quinlan, A. R. 2011. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics, 27, 3423-4.

Deffit, S. N., Yee, B. A., Manning, A. C., Rajendren, S., Vadlamani, P., Wheeler, E. C., Domissy, A., Washburn, M. C., Yeo, G. W. & Hundley, H. A. 2017. The C. elegans neural editome reveals an ADAR target mRNA required for proper chemotaxis. Elife, 6.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29, 15-21.

Dugar, G., Leenay, R. T., Eisenbart, S. K., Bischler, T., Aul, B. U., Beisel, C. L. & Sharma, C. M. 2018. CRISPR RNA-Dependent Binding and Cleavage of Endogenous RNAs by the Campylobacter jejuni Cas9. Mol Cell, 69, 893-905 e7.

Gaudelli, N. M., Komor, A. C., Rees, H. A., Packer, M. S., Badran, A. H., Bryson, D. I. & Liu, D. R. 2017. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. Nature, 551, 464-471.

Hanswillemenke, A., Kuzdere, T., Vogel, P., Jekely, G. & Stafforst, T. 2015. Site-Directed RNA Editing in Vivo Can Be Triggered by the Light-Driven Assembly of an Artificial Riboprotein. J Am Chem Soc, 137, 15875-81.

Hsu, P. D., Lander, E. S. & Zhang, F. 2014. Development and applications of CRISPR-Cas9 for genome engineering. Cell, 157, 1262-78.
Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, 337, 816-21.

Katrekar, D., Chen, G., Meluzzi, D., Ganesh, A., Worlikar, A., Shih, Y. R., Varghese, S. & Mali, P. 2019. In vivo RNA editing of point mutations via RNA-guided adenosine deaminases. Nat Methods, 16, 239-242.

Keryer-Bibens, C., Barreau, C. & Osborne, H. B. 2008. Tethering of proteins to RNAs by bacteriophage proteins. Biol Cell, 100, 125-38.
Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S. & Gabriel, A. 2004. Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res, 14, 1719-25.

Kim, E., Koo, T., Park, S. W., Kim, D., Kim, K., Cho, H. Y., Song, D. W., Lee, K. J., Jung, M. H., Kim, S., Kim, J. H., Kim, J. H. & Kim, J. S. 2017a. In vivo genome editing with a small Cas9 orthologue derived from Campylobacter jejuni. Nat Commun, 8, 14500.

Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T. & Liu, D. R. 2017b. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine

deaminase fusions. Nat Biotechnol, 35, 371-376.

Konermann, S., Lotfy, P., Brideau, N. J., Oki, J., Shokhirev, M. N. & Hsu, P. D. 2018. Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. Cell, 173, 665-676 e14.

Kosicki, M., Tomberg, K. & Bradley, A. 2018. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol, 36, 765-771.

Levanon, E. Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z. Y., Shoshan, A., Pollock, S. R., Sztybel, D., Olshansky, M., Rechavi, G. & Jantsch, M. F. 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat Biotechnol, 22, 1001-5.

Li, H. & Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Genome Project Data Processing, S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078-9.

Liao, Y., Smyth, G. K. & Shi, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30, 923-30.

Licht, K., Kapoor, U., Mayrhofer, E. & Jantsch, M. F. 2016. Adenosine to Inosine editing frequency controlled by splicing efficiency. Nucleic Acids Res, 44, 6398-408.

Liu, X. M., Zhou, J., Mao, Y., Ji, Q. & Qian, S. B. 2019. Programmable RNA N(6)-methyladenosine editing by CRISPR-Cas9 conjugates. Nat Chem Biol, 15, 865-871.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., Dicarlo, J. E., Norville, J. E. & Church, G. M. 2013. RNA-guided human genome engineering via Cas9. Science, 339, 823-6.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011, 17, 3.

Merkle, T., Merz, S., Reautschnig, P., Blaha, A., Li, Q., Vogel, P., Wettengel, J., Li, J. B. & Stafforst, T. 2019. Precise RNA editing by recruiting endogenous ADARs with antisense oligonucleotides. Nat Biotechnol, 37, 133-138.

Montiel-Gonzalez, M. F., Diaz Quiroz, J. F. & Rosenthal, J. J. C. 2019. Current strategies for Site-Directed RNA Editing using ADARs. Methods, 156, 16-24.

Montiel-Gonzalez, M. F., Vallecillo-Viejo, I. C. & Rosenthal, J. J. 2016. An efficient system for

selectively altering genetic information within mRNAs. Nucleic Acids Res, 44, e157.

Nelles, D. A., Fang, M. Y., O'connell, M. R., Xu, J. L., Markmiller, S. J., Doudna, J. A. & Yeo, G. W. 2016. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. Cell, 165, 488-96.

Nishikura, K. 2010. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem, 79, 321-49.

O'connell, M. R., Oakes, B. L., Sternberg, S. H., East-Seletsky, A., Kaplan, M. & Doudna, J. A. 2014. Programmable RNA recognition and cleavage by CRISPR/Cas9. Nature, 516, 263-6.

Phelps, K. J., Tran, K., Eifler, T., Erickson, A. I., Fisher, A. J. & Beal, P. A. 2015. Recognition of duplex RNA by the deaminase domain of the RNA editing enzyme ADAR2. Nucleic Acids Res, 43, 1123-32.

Qu, L., Yi, Z., Zhu, S., Wang, C., Cao, Z., Zhou, Z., Yuan, P., Yu, Y., Tian, F., Liu, Z., Bao, Y., Zhao, Y. & Wei, W. 2019. Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. Nat Biotechnol.

Quinlan, A. R. & Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26, 841-2.

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., Zetsche, B., Shalem, O., Wu, X., Makarova, K. S., Koonin, E. V., Sharp, P. A. & Zhang, F. 2015. In vivo genome editing using Staphylococcus aureus Cas9. Nature, 520, 186-91.

Ran, F. A., Hsu, P. D., Wright, J., Agarwala, V., Scott, D. A. & Zhang, F. 2013. Genome engineering using the CRISPR-Cas9 system. Nat Protoc, 8, 2281-2308.

Rauch, S., He, E., Srienc, M., Zhou, H., Zhang, Z. & Dickinson, B. C. 2019. Programmable RNA-Guided RNA Effector Proteins Built from Human Parts. Cell, 178, 122-134 e12.

Rousseau, B. A., Hou, Z., Gramelspacher, M. J. & Zhang, Y. 2018. Programmable RNA Cleavage and Recognition by a Natural CRISPR-Cas9 System from Neisseria meningitidis. Mol Cell, 69, 906-914 e4.

Stafforst, T. & Schneider, M. F. 2012. An RNA-deaminase conjugate selectively repairs point mutations. Angew Chem Int Ed Engl, 51, 11166-9.

Strutt, S. C., Torrez, R. M., Kaya, E., Negrete, O. A. & Doudna, J. A. 2018. RNA-dependent RNA targeting by CRISPR-Cas9. Elife, 7.

Tan, M. H., Li, Q., Shanmugam, R., Piskol, R., Kohler, J., Young, A. N., Liu, K. I., Zhang, R., Ramaswami, G., Ariyoshi, K., Gupte, A., Keegan, L. P., George, C. X., Ramu, A., Huang, N.,

Pollina, E. A., Leeman, D. S., Rustighi, A., Goh, Y. P. S., GTEx Consortium, Chawla, A., Del Sal, G., Peltz, G., Brunet, A., Conrad, D. F., Samuel, C. E., O'connell, M. A., Walkley, C. R., Nishikura, K. & Li, J. B. 2017. Dynamic landscape and regulation of RNA editing in mammals. Nature, 550, 249-254.

Thuronyi, B. W., Koblan, L. W., Levy, J. M., Yeh, W. H., Zheng, C., Newby, G. A., Wilson, C., Bhaumik, M., Shubina-Oleinik, O., Holt, J. R. & Liu, D. R. 2019. Continuous evolution of base editors with expanded target compatibility and improved activity. Nat Biotechnol, 37, 1070-1079,

Vallecillo-Viejo, I. C., Liscovitch-Brauer, N., Montiel-Gonzalez, M. F., Eisenberg, E. & Rosenthal, J. J. C. 2018. Abundant off-target edits from site-directed RNA editing can be reduced by nuclear localization of the editing enzyme. RNA Biol, 15, 104-114.

Vogel, P., Moschref, M., Li, Q., Merkle, T., Selvasaravanan, K. D., Li, J. B. & Stafforst, T. 2018. Efficient and precise editing of endogenous transcripts with SNAP-tagged ADARs. Nat Methods, 15, 535-538.

Vogel, P. & Stafforst, T. 2019. Critical review on engineering deaminases for site-directed RNA editing. Curr Opin Biotechnol, 55, 74-80.

Washburn, M. C., Kakaradov, B., Sundararaman, B., Wheeler, E., Hoon, S., Yeo, G. W. & Hundley, H. A. 2014. The dsRBP and inactive editor ADR-1 utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome. Cell Rep, 6, 599-607.

Wettengel, J., Reautschnig, P., Geisler, S., Kahle, P. J. & Stafforst, T. 2017. Harnessing human ADAR2 for RNA repair - Recoding a PINK1 mutation rescues mitophagy. Nucleic Acids Res, 45, 2797-2808.

Wong, S. K., Sato, S. & Lazinski, D. W. 2001. Substrate recognition by ADAR1 and ADAR2. RNA, 7, 846-58.

Woolf, T. M., Chase, J. M. & Stinchcomb, D. T. 1995. Toward the therapeutic editing of mutated RNA sequences. Proc Natl Acad Sci U S A, 92, 8298-302.

Wright, A. V., Sternberg, S. H., Taylor, D. W., Staahl, B. T., Bardales, J. A., Kornfeld, J. E. & Doudna, J. A. 2015. Rational design of a split-Cas9 enzyme complex. Proc Natl Acad Sci U S A, 112, 2984-9.

Yang, L. Z., Wang, Y., Li, S. Q., Yao, R. W., Luan, P. F., Wu, H., Carmichael, G. G. & Chen, L. L. 2019. Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems. Mol Cell, 76, 981-997 e7.

Chapter 2

RNA editing enzymes for discovery of RNA targets of RNA binding proteins and ribosomes

2.1 Abstract

RBPs serve as crucial regulators of gene expression through acting on all stages of RNA metabolism. To study these functional consequences of these dynamic interactions, researchers have previously been reliant on cross-linking and immunoprecipitation (CLIP)-based technologies to determine direct binding target transcripts. However, these protocols are often technically challenging, require protein-specific optimization, and can only be quantified from bulk sample. To address this, we harnessed the RNA editing enzyme APOBEC1 to identify global RNA-RBP interactions using RNA-sequencing. This simplified technique, which refer to as STAMP (Surveying Targets by APOBEC Mediated Profiling) allows for RBP-specific interaction information using RNA editing signatures alone. In addition to recapitulating known RBP target sequences. STAMP can be coupled with single-cell and long-read sequencing techniques to obtain cell type- and isoform-specific binding information. Additionally, APOBEC1 can be fused to ribosomal proteins to measure transcriptome-wide ribosome associations in cells independent of ribosome footprinting techniques. Altogether, STAMP provides a new means to probe the RNA-RBP interaction landscape in much more specific cellular contexts than

previously described.

2.2 Introduction

RBPs interact with RNA molecules from nascent transcription to decay, and play critical roles in mediating processing, stability, modification, translation, and subcellular localization (Gerstberger et al., 2014, Hentze et al., 2018). Due to their important role in mediating functional gene expression, RBPs are often implicated in several debilitating human diseases making high throughput techniques to investigate the nature of these relationships a necessity (Nussbacher et al., 2015, Gebauer et al., 2021). For years, the gold standard for identification of RBP targets across the transcriptome at single-nucleotide resolution has involved coupling high-throughput sequencing with antibody-based techniques UV-mediated crosslinking and immunoprecipitation (CLIP), whereby the protein of interest is pulled down using an antigen-specific antibody (Ramanathan et al., 2019). Although significant improvements have been made to maximize the amount of information that can be obtained from a single CLIP-seq experiment (Van Nostrand et al., 2016, Porter et al., 2021), it has still proven to be quite labor-intensive, usually requires a sizable amount of input material optimized RNA fragmentation conditions using RNase enzymes to obtain high-resolution target identification. At the same time, while there has been rapid progress in single-cell measurements of chromatin accessibility (Buenrostro et al., 2013), gene expression (Hwang et al., 2018, Tang et al., 2009), and surface protein-levels (Stoeckius et al., 2017, Shahi et al., 2017), there are currently no equivalent techniques for measuring RBP- RNA interactions at single-cell resolution.

Recently, RNA base editors have been used to great effect as tools of discovery and have demonstrated a proof-of-principle that immunoprecipitation (IP)-based techniques can be obviated for functional characterization of RNA-RBP interactions. The Target of RNA-binding proteins Identified by Editing (TRIBE) approach was the first to do so, fusing RBPs of interest to the deaminase domain from the ADAR family of RNA-editing enzymes to mark

target genes with A-to-I edit signatures (McMahon et al., 2016, Rahman et al., 2018). Although this technique can identify known RBP target genes both *in vitro* and *in vivo*, the number of quantifiable edits per gene remain quite low due to ADAR-family proteins' preference for double-stranded RNA substrates. To broaden this context of RNA target recognition, researchers have substituted ADAR proteins for APOBEC1, which is a cytidine deaminating enzyme that catalyzes the conversion of cytosine-to-uracil (C-to-U) in single-stranded RNA substrates. By fusing the N^6-Methyladenosine (m6A)-recognizing YTH-domain to APOBEC1, researchers have been able to successfully use a technique called DART-seq (Deamination Adjacent to RNA modification Target) to identify global m6A modifications both in bulk (Meyer, 2019) and single-cell (Tegowski et al., 2022) contexts. We reasoned that this editing-based strategy would also work for other RNA-recognizing moieties to allow for the IP-free identification of RBP targets across functional RNA-interaction categories using extremely low or single-cell input.

To demonstrate the power of this APOBEC-mediated approach, we developed and experimental and analytical framework called STAMP (Surveying Targets by APOBEC Mediated Profiling), which expands the capabilities of DART-seq to include broader classes of RNA recognizing entities like RBPs and ribosomes. Using STAMP, we can reproduce specific and well-characterized RNA-binding events from editing signatures alone. We also demonstrate the ability to track RNA-RBP interactions at both isoform-specific and single-cell resolution. Lastly, by coupling STAMP to specific ribosome subunits, we can extend this approach to both bulk and single-cell level detection of ribosome association. Together, these establish a novel framework to analyze an additional level of gene regulation at unprecedented levels of molecular precision.

2.3 Results

2.3.1 STAMP enables RBP binding site discovery without immunoprecipitation

To identify target transcripts of RBPs of interest, we fused the coding sequence of the cytidine deaminase enzyme APOBEC1, which is known to catalyze C-to-U base conversion of single-stranded RNA targets in mammalian cells (Rosenberg et al., 2011) (Figure 2.1A). Expression of our RBP-APOBEC1 fusion (RBP-STAMP, or STAMP) would result in the targeted deamination of cytosine bases proximal to the RBPs binding site, which would result in RNA-specific edits that could be identifiable through RNA-seq. To quantify confident edit sites across the transcriptome, we would then use the SAILOR analysis pipeline (Deffit et al., 2017) which has been modified to detect and assign confidences value for each C-to-U mismatch identified using a beta distribution that factors both site coverage and edit percentage after filtering out annotated SNPs.

To test the efficacy of our STAMP system, we profiled the RBFOX2 protein, which is a well-characterized RBP both with respect to sequence and target-specificity (Lovci et al., 2013, Yeo et al., 2009, Ponthier et al., 2006). To control expression, we generated stable, doxycycline-inducible HEK293T cell lines using lentiviral packaging and transduction; this system would allow for both temporal and dose-responsive control over our STAMP protein expression. Cells expressing low (50ng/ml doxycycline) and high (1μg/ml doxycycline) levels of RBFOX2-STAMP for 72 hours prior to harvesting had enriched C-to-U edit clusters on the 3' untranslated region (3'UTR) of the known RBFOX2 target *APP* mRNA, with edit clusters overlapping with reproducible RBFOX2 binding sites as detected by enhanced CLIP (eCLIP) of endogenous RBFOX2 (Van Nostrand et al., 2017). We also found this to be the case for the RBFOX2-APOBEC1 fusion, leading us to believe that APOBEC1 does not disrupt the binding capacity of RBFOX2 (Figure 2.1B). This signal was noticeably elevated over uninduced RBFOX2-STAMP (0ng/ml doxycycline), or control-STAMP (APOBEC1 only) at low and high

induction. Additionally, local edits were 10-fold and 25-fold more frequent than background control-STAMP edits with increasing levels of confidence cutoffs, using SAILOR scores of 0.9 and 0.999, respectively (Figure 2.1C). This together demonstrates the ability of an RBFOX2-APOBEC1 fusion to enrich for target-specific edits upon expression.

We next wished to evaluate the reproducibility of STAMP by quantifying editing levels at low and high expression. The number of edited reads per site mapping to each target gene, when normalized to total read depth and gene length (EPKM, Edited-reads Per Kilobase of transcript, per Million mapped reads), were highly reproducible, with correlations between replicates improving upon induction (Pearson R^2=0.32 at no dox treatment, to R^2=0.72 and 0.83 at low and high dox, respectively; Figure 2.2A). Irreproducible discovery rate (IDR) analysis (Li et al., 2011) around edit sites also revealed reproducible windows for RBFOX2-STAMP, and the number of these reproducible edits also increased with dox induction of RBFOX2-STAMP (figure 2.8A). At the same time, analysis of differentially expressed genes failed to detect significant gene expression changes consistent between replicates at 72 hours post induction, the latest time point assayed (figure 2.8B).

We next measured the proximity of detected edits to well conserved RBFOX2 binding sites. For the 2,852 endogenous RBFOX2 eCLIP that contain the canonical 'UGCAUG' 6-mer sequence, we determined the overall density of edits 200bp upstream and downstream of the motif sequence for both RBFOX2-STAMP and control-STAMP (background). We observed enriched edits for RBFOX2-STAMP within these windows, compared to edits from control-STAMP around the same motifs, and the proximity of edits to motifs was related to with eCLIP peak fold enrichment over size-matched input control (Figure 2.2B). This indicates that RBFOX2 RNA-binding activity is directing and enriching RBFOX2-STAMP specific edits at previously characterized sites.

To improve the meaningfulness of our cumulative editing quantifications, we developed a new set of criteria to retrieve high-confidence edit clusters for RBP-STAMP analogous to peak-calling in CLIP-seq datasets. This new algorithm (as developed by E. Kofman) attempts to

minimize the bias of background edits being called from highly expressed genes by implementing gene-specific thresholds that assumes Poisson-distributed edit scores (ε, as developed by B. Yee and described in the Methods and Materials section), calculated on a per-site basis. Sites that satisfied gene-specific ε thresholds ($p < 0.05$ with adjusted Bonferroni correction) and initial SAILOR confidence score thresholds were then merged with neighboring sites. Instances of edit sites with no neighboring edits within 100 bases in either direction were removed (workflow schematized in figure 2.8C). These additional criteria determined 5,044 edit-clusters across 21,389 for RBFOX2-STAMP (5.4% of the original unfiltered windows) while removing essentially all sites in the ¬control-STAMP sample (21 remaining, 0.04% of unfiltered windows) (figure 2.8D). In comparing these edit clusters to eCLIP peaks, we find that half of significant peaks called ($>$ 4-fold enriched over size-matched input and $p < 0.001$) overlapped with RBFOX2 edit-clusters at a SAILOR confidence threshold of 0.9 for the edit sites; this was more than two-fold higher compared to overlaps with size-matched, randomly shuffled windows selected from exons on the same target genes (Figure 2.2C). Overall, we observed 47.1% of RBFOX2-STAMP edit clusters overlapped with eCLIP peaks regardless of motif, while an additional 8.1% of all edit clusters called contained the canonical motif but did not pass eCLIP thresholding (Figure 2.2D). However, the vast majority of edit clusters were found to be within 1000 base pairs (1kb) of a nearest neighboring eCLIP peak (figure 2.8E), well within the distance reported for both TRIBE (McMahon et al., 2016) and DART-seq (Meyer, 2019). Lastly, we performed *de novo* motif discovery using high confidence edit windows identified with RBFOX2-STAMP. We could recover the canonical RBFOX2 binding 6-mer motif, and enrichment was associated with dosage of gene expression (Figure 2.2E). Comparable results were obtained for another RBP, TIA1, when expressed in our dox-inducible lentiviral system. As with RBFOX2-STAMP, we saw that the number of TIA1-STAMP edits on target genes increased with doxycycline concentration and were strongly correlated across replicates, with summary IDR analysis revealing thousands of reproducible edits (figure 2.8F). Using this same analysis pipeline, we were also able to reconstruct the reported TIA-1 binding motif (figure

2.8G). Altogether, this reinforces the versatility of STAMP in identifying true, reproducible targets of RBPs using RNA editing signatures alone.

2.3.2 Ribosome-subunit STAMP (Ribo-STAMP) edits are enriched in highly translated coding sequences

Since actively translating ribosomes are heavily associated with mRNA molecules in cells, we reasoned that APOBEC1 would also be able to edit ribosome-proximal mRNAs if tethered to the site of translation. This paradigm, which we collectively call Ribo-STAMP, should theoretically take advantage of broad incorporation of these ribosomal subunits into the 80S assembled ribosome associated with translation (Figure 2.3A). To do so, we generated HEK293T cell lines expressing APOBEC1-tagged versions of either 40S ribosomal protein 2 (RPS2) or 40S ribosomal protein 2 (RPS3). In both RPS2-STAMP and RPS3-STAMP we observed edit enrichment on protein-coding genes that are highly translated in HEK293T cells, such as *ATP5PB* (Li et al., 2018), especially when compared to control-STAMP expression (Figure 2.3B). These edits were largely exonic, and highly concordant with RPS3 eCLIP signal enrichment over size-matched input control. In comparison, RPS2-STAMP and RPS3-STAMP signal were minimally detected on highly expressed non-coding genes such as the lncRNA *MALAT1* (Figure 2.3C), which still have residual signal in RPS3 eCLIP experiments. Despite the expected breadth of a ribosomal interaction profile, we observed noticeable reproducibility of per-gene normalized edits with strong EPKM reproducibility between replicates (R^2=0.6 to 0.8) versus weaker, though still significant, correlation with no doxycycline (Figure 2.9A-C). Although enriched in coding sequences (CDS) of exons, we also noticed a fair degree of editing transcriptome-wide across 3'UTR sequences of target transcripts, probably due in part to the natural editing preference of the APOBEC1 protein itself (Figure 2.3B, F), comparison of EPKM computed from CDS only compared to EPKM values computed from both CDS and 3'UTR revealed a strong correlation (Figure 2.9D). This indicates that that 3'UTR edits need not be excluded from downstream analyses, and in some instances may provide valuable editing data

for lowly expressed genes with minimal coverage over CDS alone.

To evaluate specificity of Ribo-STAMP, we compared EPKM values to previously published ribosome profiling (Ribo-seq) data (Zhang et al., 2017) from HEK293 cells. While EPKMs for induced control-STAMP and uninduced RPS2-STAMP were poorly correlated with Ribo-seq normalized read counts (RPKM) (Pearson R^2=0.32 and R^2=0.29, respectively, Figure 2.3D) we found an improvement in correlations between EPKM values for RPS2-STAMP and Ribo-seq RPKM values with both low and high dox induction (R^2=0.41 and R^2=0.46, respectively; Figure 2.3E). Meta-gene analysis of RPS2-STAMP edits for the top quartile of ribosome-occupied genes as determined by Ribo-seq revealed enrichment of edits within the CDS compared to control-STAMP background edits and RBFOX2-STAMP edits, which showed a higher degree of 3'UTR enrichment consistent with eCLIP data (Figure 2.3F). To assess the overall responsiveness of Ribo-STAMP, we treated RPS2- and control-STAMP cells with a mammalian target of rapamycin (mTOR) pathway inhibitor Torin-1, an ATP-competitive inhibitor of mTOR kinase which should massively affect global translation through inhibition of the cell's translational machinery (Thoreen et al., 2012). Indeed, 72 hour concurrent treatment of Torin-1 and doxycycline resulted in a global suppression of both CDS and 3'UTR RPS2-STAMP edit deposition when compared to vehicle-treated samples (Figure 2.3G), a response which was not seen in control-STAMP conditions. RPS2-STAMP EPKM values were also significantly reduced upon Torin-1 treatment in the highest quartile of ribosome occupied genes as defined by both Ribo-seq (Q1 p = 1.9 e-147, Wilcoxon rank-sum test), and polysome-seq (Tan et al., 2019) (Q1 p = 7.7 e-108, Wilcoxon rank-sum test), a phenomenon what was not observed for control-STAMP lines (Figure 2.9E). Gene-level comparison of EPKM values for Torin-1 versus vehicle treated RPS2-STAMP on Q1 genes as defined by Ribo-seq revealed a disparity in editing levels with Torin-1 treatment (Figure 2.3H, Figure 2.9F), with no corresponding difference in gene expression (RPKM) (Figure 2.9G). These results demonstrate the potential for Ribo-STAMP to measure dynamic changes in translation.

2.3.3 Long-read STAMP reveals isoform-specific RBP binding

Since the experimental and analytical basis of STAMP rely solely on RNA-seq data, it can reasonably be coupled with other sequencing technologies that also have a basis in sequencing the transcriptome. We wished to see if RBP-STAMP could capture isoform-specific binding events on unique transcripts using long-read sequencing platforms, which would provide an additional level of information that is lost in CLIP-seq due to deliberate RNA fragmentation. To determine how STAMP could enable RNA target detection on full-length mRNA isoforms, we performed a 72-hour stable induction RBFOX2- and control-STAMP at 1 μg/ml doxycycline and directly sequenced cDNA long reads with the Oxford Nanopore Technologies (ONT) and PacBio (PB) sequencing platforms (Jain et al., 2016, Branton et al., 2008, Ardui et al., 2018, Rhoads and Au, 2015). Both approaches resulted in enrichment of detectable edits overlapping eCLIP peaks, and matched bulk sequencing (short read) data previously analyzed (Figure 2.4A). Additionally, both were capable of identifying the 'UGCAUG' RBFOX2 motif as previous (Figure 2.4B), though PacBio data exemplified higher specificity due to lower base-calling error rate when compared to Nanopore sequencing (Fu et al., 2019). We therefore used this platform as our primary method of analysis going forward.

To determine if we could identify isoform-specific binding events from long-read data, we calculated both RBFOX2- and control-STAMP edit fractions separately on the primary and secondary alternative polyadenylated (APA) isoforms identified across all genes (RBFOX2-STAMP n = 1,604, control-STAMP n = 1,878) in our dataset that satisfied a minimal coverage threshold of 10 reads per isoform. We observed differential isoform editing signatures for RBFOX2-STAMP between primary and secondary isoforms, differences that were not apparent in our control-STAMP condition for the same genes (Figure 2.4C, D). These examples of differential editing across isoforms insinuate that RBPs such as RBFOX2 may interact preferentially with the isoforms of different lengths, such as binding to the longer isoform of *FAR1* (Figure 2.4D). This is sensitive information that is presumedly masked in both short-read and eCLIP data, though

long-read STAMP can support observations seen at the single-molecule level. We therefore conclude that STAMP enables isoform-aware detection of RBP-RNA interactions through pairing with long-read sequencing.

2.3.4 Detection of STAMP targets at single-cell resolution

Although performed at much more granular levels of cellular resolution, single cell RNA-sequencing is also based on the principle of polyadenylated tail mRNA capture and reverse transcription prior to library preparation and sequencing. We therefor reason that STAMP might be amendable to single-cell sequencing to discover RNA-RBP interactions in individual cells. To do so, we modified our lentiviral vectors to enable capture and identification of specific RBP-STAMP proteins using 10x Genomics' Single Cell 3' v3 beads and subsequently performed 72-hour stable expression of RBFOX2- and control-STAMP using 1 μg/ml doxycycline in HEK293T cells prior to single-cell RNA sequencing library preparation. Using the modified RNA-encoded capture-sequence adjacent to the RBP open-reading frame, we could identify "capture cells" for each STAMP background. Doing so, we identified 844 individual RBFOX2-STAMP cells and 5,242 control-STAMP cells from this initial experiment.

Looking more closely at the top 200 genes in our single-cell data identified by expression ranked by transcripts-per-million (TPM), we compared editing rates between bulk and cumulative single-cell data for both RBFOX2- and control-STAMP. This yielded nearly identical edit enrichment of RBFOX2-STAMP samples in both datasets above controls and demonstrated how per gene editing signatures could be determined both in aggregate and single-cell datasets (Figure 2.5A). This is more clearly demonstrated for *UQCRH*, a known eCLIP target gene of RBFOX2 that experiences similar editing in both bulk and single-cell RBFOX2-STAMP experiments (Figure 2.5B). Overall, we saw high concordance (80%) in the target genes that contained filtered high-confidence RBFOX2-STAMP edit clusters obtained from single-cell and bulk datasets (Figure 2.10A) and found that 60% of single-cell clusters directly overlapped with bulk RBFOX2-STAMP clusters, while about 70% fell within 400 bp of bulk edit-clusters (Figure

2.5C). These were also in agreement with eCLIP peaks with 73% of single-cell STAMP targets containing significant RBFOX2-APOBEC1 eCLIP peaks ($P<0.001$; Figure 2.5D). As with bulk RBFOX2-STAMP, most (68.7%) single-cell RBFOX2-STAMP edit-clusters overlapped with eCLIP peaks and/or harbored the RBFOX2 binding motifs (Figure 2.5E). For the large number of clusters that did not directly overlap eCLIP peaks, many were still present in target genes generally within 1000 bp of an eCLIP peak (Figure 2.5F). As we also saw with bulk data, single-cell RBFOX2-STAMP eCLIP peak capture rate was associated with target expression level (Figure 2.10B). Interestingly enough, *de novo* motif analysis from edit-clusters by randomly down-sampling the numbers of single cells analyzed was still capable of identifying the canonical 'UGCAUG' motif with significance (Figure 2.5G). Although not reaching statistical significance, we could detect motif presence even to the resolution of one cell, showcasing the potential power of a single-cell STAMP experiment.

While single-plex RBP-STAMP experiments reproduce findings found in bulk data, individual lentiviral barcodes would allow for multiple RBPs to be assayed simultaneously. We also performed a separate 72-hour TIA-1 single-cell experiment at 1 μg/ml doxycycline, mixing and equal number from each cell population prior to library preparation. This yielded a single-cell experiment consisting of 8,117 cells from RBFOX2-, TIA1-, or control-STAMP backgrounds. We then used the protein-specific open reading frame sequences to demultiplex and distinguish individual STAMP backgrounds. Cells harboring protein-specific capture sequences could be clearly distinguished from each other and from control-STAMP cells by UMAP visualization of ε editing scores (Figure 2.6B), while remaining quite difficult to separate based upon gene expression alone (Figure 2.6 A, Figure 2.10E). This is not entirely surprising given the relative homogeneity of HEK293T cells, though was very interesting that RBP-specific editing signatures allowed for clearer STAMP-specific clustering. Due to the sparsity of cells whose identity could clearly be distinguished using capture sequences alone (Figure 2.6A-B, Figure 2.10C), we could use Louvain clustering (implemented by A. Chaim) on corresponding ε score profiles to further delineate an RBFOX2-population (n = 6,003 cells), a TIA1-population (n = 1,841 cells) and a

background-population (n = 6,623 cells) for our downstream analysis (Figure 2.10D) while still overlapping based upon gene signatures (Figure 2.10E). These RBFOX2 cells could still identify the RBFOX2 motif via HOMER, whole background cells failed to do so (Figure 2.10F). By re-clustering alongside control-STAMP cells, we could identify more inclusive clusters based upon similarities in ε score (Figure 2.6C). These new RBFOX2- and TIA1-STAMP clusters displayed distinct editing profiles when compared to control-STAMP, and could identify several differentially edited genes across cell populations using ε score (Figure 2.6D). After identifying 25 genes with the strongest degree of differential editing across cell populations, we were able to further rank cells based on summed ε scores, choosing cells with the most robust editing per gene. We found that the top 5 cells for each RBP displayed edit enrichment on the shared RBFOX2- and TIA1-STAMP target *NPM1*, which was also detected as a TIA1 target by eCLIP and bulk TIA1-STAMP (Figure 2.6E). Edit enrichments for individual cells were specific to TIA1-STAMP on the *BTF3* target gene, and to RBFOX2-STAMP on the *CFL1* target gene (Figure 2.6E), demonstrating that the targets and binding sites of multiplexed RBP-STAMP fusions can be delineated from edit signatures within single-cell experiments.

Lastly, we wished to see if Ribo-STAMP could be used to quantify ribosomal association of mRNA at the single-cell level. We again performed 72-hour induction of RPS2 at 1 μg/ml doxycycline prior to single-cell capture and sequencing. As we had done for our bulk sequencing analysis, we calculated EPKM values for protein-coding genes in both RPS2- and control-STAMP backgrounds on a per-cell basis. EPKM-based UMAP representation (Figure 2.7A) followed by Louvain clustering (Figure 2.7B) revealed a group of RPS2-STAMP (green, RPS2-population) cells that was clearly distinct from a population of background cells that contained a mixture of both control-STAMP and RPS2-STAMP cells (orange, background-population). In this new RPS2-population, we see that EPKM values (from CDS and 3'UTR of each gene) aggregated from the 3,917 single cells identified correlated meaningfully (R^2=0.53) with genome-wide EPKM values from bulk Ribo-STAMP (Figure 2.11A). We also note that like bulk data, EPKM values computed across CDS and 3'UTR correlated very strongly (R^2=0.81) with EPKM values

calculated from CDS sequences alone(Figure 2.11B), therefore we included 3'UTR-derived edit measurements. We also noticed that while RNA abundance in Ribo-STAMP experiments (RPKM) were in relatively good agreement with total input RNA abundance taken from our polysome-seq dataset (R^2=0.54, Figure 2.11C), mRNA edits remained positively correlated with these measurements but to a lesser degree (R^2 =0.32, Figure 2.11D). In contrast, Ribo-STAMP edits remained better correlated with polysome-enriched RNA abundance overall (R^2=0.51, Figure 2.7C), indicating that single-cell Ribo-STAMP, like single-cell RBP-STAMP, recapitulates results from bulk experiments and correlates well with standard measurements from orthogonal bulk approaches.

After performing our initial RPS2-STAMP, we combined Ribo-STAMP with RBP-STAMP to define ribosome association and RBP binding sites in parallel after merging all control-, RBFOX2-, TIA1 and RPS2-STAMP single-cell edits matrices. UMAP visualization of ε scores revealed that control-STAMP cells overlapped with a subpopulation of captured RBFOX2-, TIA1- and RPS2-STAMP cells (Figure 2.11E) insinuating that some cells may have similar levels of background editing. Repeating our Louvain clustering method and projecting onto UMAP space helped us to define four distinct groups of single cells for our analysis: RPS2-population cells (n = 3,621 cells), RBFOX2-population cells (n = 7,000 cells), containing the majority (92%) of RBFOX2 cells identified by capture sequencing, TIA1-population cells (n = 1,312 cells) containing the majority (57%) of TIA1 capture cells, and a background-population (n = 20,655 cells), composed of control-STAMP cells and any cells that overlap spatially with control-STAMP cells (Figure 2.7D, Figure 2.11F). The ε score-derived RPS2-population experienced 90% similarity to our EPKM-derived RPS2-population (Figure 2.11G), leading us to conclude that both metrics were comparable to one another when determining ribosome-associated edit signatures. Additionally, we saw significant, positive correlation of EPKM values calculated from our RPS2-STAMP Louvain cluster and polysome-seq measurements (R^2=0.47, Figure 2.7E). Metagene plotting of edits from these four subgroups for the top quartile of ribosome occupied genes (from Ribo-seq, n = 4,931 genes) demonstrate CDS-specific enrichment for

single-cell RPS2-STAMP edits compared to more 3'UTR-centric enrichment for single-cell RBFOX2- and TIA1-STAMP, though with noticeable UTR editing coverage (Figure 2.7F), in agreement with our results from bulk (Figure 2.3F). Differential ε score analysis showed distinct editing signatures for RPS2-, RBFOX2- and TIA1-population cells compared to the background-population (Figure 2.7G). To illustrate, we found that the top 10 cells ranked by summed ε score exhibited the expected specific editing signatures on *RPL12*, *RPL30*, and *RPL23A* target transcripts, which were strongly associated with RPS2-STAMP (Figure 2.7H). These results highlight the capacity of STAMP to reveal RBP targets and ribosome association in parallel at single-cell resolution.

2.4 Discussion

In this study, we have developed experimental and computational methods called STAMP to assess RNA-RBP interactions in an immunoprecipitation-free manner. By fusing an RBP of interest to the APOBEC1 protein, which mediates C-to-U deamination of interacting transcripts we could reconstruct global RBP-binding events across the transcriptome using simple RNA-seq and mutation quantification using the SAILOR software package. While STAMP was originally designed to investigate RBP-specific target transcripts, we have also developed Ribo-STAMP, which is capable of recognizing ribosome-associated transcripts through APOBEC1-tagging of ribosomal subunits (here, RPS2 and RPS3). Together, this was an extension of previous techniques TRIBE/HyperTRIBE, which uses ADAR proteins to edit the transcriptome, and DART-seq, which uses YTH-domain fused APOBEC1 to identify sites of m6A modification. In doing so, we combined the respective positives of each of these individual techniques (breadth of binding targets; efficiency of RNA editor) while avoiding the drawbacks (requirement of double-stranded substrate; narrow molecular context). This allowed us to successfully map the global interaction landscapes of several RBPs (RBFOX2 and TIA1) as well as ribosomes. Our editing data matches well with previously performed CLIP and ribosome-profiling experiments,

respectively, without the laborious workflow necessary to achieve usable libraries.

Ribo-STAMP itself offers distinct advantages over other ribosome-association techniques. In addition to ease-of-use, Ribo-STAMP uses edited and non-edited reads to reflect ribosome-associated and input gene expression values simultaneously, allowing to account for two levels of gene expression from a single library. We also find that Ribo-STAMP is sensitive to detecting translational perturbations, as we could detect large changes in translational downregulation at the editing level following mTOR pathway inhibition independent of RNA abundance. We envision that these simultaneous RNA and ribosomal readouts will be extremely useful in more complex and heterogeneous cellular contexts to address questions concerning cell identity or disease states at both the level of transcription and translation. To enable dissemination of our single-cell STAMP technologies, we also developed computational methods that demultiplex multiple RBPs by clustering cells using only edit signatures, which we can validate using 10x feature barcoding technology.

When comparing to another analogous technique, TRIBE, we find that STAMP has several distinct advantages. For one, TRIBE generally yields only gene-level target information without binding site resolution, with only one to two edits on average detectable in any given target gene (Nguyen et al., 2020, Xu et al., 2018, Rahman et al., 2018, McMahon et al., 2016). The scarcity of edits is likely due to higher constraints of ADAR target recognition, which prefers double-stranded RNAs with a bulged mismatch adjacent to the target adenosine residue, which occurs relatively infrequently transcriptome-wide (Song et al., 2020, McMahon et al., 2016, Matthews et al., 2016), and may be even less frequent with actively translating ribosomes. APOBEC enzymes, on the other hand, can access cytosines in single-stranded RNA, which constitute 25%-35% of nucleotides in any given mammalian transcript and produce clusters of edits proximal to target sites, which allow for more site-specific characterization such as *de novo* motif discovery. This infrequency and need for scale make applications like single-cell TRIBE infeasible due to the reduced transcriptome-wide coverage seen in single-cell applications. Only with more frequent base editors like APOBEC1 are such experimental paradigms possible.

Although antibody-based methodologies such as CLIP and RIP are staples used to identify RNA binding sites across the transcriptome, our goal was to obviate these pull-down based approaches to lighten the experimental burden. STAMP approach offers several advantages over CLIP, both technical and analytical. CLIP-based protocols tend to be restrictive due to requirements for large amounts of input material, frequently needing millions of cells for a successful experiment. Here we demonstrate that STAMP can be used reliably at single-cell resolution to identify RNA targets, binding sites, and even extract motifs from small amounts of material or number of cells. Additionally, STAMP enables researchers to multiplex identification of RBP binding sites and global measurement of gene expression into one assay due to reliance on only RNA for its readout. CLIP also requires RNA fragmentation to separate bound and unbound molecules, making direct discovery of isoform-dependent events difficult to resolve. Here we show here that STAMP allows long-read assessment to distinguish RBP binding on different transcript isoforms, a first in the field of RNA-RBP mapping. Furthermore, direct RNA sequencing has recently been demonstrated to be RNA-modification sensitive (Lorenz et al., 2020), which opens the possibility of using STAMP to detect modification-sensitive RNA-protein interactions along with binding information.

Although we demonstrate an overall improvement in detecting RBP and ribosomal targets using APOBEC1, our current STAMP system may be subject to off-target consequences due to duration and super-physiological levels of fusion protein expression. False positives of RBP-STAMP may occur from exogenous expression of our STAMP transgenes as well, possible due to promiscuous interactions and turnover rate of individual molecules of mRNA. For example, while we found that while most RBFOX2-STAMP clusters overlapped with eCLIP peaks and target genes, we did note a number of potential non-target genes accumulating detectable edit clusters. Our strategy for Ribo-STAMP to 'mark' the native translatome using RNA editing may also lead to alterations in coding sequence of target mRNA, leading to unintended downstream gene perturbations. Additionally, although a dox-inducible system allows a certain level of control over STAMP expression, the 12-24 hours necessary to detect expression is still somewhat

tantamount to constitutive expression and may minimize our ability to detect rapid changes in translational dynamics which occur on a matter of hours, if not minutes. Post-translational means of tuning STAMP levels, such as tagging with a small-molecule-stabilized destabilization domain (Iwamoto et al., 2010), may allow for more precise expression windows by controlling the rate of STAMP accumulation following transcriptional induction (Figure 2.12A). We have shown the potential efficacy of such as system by tagging our Ribo-STAMP system with an N-terminal ecDHFR destabilization domain, which can only successfully express our fusion protein with co-treatment of doxycycline and the small molecule trimethoprim (TMP; Figure 2.12B, C). Alternative approaches to optimize expression for future studies may include use of a native promoter by knocking in the APOBEC deaminase domain in frame on one target cell allele, or transient transfection of synthetic mRNAs that code for the fusion for immediate translation in the cytoplasm.

Looking forward, we hope to deploy STAMP for the discovery of RBP- and ribosome-associations in multiple cellular and disease-based contexts, as well as engineer newer and more tunable versions of STAMP technology to further optimize the cellular and molecular resolution of our approach to functional genomics. Although it serves as an excellent proof-of-principal, the vast majority of our experiences to date have been in HEK293T cells stably transduced with RBP-or Ribo-STAMP through lentivirus, and we eagerly hope to explore alternative approaches beyond these homogeneous cell lines. We have recently shown Ribo-STAMP's potential to assess alterations in translation in highly heterogeneous cell samples by tracking expression changes in solid tumors following depletion of the m6A-reader protein YTDHF2 (Einstein et al., 2021). We hope to continue this trajectory by applying STAMP to complex cellular architectures, such as the animal and organoid models, allowing us to overcome major technical hurdles in profiling transcriptional and translational landscapes of development and disease phenotypes at the cell type-specific level.

2.5 Methods and Materials

2.5.1 Plasmid Construction

For the generation of stable cell lines, all RBP-STAMP mammalian expression constructs were in one of two lentiviral Gateway (Invitrogen) destination vector backbones: 1) pLIX403_APOBEC_HA_P2A_mRuby or 2) pLIX403_Capture1_APOBEC_HA_P2A_mRuby. pLIX403_APOBEC_HA_P2A_mRuby was cloned by amplification (Cloneamp, Takara Bio) of APOBEC1_HA_P2A cassette after removal of the YTH cassette from APOBEC1-YTH (gift from Kate Meyer) originally cloned from pCMV-BE1 plasmid (a gift from D. Liu; Addgene plasmid no. 73019). APOBEC_HA_P2A was inserted into the pLIX403 inducible lentiviral expression vector adapted from pLIX_403 (deposited by David Root, Addgene plasmid # 41395) to contain TRE-gateway-mRuby and PGK-puro-2A-rtTA upstream of mRuby by Gibson assembly reaction of PCR products (Cloneamp, Takara Bio). pLIX403_Capture1_APOBEC_HA_P2A_mRuby was constructed by insertion of a synthetic gene block (Integrated DNA technologies, IDT) containing 10x Feature Barcode Capture Sequence 1 with Gibson assembly reaction into MluI digested backbone pLIX403_APOBEC_HA_P2A_mRuby in frame and immediately upstream of the APOBEC1 ORF. RBP open reading frames (ORFs) were obtained from human Orfeome 8.1 (2016 release) donor plasmids (pDONR223) when available, or amplified (Cloneamp, Takara Bio) from cDNA obtained by SuperSript III (Invitrogen) RT-PCR of HEK293XT cell purified RNA (Direct-zol, Zymogen) and inserted into pDONR223 by Gateway BP Clonase II reactions (Invitrogen). Donor ORFs were inserted in frame upstream of APOBEC1 or Capture Sequence 1 APOBEC1 by gateway LR Clonase II reactions (Invitrogen). For transient transfections of HEK293T cells, constructs were modified from pCMV BE1-YTH-HA plasmid (a gift from Kate Meyer modified from D. Liu; Addgene plasmid no. 73019; http://n2t.net/addgene:73019) by removal (control-STAMP) or replacement (RBFOX2-STAMP) of YTH cassette with RBFOX2 open reading frame by PCR and Gibson assembly reactions. To make the dual-induction STAMP vector in Figure 2.12, the ecDHFR sequence was ordered in a Gblock (IDT) and cloned into the

pLIX403_RPS2_Capture1_APOBEC_HA_P2A_mRuby plasmid backbone.

2.5.2 Human cell culture conditions and maintenance

All stable STAMP cell lines were generated using human lenti-X HEK293T cells (HEK293XT, Takara Bio) which are derived from transformed female human embryonic kidney tissue. Cells were maintained in DMEM (4.5 g/L D-glucose) supplemented with 10% FBS (Gibco) at 37^{o} C with 5% CO2. Cells were periodically passaged once at 70-90% confluency by dissociating with TrypLE Express Enzyme (Gibco) at a ratio of 1:10. The stable HEK293XT cell lines RBFOX2-STAMP, TIA1-STAMP, SLBP-STAMP, RPS2-STAMP, RPS3-STAMP, and control-STAMP were generated as described in Generation of STAMP stable cell lines section. by transducing 1 million cells with 8μg/ml polybrene and 1ml viral supernatant in DMEM+10%FBS at 37C^{o} for 24 hours, followed by subsequent puromycin resistance selection (2μg/ml).

2.5.3 Generation of stable STAMP cell lines

Lentivirus was packaged using HEK293XT cells seeded approximately 24 hours prior to transfection at 30-40% in antibiotic-free DMEM and incubation at 37^{o} C, 5% CO_2 to 70-90% confluency. One hour prior to transfection DMEM was replaced with OptiMEM media transfection was performed with Lipofectamine 2000 and Plus reagent according to manufacturer's recommendations at a 4:2:3 proportion of lentiviral vector: pMD.2g: psPAX2 packaging plasmids. 6 hours following transfection, media was replaced with fresh DMEM + 10% FBS. 48 hours after media replacement, virus containing media was filtered through a 0.45 μm low protein binding membrane. Filtered viral supernatant was then used directly for line generation by transducing 1 million cells (1 well of 6 well dish) with 8μg/ml polybrene and 1ml viral supernatant in DMEM+10%FBS at 37^{o} C for 24 hours. After 24 hours of viral transduction, cells were split into 2g/L puromycin and selected for 72 hours before passaging for storage and downstream validation and experimentation.

2.5.4 STAMP editing

For stable cell STAMP fusion protein expression cells were induced with 50ng/ml (low) or 1μg/ml (high) doxycycline in DMEM for 24-72 hours, followed by Trizol extraction and Direct-zol miniprep (Zymo Research) column purification in accordance with manufacturer protocol. Uninduced cells of the same genetic background were used as negative controls. For transient transfections, 1 million cells were transfected with 2μg expression construct using Fugene HD (Promega) according to manufacterer's protocol. Upon Agilent TapeStation quantification, 500ng RNA was used as input material to make total RNA-seq libraries with either TruSeq Stranded mRNA Library Prep (Illumina) or KAPA RNA HyperPrep Kit with RiboErase (Roche) following the provided protocols. For mTOR perturbation experiments, cells were treated with 100nM Torin-1 (Cell Signaling) or DMSO vehicle control alongside 1μg/ml doxycycline induction and harvested for RNA after 72 hours 37^{o} C incubation.

For destabilization domain experiments, cells were treated the same as before, but subjected to treatment with 20 μM trimethoprim (TPM, Sigma-Aldrich) either concurrently or 22-23 hours following doxycycline induction (50 and 100 ng/ml). At the time of harvest, cells were resuspended in RIPA lysis buffer and subjected to Western Blot analysis using Anti-HA Tag Antibody (Ms mAb, Covance #MMS-101R; 1:1000) and Anti-GAPDH Antibody (Ms mAb, Abcam ab8245; 1:5000).

2.5.5 eCLIP

All STAMP fusion (RBFOX2-, TIA1-, and SLBP-APOBEC1) eCLIPs were conducted following induction or transient transfections and IP was conducted using Anti-HA tag antibody - ChIP Grade (Abcam, ab9110). eCLIP experiments were performed as previously described in a detailed standard operating procedure (Van Nostrand et al., 2016), which is provided as associated documentation with each eCLIP experiment on the ENCODE portal (https://www.encodeproject.org/documents/fa2a3246-6039-46ba-b960-

17fe06e7876a/@@download/attachment/CLIP_SOP_v1.0.pdf). In brief, 20 million crosslinked cells were lysed and sonicated, followed by treatment with RNase I (Thermo Fisher) to fragment RNA. Antibodies were pre-coupled to species-specific (anti-rabbit IgG) Dynabeads (Thermo Fisher), added to lysate, and incubated overnight at 4^{o} C. Prior to IP washes, 2% of sample was removed to serve as the paired input sample. For IP samples, high- and low-salt washes were performed, after which RNA was dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adaptor was ligated with T4 RNA ligase (NEB). Ten per cent of IP and input samples were run on an analytical PAGE Bis-Tris protein gel, transferred to PVDF membrane, blocked in 5% dry milk in TBST, incubated with the same primary antibody used for IP (typically at 1:4,000 dilution), washed, incubated with secondary HRP-conjugated species-specific TrueBlot antibody (Rockland), and visualized with standard enhanced chemiluminescence imaging to validate successful IP. Ninety per cent of IP and input samples were run on an analytical PAGE Bis-Tris protein gel and transferred to nitrocellulose membranes, after which the region from the protein size to 75 kDa above protein size was excised from the membrane, treated with proteinase K (NEB) to release RNA, and concentrated by column purification (Zymo). Input samples were then dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3' RNA adaptor was ligated with T4 RNA ligase (NEB) to synchronize with IP samples. Reverse transcription was then performed with AffinityScript (Agilent), followed by ExoSAP-IT (Affymetrix) treatment to remove unincorporated primer. RNA was then degraded by alkaline hydrolysis, and a 3' DNA adaptor was ligated with T4 RNA ligase (NEB). qPCR was then used to determine the required amplification, followed by PCR with Q5 (NEB) and gel electrophoresis to size-select the final library. Libraries were sequenced on the HiSeq 2000, 2500, or 4000 platform (Illumina). Each ENCODE eCLIP experiment consisted of IP from two independent biosamples, along with one paired size-matched input (sampled from one of the two IP lysates before IP washes). Reproducible eCLIP peaks were called using the latest release of the core pipeline (https://github.com/yeolab/eclip), followed by a peak merging sub-workflow to identify reproducible peaks (https://github.com/YeoLab/merge_peaks).

2.5.6 RNA-seq

Bulk RNAseq was sequenced single-end 100nt and trimmed using cutadapt (v1.14.0). Trimmed reads were filtered for repeat elements using sequences obtained from RepBase (v18.05) with STAR (2.4.0i). Reads that did not map to repeats were then mapped to the hg19 assembly with STAR, sorted with samtools (v1.5) and quantified against Gencode (v19) annotations using Subread featureCounts (v1.5.3). Genes with zero counts summed across all samples were removed prior to performing differential expression analysis using DESeq2 (v1.26.0) (Love et al., 2014).

Differential Expression (DESeq2-Supplementary table 3): To calculate differential expression from RNA-seq data, we used DESeq2 (v1.26.0), which uses a negative binomial regression model and Bayesian shrinkage estimation dispersions and fold change to estimate differentially expressed genes (Love, et al. 2014). Significance of logarithmic fold changes are determined by a Wald test to approximate p-values, and genes passing an independent filtering step are adjusted for multiple testing using the Benjamini-Hochberg procedure to yield a false discovery rate (FDR). Genes with an FDR less than 0.05 were considered statistically significant.

2.5.7 SAILOR calls for C-to-U edits

Resulting BAM files were each used as inputs to SAILOR (v1.1.0) to determine C>U edit sites across the hg19 assembly. Briefly described, SAILOR filters potential artifacts and known SNPs (dbSNP v147) and returns a set of candidate edit sites evidenced by the number of C>U conversions found among aligned reads. We used an adapted Bayesian "inverse probability model" (Li et al., 2008) to identify high-confidence A-to-I editing sites from the RNA-seq data, where a confidence value based on the number of reads is associated with each predicted site. Sites were transformed into broader "window" by opening a 51-nucleotide window centered on each site.

2.5.8 Edit distribution, EPKM and ε score method details

We describe an "ε score" fraction formula:

$$\sum_{p=1}^{i} \frac{\left(\sum_{cu=0}^{m} Y_c u\right)}{\left(\sum_{c=1}^{n} Y_c\ \right)}$$

where i represents the number of C positions p in a given coordinate window, with $Y_c u$ and Y_c representing the depth of C>U coverage m and total coverage n at each position, respectively, which considers read coverage, edit frequency (ie. how often a C>U conversion is found) and edit potential (ie. how C-rich a given region is). To find the ε score for a given window, we calculated the ratio between the number of (post-SAILOR-filtered) C>U read conversions to the total (post-SAILOR-filtered) coverage across every C found within the window.

To calculate Edits per kilobase of transcript per million mapped reads (EPKM) per gene, we used cumulative edit counts (T coverage over each edit site called) as determined by SAILOR (v1.1.0). We summed region-specific (either CDS or CDS+3'UTR as defined by hg19 v19 Gencode annotations) edit counts per gene and divided by "per million" mapped read counts to either CDS or CDS+3UTR, respectively for all genes with read counts greater than 0 as defined by Subread featureCounts (v1.5.3). We then normalized this number to the length of either the CDS or CDS+3'UTR of each gene in kilobases (kb). To assess the relationship between RPS2-STAMP and mRNA translation, we compared these per gene EPKM values to normalized read units (Reads per kilobase of transcript per million mapped reads, RPKM) for ribosome protected transcripts assessed in Ribo-seq (GSE94460) and polysome-seq (GSE109423). For these analyses we included all genes that with detected read counts in either our RPS2-STAMP or ribosome-occupancy datasets.

2.5.9 Edit-cluster identification and de-noising

De-noising of STAMP edit data was implemented via a combination of filters designed to retain high-confidence STAMP-edited regions, followed by merging the resulting sites into

coherent "peaks." The first filter (Poisson-based filter) models the number of edited Cs relative to total C coverage as a Poisson process. Given that total edit count on any given gene correlates with expression of that gene, a gene-specific background proportion of edited Cs due to off-target effects is also assumed. By dividing the total number of C¿T conversions by the total number of reads at C positions for each gene, a Poisson parameter is established for each gene, representing this background proportion. Then each edit site is individually evaluated by whether its proportion of edited Cs falls enough far to the right on its own gene's Poisson distribution, using a baseline p-value of 0.05 with a Bonferroni correction based on the number of edit sites being evaluated on that gene, with increased stringency achieved by further dividing this per-gene adjusted p-value by a constant factor. The second approach (score-based filter) makes use of the per-site beta-distribution-derived confidence score described earlier, filtering out any edit sites with a score less than 0.999. The final approach (isolated site filter) is based on the observation that STAMP sites overlapping with the most confident eCLIP peaks tend to be found in clusters rather than isolated, and as such any edit sites with zero neighboring sites within 100 bp in either direction are filtered out. STAMP edit-clusters are generated by merging sites found within 100 bp of each other using bedtools. We performed de-novo motif finding using HOMER (v4.9.1).

Peaks exhibiting l2fc$>$2 and l10p$>$3 from C-Terminus RBFOX2-APOBEC fusion eCLIP data were shuffled within the 5'UTR, CDS, and 3'UTR regions of their respective genes, over 40 permutations. These peak permutations were then expanded by 200 bp on each flank and intersected with de-noised STAMP edit-clusters. The same flank-expansion and intersection was conducted for the original experimentally-derived eCLIP peaks. Six different versions of STAMP site "de-noising" are reflected by the x axis labels, where the decimal value reflected the confidence score used for filtering, and the "filtered" suffix reflects application of the additional isolated site filter.

Figure 2.3C: RPS2-STAMP at 0, 50, and 1000ng doxycycline treatments compared to corresponding control-STAMP datasets were compared across genesets taken from GSE94460 and (Zhang et al., 2017). Similar comparisons were done using normalized occupancy ratios from

(Tan et al., 2019) (using X3 values, which closely approximate native ribosome occupancy levels in 293T cells). Figure 2F: Metagene profiles comparing edits (conf $\geq$ 0.5) in RPS2-, RBFOX2-, and control-STAMP were generated using metaPlotR (https://github.com/olarerin/metaPlotR) with the highest occupancy via ribo-seq transcripts from GSE112353, with the top quartile of expressed transcripts being used, although no expression filtering or transcript-to-gene mapping was needed as transcript-level annotations were required (Q1 n=4,677). Figure 2.3D-E: Metagene plots were generated in similar fashion to Figure 3G, comparing all replicates of Torin-1 treated RPS2-STAMP, RPS2-STAMP vehicle treated. Figure 2.3F: From GSE94460, genes were ranked in descending order according to their replicate-averaged TPM-normalized occupancy counts. To consolidate annotations, transcripts that were found with the highest occupancy were kept. Additionally, only genes included in both our analysis (minimally expressed protein coding genes TPM>0 in either RPS2- or control-STAMP, n=16,128) and the GSE112353 dataset (n=19,724) were used. The remaining genes (n=15,485) were quartiled according to occupancy score, such that "quartile 1" represents genes with the highest ribosome occupancy. EPKM values across CDS and 3'UTR exons within these quartiles were compared using a Wilcoxon rank-sum test to determine significance.

2.5.10 Irreproducible Discovery Rate

Irreproducible Discovery Rate (IDR) was employed to determine reproducible edit windows between experimental replicates (Li Q et al., Annals of Applied Statistics. 2011). After pre-filtering SAILOR outputs for a minimum confidence score ($\geq$ 0.5), we created 51nt windows around candidate C>U sites and calculated reproducibility scores for each window using IDR (v2.0.2). Scaled score (-125*log2(IDR_score)) were converted to linear values and plotted, with unscaled scores $\geq$ 0.05 considered as reproducible sites.

2.5.11 RNA Isolation and PolyA selection for Nanopore and PacBio Sequencing

At 80% confluency in 10cm plates, cells were washed with PBS and harvested in 1mL of TRIzol reagent (Thermo Fisher) or Direct-zol kit with DNase treatment (Zymo Research). Total RNA was extracted following the manufacturer's protocol. 20 μg of total RNA was poly-A selected using a poly-A magnetic resin kit (NEB E7490L). RNA was then analyzed by high-sensitivity RNA Tapestation (Agilent #5067-5579) to confirm poly-A selection and RNA quality.

2.5.12 Direct cDNA Nanopore Sequencing

100ng of poly-A selected RNA was used as input for the Nanopore direct cDNA sequencing kit (SQK- DCS109). cDNA was prepared following the manufacturer's protocol. Sequencing was carried out on using Oxford Nanopore PromethION flow cells (FLO- PRO002) for 48 hours. Data was base called in real time on the PromethION Guppy base callers with the high accuracy setting. Total reads (in millions) were: RBFOX2=24.9, APOBEC_control= 8.4.

2.5.13 Nanopore Read Base and Edit Calling

All Nanopore reads were aligned to both hg19 and ENSEMBL's cDNA reference genomes using Minimap2 (Li, 2018) with default RNA parameters. Theses alignments are referred to genomic and cDNA respectively. Edits were called using Bcftools mpileup with settings "-Q 5 -d 8000 -q 1" followed by filtering each position for reference C positions on the appropriate strand. cDNA alignments were assumed to be positive stranded and genome alignments were intersected with gene annotations to determine strand. Sites with ambiguous strand information and/or fewer than 10 reads were removed. Edit fractions were determined for sites with C to U mutations by the fraction (# of mismatches)/(#of mismatches + # of matches). Confidence scores and SNP removal were done via custom implementation of the SAILOR scripts. A final list of RBFOX2 -STAMP sites was made by subtracting all sites found in the control-STAMP

with a confidence score of 0.99 or greater. Isoform specific binding were detected by summing the number of RBFOX2 unique sites and all sites identified in the control-SAMP The top two expressing isoforms, as determined by average coverage across C positions with at least 10 reads, were selected for further analysis and isoforms comparing the largest difference in edits were compared by hand.

2.5.14 Direct cDNA PacBio sequencing

Technical triplicate samples containing 1 μg total RNA were extracted from HEK293T cells expressing control-STAMP and HEK293T, cells expressing the RBFOX2-STAMP fusions, and following 1μg/ml doxycyline induction for 72 hours. RNA extraction was completed using Direct-zol, Zymogen. All STAMP samples were assayed for quality and the all sample RNA integrity numbers (RINs) were greater than 9. Long read cDNA libraries were prepared according to the PacBio Iso-Seq Express protocol with 300 ng of total RNA and amplified for 13-15 cycles with the following forward and reverse primers(Forward: 5'- GGCAATGAAGTCGCAGGGTTG - 3'; Reverse: 5'- AAGCAGTGGTATCAACGCAGAG – 3'). The double stranded cDNA for each sample was converted to sequencing libraries as recommended (PacBio SMRTbell Express Template Prep Kit 2.0) but with separate barcoded adapters for each sample (PacBio Barcoded Overhang Adapter Kit).

All of the samples were pooled in an equimolar fashion and sequenced on a SMRTcell 8M with the PacBio Sequel II instrument (2.0 chemistry/2.1 polymerase with 2 hour pre-extension and 30 hour movie times). After barcodes were demultiplexed, the initial data was used to re-balance the pooling by barcode counts before further sequencing. In total, the samples were sequenced over 5 SMRTcell 8M. The PacBio Sequel II system was used for all sample sequencing.

Following sequencing, the circular consensus sequence (CCS) reads for each set of technical replicates were processed using the Isoseq v3 pipeline (Gordon et al., 2015) (https://github.com/PacificBiosciences/IsoSeq) to generate full-length non-concatamer reads

in fasta format. For this step, software package lima v2.0.0 was used with parameters: –isoseq and –dump-clips. In addition, isoseq3 v3.4.0 refine was used with parameters: –require-polya. Fasta files for each set of technical replicates were the pooled together and the full-length non-concatemer reads for each sample were aligned to the hg19 reference genome using minimap2 v2.17-r941 using parameters: –ax-splice, -uf, –secondary=no, -t 30. Cupcake v18.1.0 (https://github.com/Magdoll/cDNA_Cupcake/wiki) script collapse_isoforms_by_sam.py was then run using the pooled full-length non-concatemer fasta file and aligned SAM file for each sample with parameter –dun-merge-5-shorter to collapse redundant isoforms. This step was completed to collapse high quality isoforms into unique isoforms informed by genome alignment. Following this, SQANTI3 v1.6 (Gordon et al., 2015) (https://github.com/ConesaLab/SQANTI3) script sqanti3_qc.py was used to compare the collapsed isoform results from Cupcake to the Gencode hg19 (v19) annotation in order to characterize the collapsed isoforms.

Edit/C were quantified for each sample using the SAILOR computational tool without filtering reads for RBFOX2-APOBEC1 and APOBEC1-control samples. Edited positions having a confidence score of greater than or equal to 0.99 were then used to elucidate motifs using HOMER tool findMotifsGenome.pl v4.9.1 (Heinz et al., 2010).

A custom script was generated in order to quantify the percent edited reads in the 3' exonal region for each sample. Only previously annotated genes were considered, using the Gencode hg19 (v19) annotation as the reference. For each gene, the isoforms associated with the gene were first determined based on assignment by the SQANTI3 isoform classification pipeline. Only genes with two or more isoforms were considered. Following this, the reads associated with each isoform were determined and categorized using the .group.txt file generated by Cupcake. Samtools v1.9 tool bamtobed was used to generate a BED file based on the aligned reads for each sample. For each sample, start and end coordinates for each read associated with the gene were extracted from the BED file and used to group reads into bins based on the coordinate of the 3' end of each read, applying a leniency of a 10bp window. Only bins corresponding to the dominant 3' exon start site were considered in order to filter for bins that would support

instances of alternative polyadenylation (APA). Edits in a read were counted across the region of a read between the dominant 3' exon start site and the end site corresponding to the respective bin. Edits located at potential SNP positions (positions where $\geq$ 50% of the reads in the bin contained an edit) were not considered. The proportion of reads containing one or more edits within the selected region corresponding to each respective bin was then quantified. Further filtering involved only comparing the two bins with the most reads for each gene and filtering out genes in which the bin with the most reads had more than five times the number of reads in the second bin.

2.5.15 Single cell RNA-seq

For the single cell RNA sequencing of transduced cells. Following 72 hours of doxycycline treatment (1 μg/mL), cells were trypsinized (TrypLE, Invitrogen), counted and resuspended at a density of 1,000 cell/μL in 0.04% BSA in PBS. Single cells were processed through the Chromium Single Cell Gene Expression Solution using the Chromium Single Cell 3' Gel Bead, Chip, 3' Library and 3' Feature Barcode Library Kits v3 (10X Genomics) as per the manufacturer's protocol. Sixteen thousand total cells were added to each channel for a target recovery of 10,000 cells. The cells were then partitioned into Gel Beads in Emulsion in the Chromium instrument, where cell lysis and barcoded reverse transcription of RNA occurred, followed by amplification with the addition of "Feature cDNA Primers 1" (for the mixed RBFOX2:TIA1-STAMP), fragmentation, end-repair, A-tailing and 5' adaptor and sample index attachment as indicated in the manufacturer's protocol for 3' expression capture. 3' feature barcode libraries were prepared as described by the manufacturer's protocol, following cDNA amplification, the Ampure cleanup supernatant was saved, amplified with Feature and Template Switch Oligo primers and finally indexed. Agilent High Sensitivity D5000 ScreenTape Assay (Aglient Technologies) was performed for QC of the libraries. 3' polyA and feature libraries were sequenced on an Illumina NovaSeq 6000. For 3' polyA de-multiplexing, alignment to the hg19 and custom hg19 + lentiviral-genes transcriptomes and unique molecular identifier (UMI)-collapsing were

performed using the Cellranger toolkit (version 2.0.1) provided by 10X Genomics. Cells with at least 50,000 mapped reads per cell were processed. Analysis of output digital gene expression matrices was performed using the Scanpy v1.4.4 package (Wolf et al., 2018). Matrices for all samples were concatenated when necessary and all genes that were not detected in at least 0.1% of single cells were discarded. Cells with fewer than 1,000 or more than 7,000 expressed genes as well as cells with more than 50,000 unique transcripts or 20% mitochondrial expressed genes were removed from the analysis. Transcripts per cell were normalized to 10,000, added a unit and logarithmized ("ln(TPM+1)") and scaled to unit variance (z-scored). Top 2,000 variable genes were identified with the filter_genes_dispersion function, flavor='cell_ranger'. PCA was carried out, and the top 40 principal components were retained. With these principal components, neighborhood graphs were computed with 10 neighbors and standard parameters with the pp.neighbors function. Single cell edits were called by first computing the MD tag from Cellranger outputs (possorted_genome_bam.bam) using Samtools calmd and splitting every read according to their cell barcode. SAILOR, ε score, region EPKM and motif analysis were run for each cell (or the aggregate of reads for all cell barcodes within defined Louvain clusters) in similar fashion to bulk RNAseq. Reads belonging to each cluster of barcodes were combined using a custom script and treated similarly. Analysis of output digital gene edit matrices was performed using the Scanpy v1.4.4 package (Wolf et al., 2018). Matrices for all samples were concatenated and all genes that were not edited in at least 2 single cells were discarded, leaving 1,748, 1,542, 1,949 and 1,862 edited genes for further analyses for HEK293T-control-, HEK293T-RBFOX2:TIA1-, HEK293T-RPS2-STAMP and RPS2-cluster, respectively. Cells with fewer than 10 edited genes were removed from the analysis. EPKM or ε scores for each cell were normalized to 10,000, added a unit and logarithmized ("ln(TPM+1)") and scaled to unit variance (z-scored). PCA was carried out, and the top 40 principal components were retained. With these principal components, neighborhood graphs were computed with 10 neighbors and standard parameters with the pp.neighbors function. Louvain clusters were computed with the tl.louvain function and standard parameters. Following visual inspection, subsets of Louvain clusters were merged guided by

their overlap (or lack thereof) with control-STAMP cells in order to define RBP-specific clusters. Single cell and mean ε scores per sample heatmaps were generated with the pl.heatmap and pl.matrixplot functions, respectively. Differentially edited genes were determined for each set of Louvain (or modified) clusters with the tl.rank_gene_groups function (method='wilcoxon').

Figure 4G: Edits were called from groups of randomly selected RBFOX2-capture sequence barcodes (n=1..49, 100, 200, 300, 400, 500, 600, 700, 800, 844 cells) and processed using SAILOR pipeline. To discover whether or not sites are globally enriched for known binding motifs, we re-calculated the confidence score using the same e score (number of C>U read conversions over the total coverage across all C's within a window) across all 51nt windows surrounding each candidate edit site and filtered these windows using various scores (0.99 and 0.999). We performed de-novo motif finding using HOMER (v4.9.1) using these filtered windows and a shuffled background for each UTR, CDS, intron, and total genic region (findMotifs.pl foreground.fa fasta outloc -nofacts -p 4 -rna -S 20 -len 6 -noconvert -nogo -fasta background.fa), resulting in a set of fasta sequences corresponding to each 51nt edit window as well as a corresponding random background. This was repeated 10 times. HOMER was then run on each of the 580 real/random sequences to find enriched de-novo motifs. The most significant motif that most resembled the canonical RBFOX2 motif (UGCAUG) was then used as a pivot, and significance was re-calcuated for this motif for each foreground/background group and trial (findMotifs.pl foreground.fa fasta output/ -nofacts -p 4 -rna -S 20 -len 6 -noconvert -nogo -known -fasta background.fa -mknown UGCAUG.motif).

2.6 Acknowledgements

Chapter 2, in part, is adapted from material as it appears in Nature Methods 2021. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, Jagannatha P, Dong KD, Madrigal AA, Underwood JG, Yeo GW. Nature Press, 2021. The dissertation author was an investigator and author of this paper.

We thank Dr. Gabriella Viero at the Institute of Biophysics, CNR Unit at Trento for her helpful advice concerning analysis of ribo-seq data. We thank Dr. Todd Michael at the J. Craig Venter Institute for use of the Oxford Nanopore PromethION system. We are grateful to Dr. Kate D. Meyer for her gift of the YTH-APOBEC1 construct. We are grateful to the La Jolla Institute's Immunology Sequencing Core and the IGM Genomics Center, University of California San Diego, for use of the 10X Chromium and Illumina sequencing platforms. This work was partially supported by National Institutes of Health HG004659 and HG009889 to G.W.Y. I.A.C. is a San Diego IRACDA Fellow supported by NIH/NIGMS K12 GM068524 Award. R.J.M. was supported in part an institutional award to the UCSD Genetics Training Program from the National Institute for General Medical Sciences, T32 GM008666 and a Ruth L. Kirschstein National Research Service Award (1-F31-NS111859-01A1). K.W.B is a University of California President's Postdoctoral Fellow supported by NIH/NINDS K22NS112678 K22 Award.

2.7 Figures

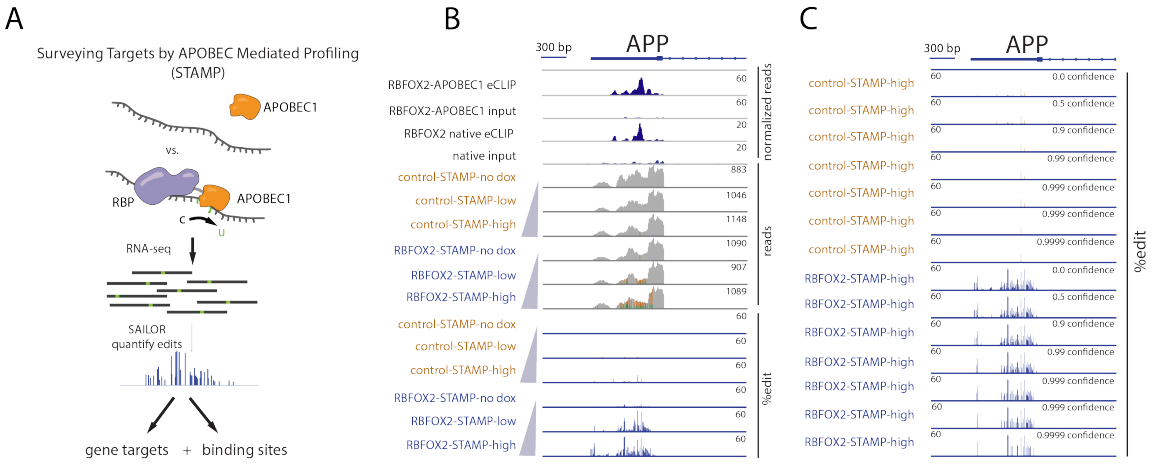
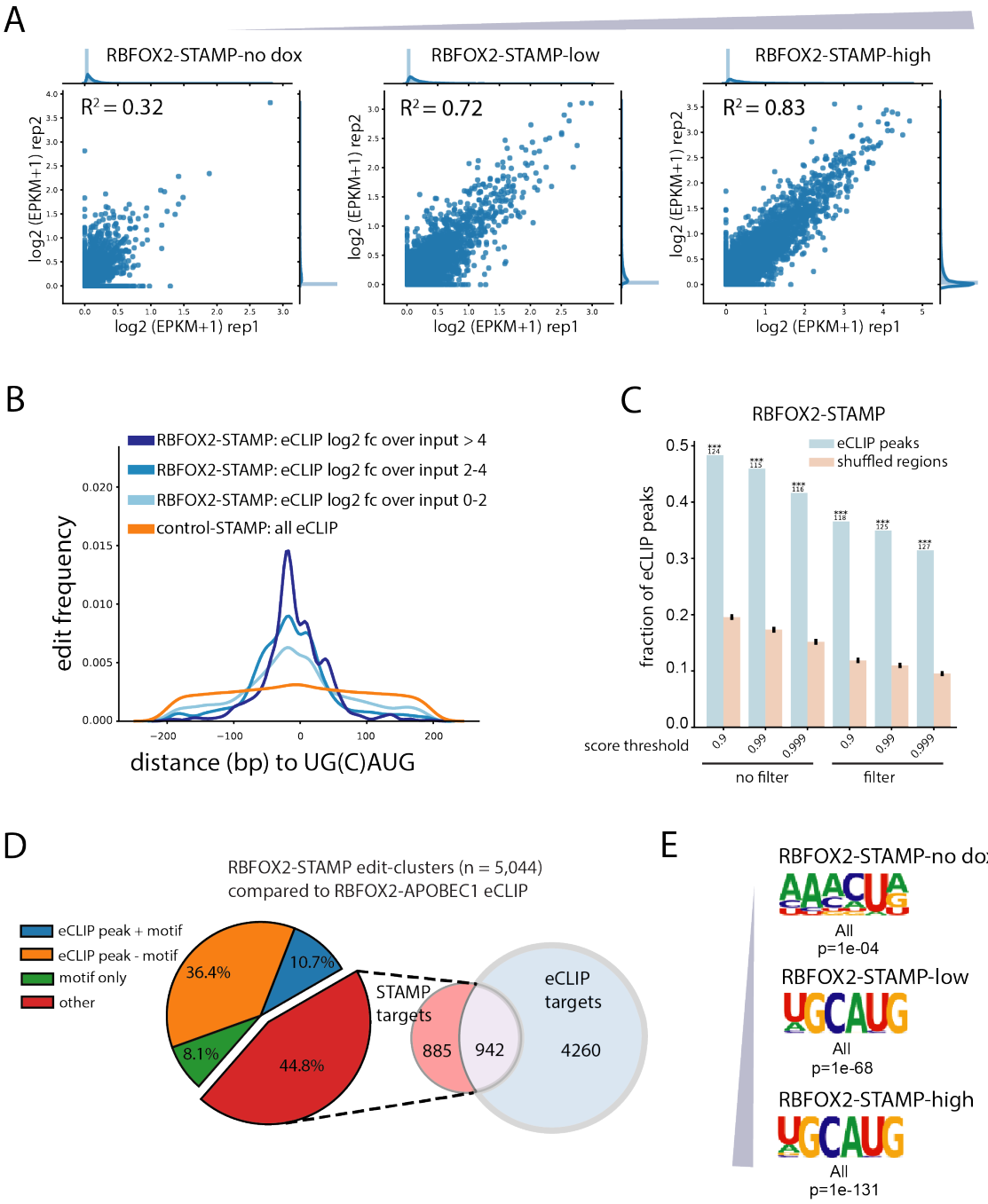


Figure 2.1. RBP-STAMP edits mark specific RBP binding sites. A, Surveying Targets by APOBEC Mediated Profiling (STAMP) strategy fuses rat APOBEC1 module to an RBP of interest to deposit edits at or near RBP binding sites. C-to-U mutations from either APOBEC1-only control (control-STAMP) or RBP fusion (RBP-STAMP) can be detected by standard RNA-sequencing and quantified using our SAILOR analysis pipeline. B, Integrative genome viewer (IGV) browser tracks showing RBFOX2 and RBFOX2-APOBEC1 eCLIP peaks on the target gene APP, compared with control- and RBFOX2-STAMP signal and SAILOR quantified edit fraction for increasing induction levels of fusions (doxycycline: 0ng = none, 50ng = low, or 1μg/ml = high, 72 hours). C, IGV tracks showing 72-hour high-induction control- and RBFOX2-STAMP signal on the APP target gene at increasing confidence levels.

Figure 2.2. RBFOX2-STAMP is reproducible and matches interaction maps via eCLIP
A, RBFOX2-STAMP replicate correlations for the edited read counts per target normalized for length and coverage (EPKM). B, RBFOX2-STAMP and control-STAMP (background) edit frequency distribution within a 400 bp window flanking RBFOX2 eCLIP binding-site motifs, split into increasing levels of log2 fold enrichment of eCLIP peak read-density over size-matched input. C, Fraction of RBFOX2-APOBEC1 eCLIP peaks (log2fc>2 and -log10p>3 over size-matched input) with RBFOX2-STAMP edit-clusters, compared to size-matched shuffled regions, calculated at different edit site confidence levels before and after site filtering (see Materials and Methods for filtering procedure). Numbers atop bars are Z-scores computed comparing observed with the distribution from random shuffles. *** denotes statistical significance at p = 0, one-sided exact permutation test. D, Pie chart showing the proportion of filtered RBFOX2-STAMP edit-clusters overlapping either 1) RBFOX2-APOBEC1 fusion high-confidence eCLIP peaks (log2fc>2 and -log10p>3) containing the conserved RBFOX2 binding motif, 2) equally stringent eCLIP peaks not containing the conserved motif, 3) the conserved motif falling outside of eCLIP peaks, or 4) neither eCLIP peaks or conserved motifs. E, Motif enrichment using HOMER and shuffled background on RBFOX2-STAMP edit-clusters for increasing RBFOX2-STAMP induction levels.


A
RBFOX2-STAMP-no dox
RBFOX2-STAMP-low
RBFOX2-STAMP-high
R² = 0.32
R² = 0.72
R² = 0.83
log2 (EPKM+1) rep2
log2 (EPKM+1) rep1
B
RBFOX2-STAMP: eCLIP log2 fc over input > 4
RBFOX2-STAMP: eCLIP log2 fc over input 2-4
RBFOX2-STAMP: eCLIP log2 fc over input 0-2
control-STAMP: all eCLIP
edit frequency
distance (bp) to UG(C)AUG
C
RBFOX2-STAMP
eCLIP peaks
shuffled regions
fraction of eCLIP peaks
score threshold
no filter
filter
D
RBFOX2-STAMP edit-clusters (n = 5,044)
compared to RBFOX2-APOBEC1 eCLIP
eCLIP peak + motif
eCLIP peak - motif
motif only
other
36.4%
10.7%
8.1%
44.8%
STAMP targets
eCLIP targets
885
942
4260
E
RBFOX2-STAMP-no dox
All
p=1e-04
RBFOX2-STAMP-low
All
p=1e-68
RBFOX2-STAMP-high
All
p=1e-131

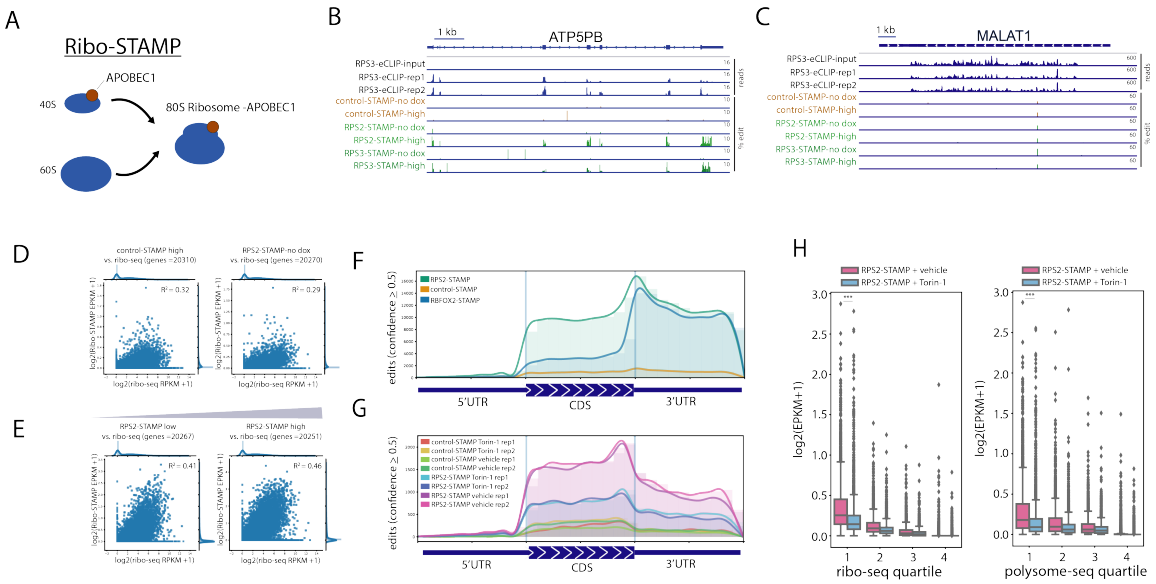


Figure 2.3. Ribo-STAMP edits mark highly translated coding sequences. A, Schematic of RPS-tagging with APOBEC1 and Ribo-STAMP. B, IGV browser tracks displaying coding sequence edit frequency from control, RPS2-STAMP, and RPS3-STAMP at no-induction or 72-hour high-induction on the ATP5BP gene locus, with RPS3 eCLIP and input reads are shown for comparison. C, IGV browser tracks as in A on the noncoding RNA MALAT1, showing no enrichment for RPS3 eCLIP reads, RPS2- or RPS3-STAMP edits. D, Genome-wide scatterplot comparison of control- (left) and RPS2-STAMP EPKM without induction (right), or and E, increasing levels of doxycycline to ribo-seq ribosome protected fragment (RPF) RPKM for increasing levels of RPS2-STAMP. F, Metagene plot showing edit ($\geq$ 0.5 confidence score) distribution for high-induction RPS2-STAMP compared to control-STAMP and RBFOX2-STAMP across 5'UTR, CDS and 3'UTR gene regions for the top quartile (n=4,931) of ribosome occupied genes (ribo-seq). G, Metagene plot as in F showing edit ($\geq$ 0.5 confidence level) distribution for vehicle-treated 72-hour high-induction RPS2-STAMP compared to replicate Torin-1 treated 72-hour high-induction RPS2-STAMP across 5'UTR, CDS and 3'UTR gene regions for the top quartile of ribosome occupied genes. H, Comparison of EPKM from combined replicates (n = 2) vehicle treated 72-hour high-induction RPS2-STAMP compared to Torin-1 treated 72-hour high-induction RPS2-STAMP showing significant signal reduction for top ribosome occupied quartile genes containing Torin-1 sensitive TOP genes as detected by ribo-seq (Q1 p = 1.9 e-147, n = 3589 genes, Wilcoxon rank-sum one-sided) and polysome profiling (Q1 p = 7.7 e-108, n = 3589 genes, Wilcoxon rank-sum one-sided).

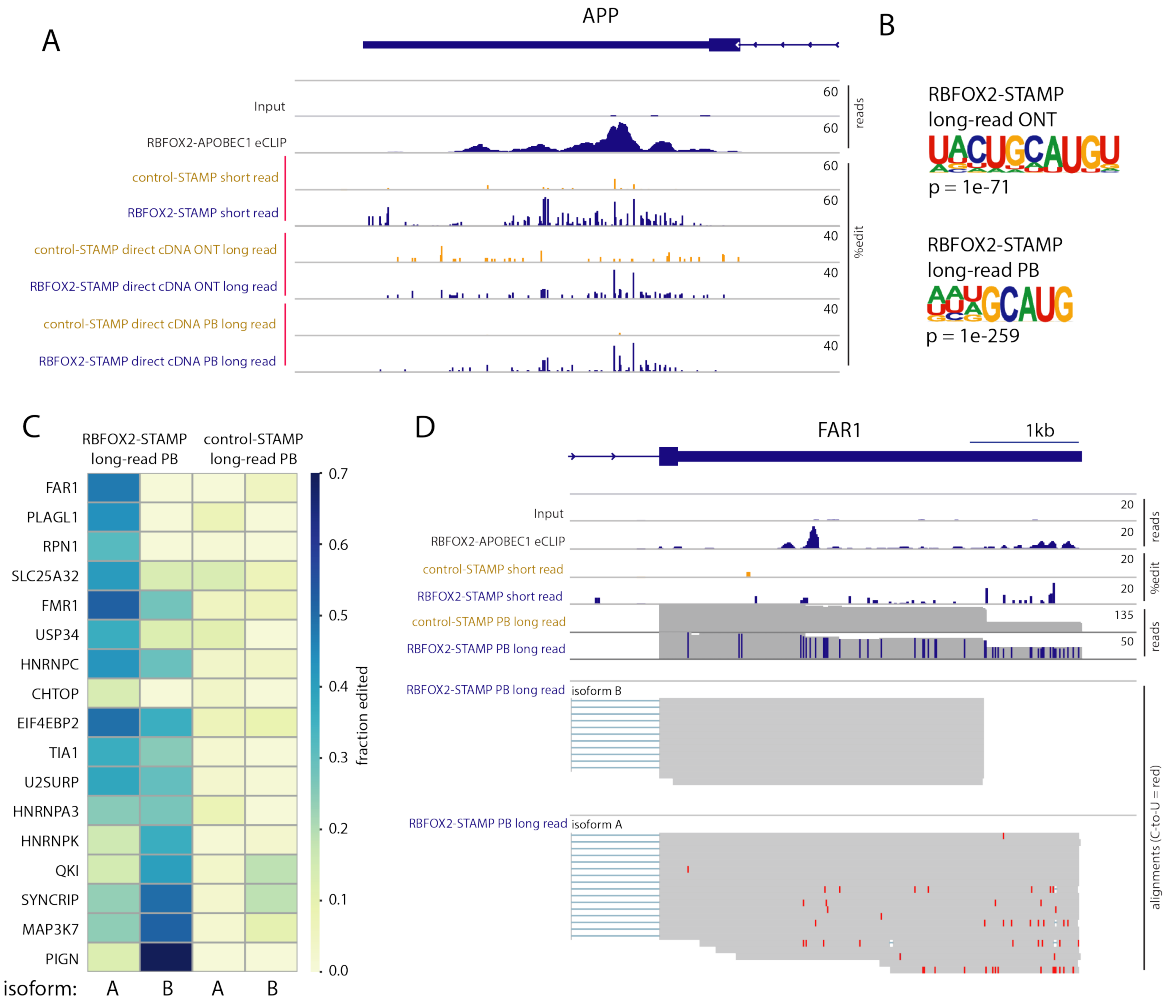


Figure 2.4. Long-read STAMP reveals isoform specific binding profiles. A, IGV tracks showing RBFOX2 eCLIP peak on the target gene APP, compared with 72-hour high-induction control- and RBFOX2-STAMP SAILOR quantified edit fractions for both long-read (Oxford Nanopore Technologies (ONT) or PacBio (PB)) direct cDNA, and short read (NGS) outputs. B, HOMER motif analysis of RBFOX2-STAMP long-reads (ONT and PB) for edits above 0.99 confidence. C, Heatmap of control- and RBFOX2-STAMP edit fractions on the 2 primary alternative polyadenylation (APA) isoforms for the top differentially edited RBFOX2-STAMP APA targets. D, IGV tracks showing RBFOX2-APOBEC1 eCLIP peaks, control- and RBFOX2-STAMP short-read edit frequencies, and control- and RBFOX2-STAMP long-read (PB) alignments on the 2 primary isoforms of the target gene *FAR1*, with red colored C-to-U conversions on different isoforms.

Figure 2.5. STAMP allows RBP binding site detection at single-cell resolution., A, Edit fraction comparison of bulk 72-hour high-induction control- and RBFOX2-STAMP with single-cell control- and RBFOX2-STAMP across the top 200 genes ranked by transcripts per million (TPM) from bulk RBFOX2-STAMP RNA-seq. B, IGV tracks showing the RBFOX2 eCLIP peak on the target gene *UQCRH*, compared with RBFOX2-STAMP edit fractions for the top 10 control- and RBFOX2-STAMP cells ranked by summed ε scores. C, Evaluation of percentage overlap between bulk and single-cell edit-clusters showing that 60-75% of single-cell edit clusters overlap bulk edit clusters over increasing cluster-flanking regions. D, Overlap between RBFOX2-APOBEC1 eCLIP target transcripts (peaks log2fc>2 and -log10p>3 over input) and single-cell RBFOX2-STAMP edit-cluster containing target transcripts. E, Pie chart showing the proportion of single-cell RBFOX2-STAMP edit-clusters overlapping either 1) RBFOX2-APOBEC1 fusion high-confidence eCLIP peaks (log2fc>2 and -log10p>3 over input) containing the conserved RBFOX2 binding motif (GCAUG), 2) equally stringent eCLIP peaks not containing the conserved motif, 3) the conserved motif falling outside of eCLIP peaks, or 4) neither eCLIP peaks nor conserved motifs. F, Cumulative distance measurement from single-cell RBFOX2-STAMP distal edit-clusters to eCLIP peaks on targets genes. G, -log10 of p-values (n = 10 trials) for motifs extracted by HOMER (v4.9.1) using RBFOX2-STAMP $\geq$ 0.99 confidence level edits from randomly sampled cells showing RBFOX2 motif detection to 1 cell resolution.

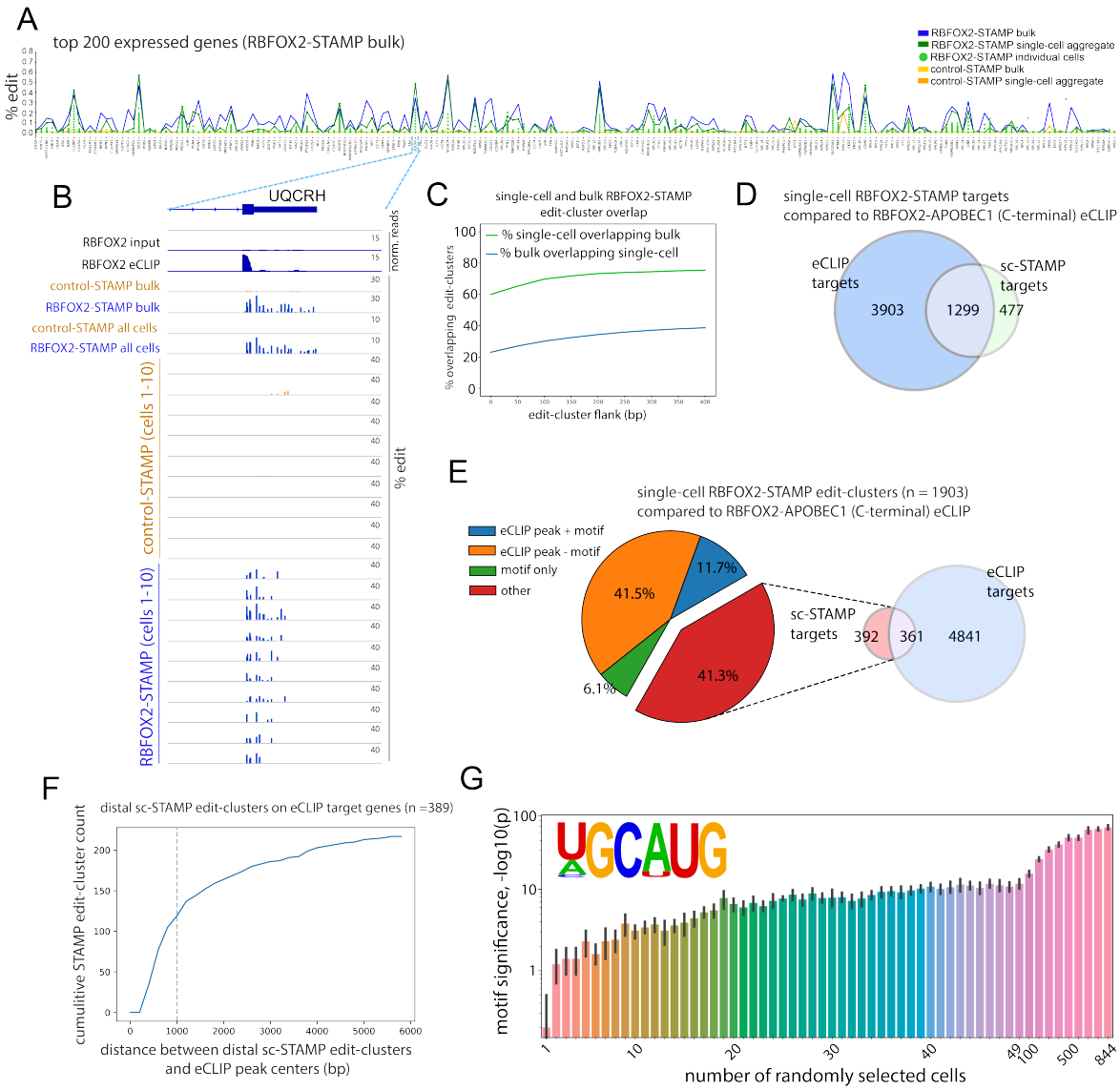

A
top 200 expressed genes (RBFOX2-STAMP bulk)
% edit
RBFOX2-STAMP bulk
RBFOX2-STAMP single-cell aggregate
RBFOX2-STAMP individual cells
control-STAMP bulk
control-STAMP single-cell aggregate
B
UQCRH
RBFOX2 input
RBFOX2 eCLIP
control-STAMP bulk
RBFOX2-STAMP bulk
control-STAMP all cells
RBFOX2-STAMP all cells
norm. reads
control-STAMP (cells 1-10)
RBFOX2-STAMP (cells 1-10)
% edit
C
single-cell and bulk RBFOX2-STAMP edit-cluster overlap
% single-cell overlapping bulk
% bulk overlapping single-cell
% overlapping edit-clusters
edit-cluster flank (bp)
D
single-cell RBFOX2-STAMP targets compared to RBFOX2-APOBEC1 (C-terminal) eCLIP
eCLIP targets
sc-STAMP targets
3903
1299
477
E
single-cell RBFOX2-STAMP edit-clusters (n = 1903) compared to RBFOX2-APOBEC1 (C-terminal) eCLIP
eCLIP peak + motif
eCLIP peak - motif
motif only
other
41.5%
11.7%
41.3%
6.1%
sc-STAMP targets
392
361
4841
eCLIP targets
F
distal sc-STAMP edit-clusters on eCLIP target genes (n =389)
cumulitive STAMP edit-cluster count
distance between distal sc-STAMP edit-clusters and eCLIP peak centers (bp)
G
motif significance, -log10(p)
number of randomly selected cells

Figure 2.6. Deconvolution of multiple RBPs and cell-type specific targets. A, Uniform Manifold Approximation and Projection (UMAP) analysis of gene expression from merged 72-hour high-induction control- and RBFOX2:TIA1-STAMP cells with capture sequence RBFOX2-STAMP (blue, n = 844) and TIA1-STAMP cells (red, n = 527) highlighted. B, UMAP analysis using ε score rather than gene expression after merging 72-hour high-induction control-STAMP cells (orange). C, UMAP plot as in B color-coded by ε score Louvain clustering into RBFOX2-population (blue), TIA1-population (red) and background-population (gray) populations with control-STAMP cells (orange) overlaid. D, Heatmap of normalized ε score signatures for RBFOX2- and TIA1-population cells compared to control-STAMP and background cells on the top 25 differentially edited gene targets. E, IGV browser tracks showing SAILOR quantified edit fractions for the top 5 control-, RBFOX2-, and TIA1-STAMP cells (ranked by summed ε scores) on the *NPM1*, *BTF3*, and *CFL1* gene targets.

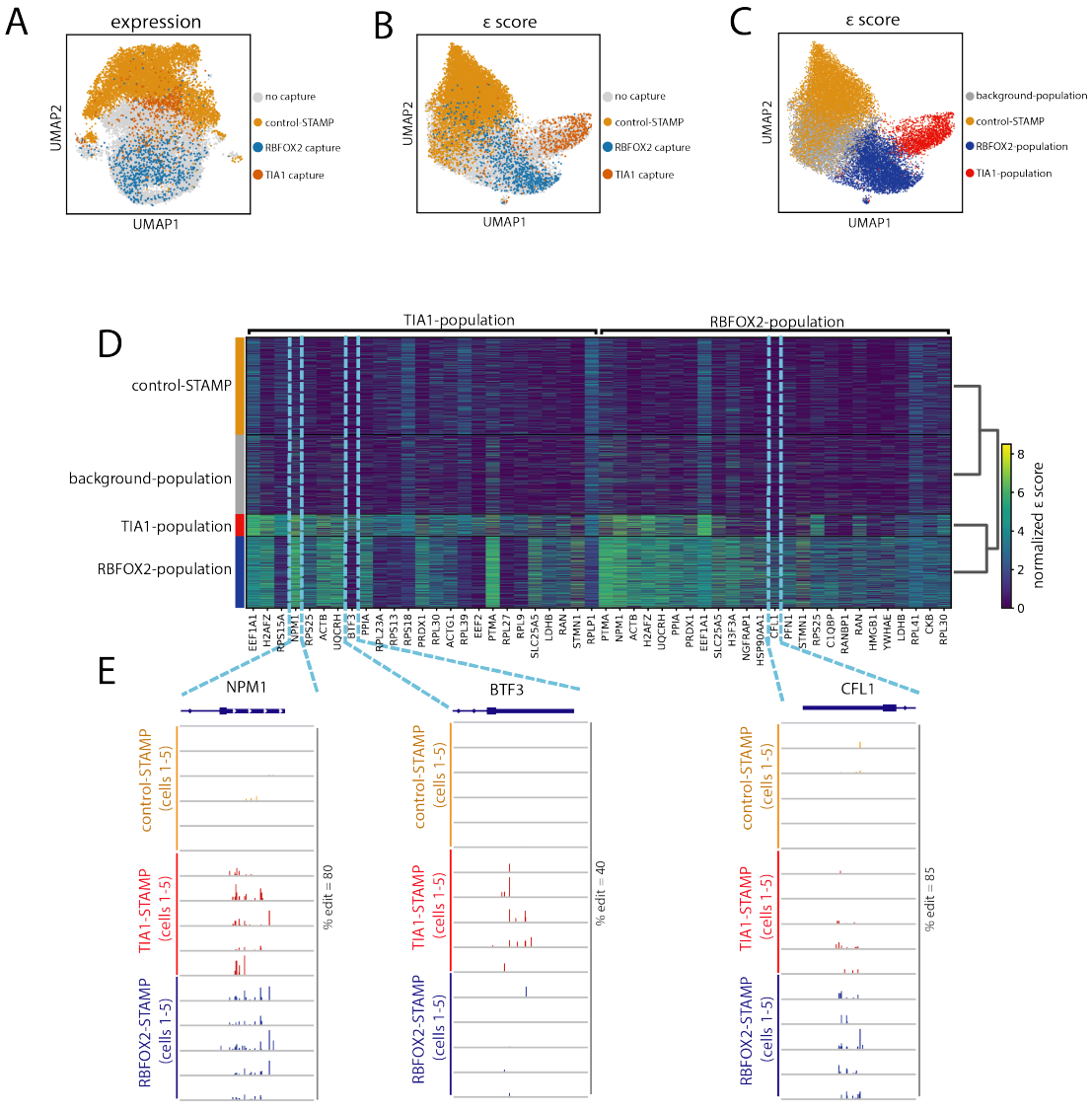
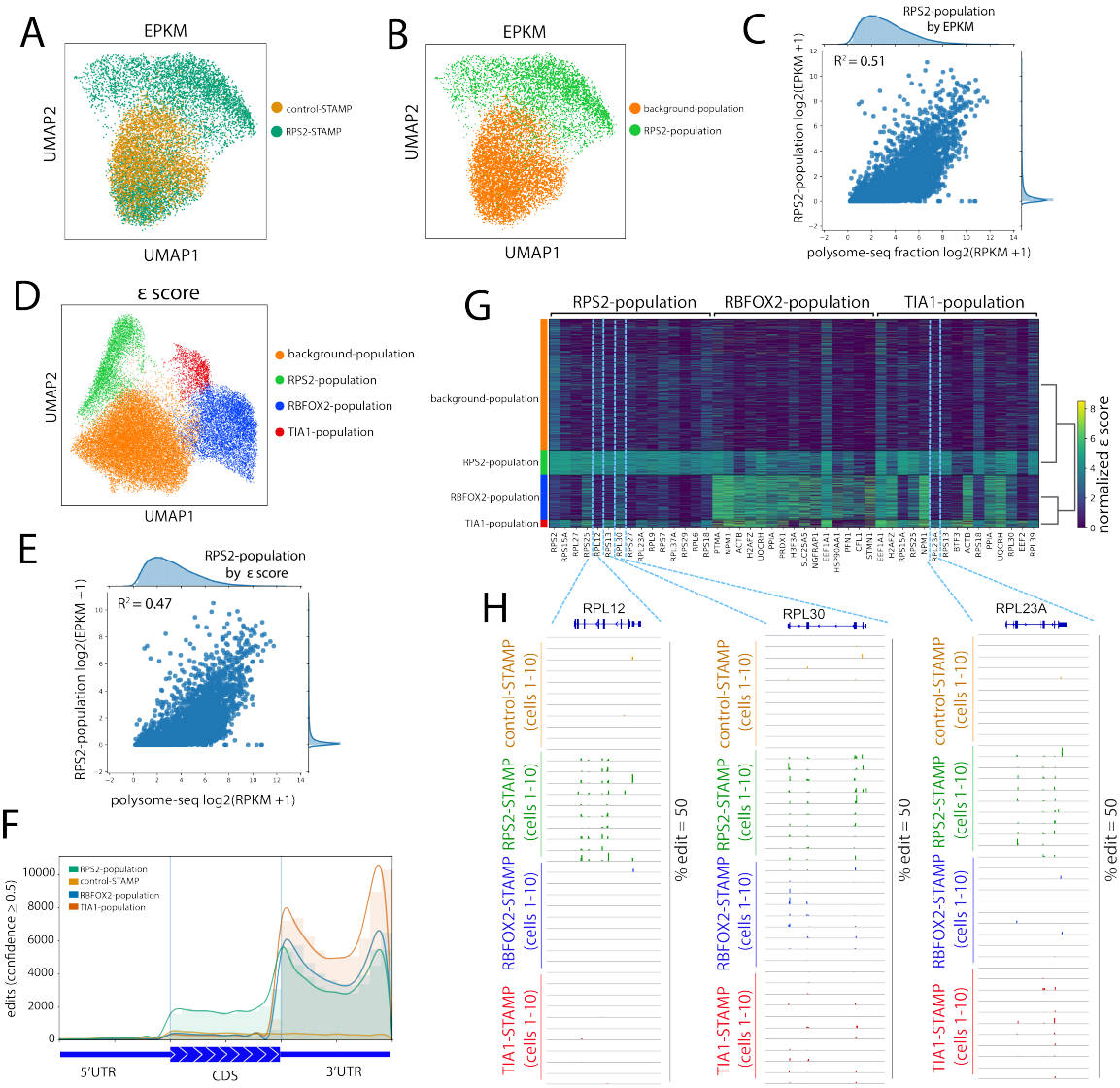

A
expression
UMAP2
UMAP1
no capture
control-STAMP
RBFOX2 capture
TIA1 capture
B
ε score
UMAP2
UMAP1
no capture
control-STAMP
RBFOX2 capture
TIA1 capture
C
ε score
UMAP2
UMAP1
background-population
control-STAMP
RBFOX2-population
TIA1-population
D
TIA1-population
RBFOX2-population
control-STAMP
background-population
TIA1-population
RBFOX2-population
normalized ε score
0
2
4
6
8
EEF1A1 H2AFZ RPS15A NPM1 RPS25 ACTB UQCRH BTF3 PPIA RPL23A RPS13 RPS18 PRDX1 RPL30 ACTG1 RPL39 EEF2 PTMA RPL27 RPL9 SLC25A5 LDHB RAN STMN1 RPLP1
PTMA NPM1 ACTB H2AFZ UQCRH PPIA PRDX1 EEF1A1 SLC25A5 H3F3A NGFRAP1 HSP90AA1 CFL1 PFN1 STMN1 RPS25 C1QBP RANBP1 RAN HMGB1 YWHAE LDHB RPL41 CKB RPL30
E
NPM1
BTF3
CFL1
control-STAMP (cells 1-5)
TIA1-STAMP (cells 1-5)
RBFOX2-STAMP (cells 1-5)
% edit = 80
% edit = 40
% edit = 85

Figure 2.7. Ribo-STAMP reveals ribosome occupancy from individual cells A, UMAP analysis of EPKM for 72-hour high-induction RPS2-STAMP (green), control-STAMP (orange). B, UMAP analysis of cells shown in A with EPKM Louvain clustering into background-population and RPS2-population. C, Comparison of EPKM-derived RPS2-population CDS+3'UTR EPKM values with poly-ribosome-fraction-enriched polysome-seq RPKM values. D, UMAP plot color-coded by ε score Louvain clustering into background-cluster (orange), RBFOX2-cluster (blue), TIA1-cluster (red), and 677 RPS2-cluster (green) from merged 72-hour high-induction STAMP experiments. E, Comparison of ε score-derived RPS2-population CDS+3'UTR EPKM values with poly-ribosome-fraction-enriched polysome-seq RPKM values. F, Metagene plot showing distribution for aggregate cell edits ($\geq$ 0.5 confidence level) from control-STAMP, RPS2-cluster, TIA1-cluster, and RBFOX2-cluster cells across 5'UTR, CDS and 3'UTR gene regions for the top quartile of ribosome occupied genes. G, Heatmap of normalized ε score signatures for RPS2-population, RBFOX2-population, and TIA1-population cells compared to background cells on the top 15 differentially edited gene targets. H, IGV browser tracks showing edit fractions for the top 10 control-, RPS2-, RBFOX2-, and TIA1-STAMP cells (ranked by summed ε scores) on the *RPL12*, *RPL30* and *RPL23A* gene targets.


A
EPKM
UMAP2
UMAP1
control-STAMP
RPS2-STAMP
B
EPKM
UMAP2
UMAP1
background-population
RPS2-population
C
RPS2-population
by EPKM
R² = 0.51
RPS2-population log2(EPKM +1)
polysome-seq fraction log2(RPKM +1)
D
ε score
UMAP2
UMAP1
background-population
RPS2-population
RBFOX2-population
TIA1-population
E
RPS2-population
by ε score
R² = 0.47
RPS2-population log2(EPKM +1)
polysome-seq log2(RPKM +1)
F
RPS2-population
control-STAMP
RBFOX2-population
TIA1-population
edits (confidence ≥ 0.5)
5'UTR
CDS
3'UTR
G
RPS2-population
RBFOX2-population
TIA1-population
background-population
RPS2-population
RBFOX2-population
TIA1-population
normalized ε score
H
RPL12
RPL30
RPL23A
control-STAMP (cells 1-10)
RPS2-STAMP (cells 1-10)
RBFOX2-STAMP (cells 1-10)
TIA1-STAMP (cells 1-10)
% edit = 50

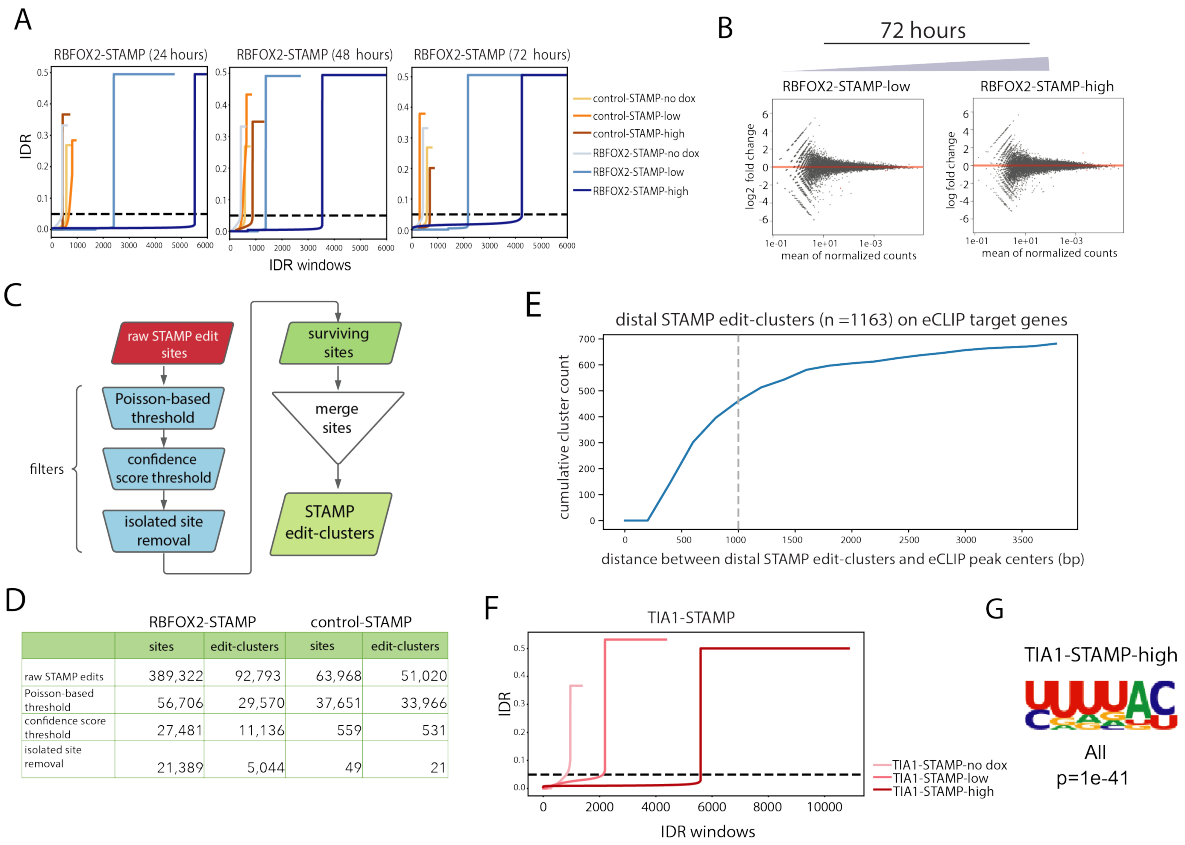


Figure 2.8. Additional RBP-STAMP reproducibility and concordance with eCLIP. A, Irreproducible Discovery Rate (IDR) analysis comparing ≥ 0.5 confidence edit windows for increasing levels of RBFOX2-STAMP at 24, 48 and 72 hours. B, Differential expression (DEseq2) analysis of RBFOX2-STAMP for increasing levels of RBFOX2-STAMP at 72 hours. C, STAMP edit-site filtering and cluster-calling workflow. D, Number of control- and RBFOX2-STAMP edit sites and clusters retained after each filtering step in C. E, Cumulative distance measurement from RBFOX2-STAMP distal edit-clusters to eCLIP peaks on targets genes. F, Irreproducible Discovery Rate (IDR) analysis comparing $0.5 \geq$ confidence level edit windows for increasing levels of TIA1-STAMP at 72 hours. G, Motif enrichment using HOMER and shuffled background on TIA1-STAMP edit-clusters.

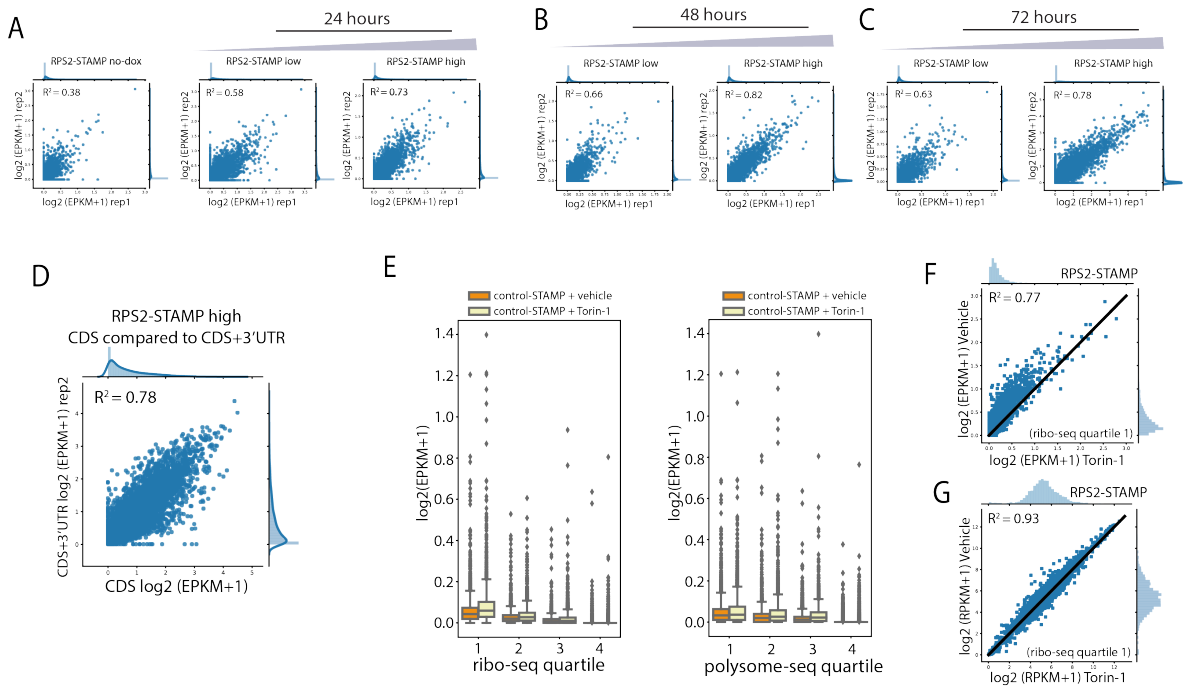


Figure 2.9. Ribo-STAMP reproducibility and response to mTOR pathway perturbations A–C, Scatterplot comparisons of CDS+3'UTR EPKM values from RPS2-STAMP replicate experiments showing high, dose-dependent correlation at 24 (A), 48 (B) and 72 hours (C). D, Scatterplot comparison of CDS EPKM values with CDS+3'UTR EPKM values for RPS2-STAMP. E, Comparison of EPKM from vehicle treated 72-hour high-induction control-STAMP compared to Torin-1 treated 72-hour high-induction control-STAMP showing no significant signal reduction for top ribosome occupied quartile genes containing Torin-1 sensitive TOP genes as detected by ribo-seq (Q1 p=1.0, n=3589 genes, Wilcoxon rank-sum one-sided) and polysome profiling (Q1 p=1.0, n=3589 genes, Wilcoxon rank-sum one-sided). F, Scatterplot comparison of CDS+3'UTR EPKM values on Ribo-seq top quartile genes (n=3589) for Torin-1 treated and vehicle treated RPS2-STAMP 72-hour high (1μg/ml) doxycycline inductions as in Figure 2.3H. G, Scatterplot comparison of CDS+3'UTR RPKM values on Ribo-seq quartile-1 genes (n=3589) for Torin-1 treated and vehicle treated RPS2-STAMP 72-hour high (1μg/ml) doxycycline inductions.

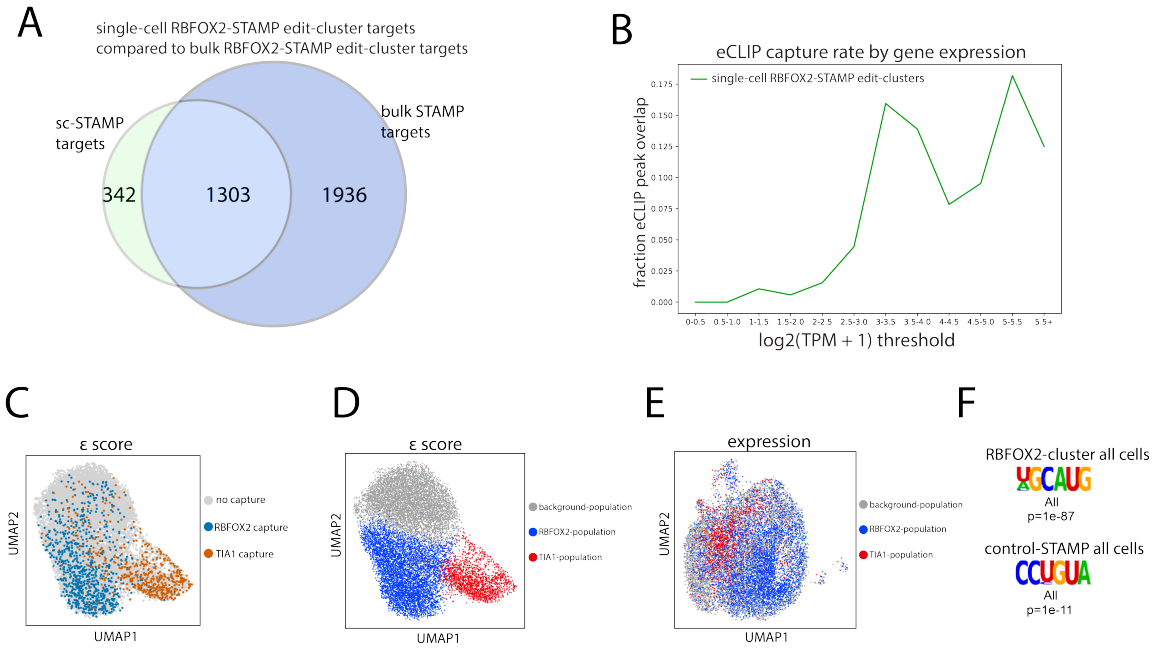


Figure 2.10. Single-cell RBP-RNA interaction detection by STAMP. A, Overlap between single-cell and bulk RBFOX2-STAMP target genes containing edit-clusters. B, Fraction of RBFOX2-APOBEC1 eCLIP peaks overlapping low and high induction single-cell RBFOX2-STAMP edit-clusters at increasing expression (TPM) thresholds. C, UMAP plot using ε score from RBFOX2-STAMP and TIA1-STAMP mixture with capture sequence RBFOX2-STAMP (blue, n=844) and TIA1-STAMP cells (red, n=527) highlighted. D, UMAP plot as in A color-coded by Louvain clustering into RBFOX2-cluster (blue), and TIA1-cluster (red), or background-cluster (gray) populations. E, UMAP plot of gene expression for ε score Louvain clusters defined in (D). F, Motif enrichment using HOMER from ≥ 0.99 confidence edits from combined RBFOX2-cluster and control-STAMP cells.

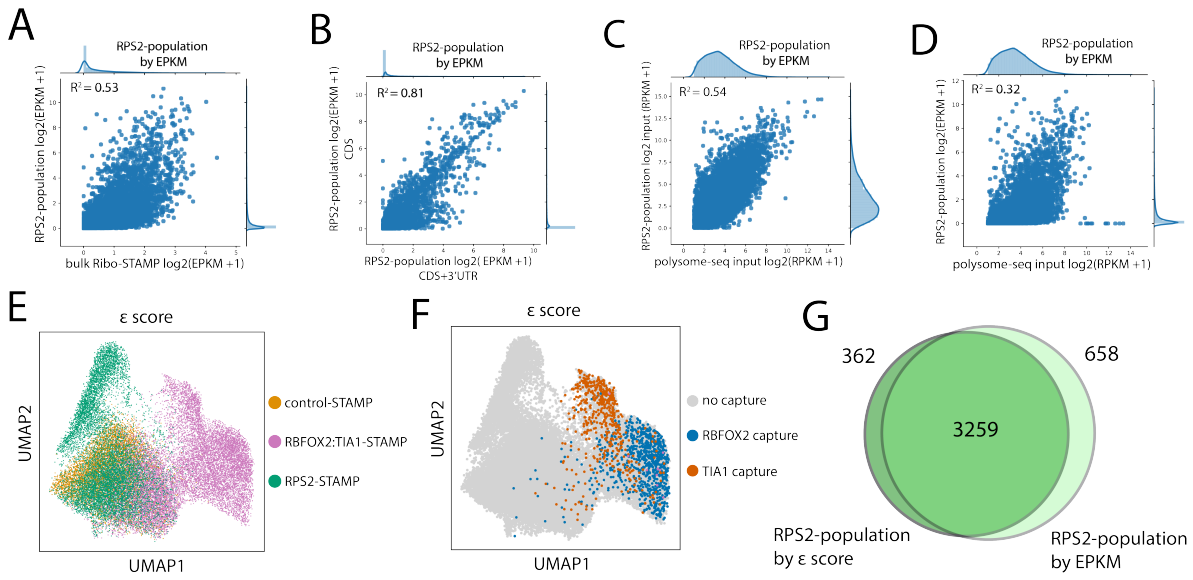


Figure 2.11. Single Ribo-STAMP detects ribosome occupancy from individual cells A, Genome-wide comparison of CDS+3'UTR EPKM values for bulk and single-cell EPKM-derived RPS2-population. B, Comparison of EPKM-derived RPS2-population CDS and CDS+3'UTR EPKM values. C, Comparison of EPKM-derived RPS2-population total mRNA RPKM values with total mRNA RPKM values from polysome-seq input. D, Comparison of EPKM-derived RPS2-population CDS+3'UTR EPKM values with total mRNA RPKM values from polysome-seq input. E, UMAP analysis of ε score from merged 72-hour high-induction RPS2-STAMP (green), control-STAMP (orange) and mixed-cell RBFOX2:TIA1-STAMP (purple) single-cell experiments. F, UMAP plot as in E with only capture sequence RBFOX2-STAMP (blue, n=844) and TIA1-STAMP cells (red, n=527) highlighted. G, Individual cell barcode overlap for EPKM-derived and ε score-derived RPS2-populations.

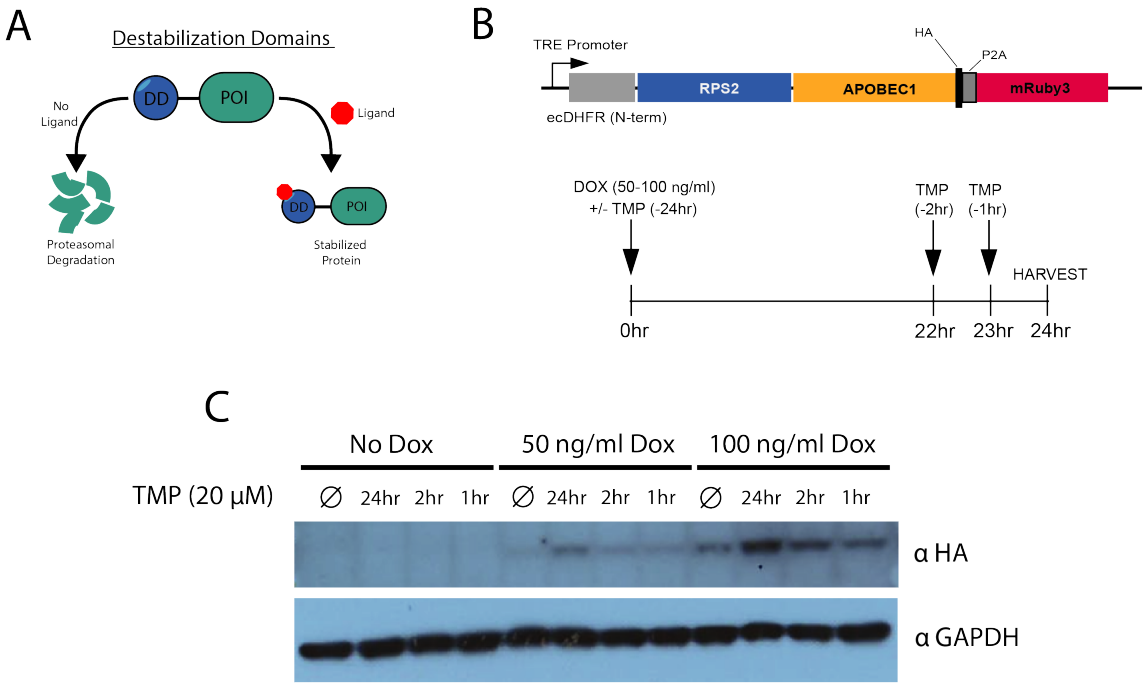


Figure 2.12. Destabilization domain tag of STAMP allows for post-transcriptional regulation. A, Schematic for destabilization domain-mediated (DD) post-transcriptional control of STAMP proteins. Expression of STAMP in the absence of the small molecule TMP results in constant proteosomal turnover, while small molecule supplementation stabilizes the structure of the domain, leading to protein-level expression. B, Map of new inducible STAMP lentiviral vector. Protein expression will be controlled by doxycycline via tet-responsive element (TRE)-containing promoter, as well as TMP using an N-terminal ecDHFR DD (top). Time course for proof-of-principle, dual induction STAMP expression experiment. C, Western blot demonstrating efficacy of 24 hours of induction by doxycycline (50 and 100 ng/ml) followed by treatment of 20 μM TMP for 1, 2, and 24 hours.

References

Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res, 46, 2159-2168.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M. & Schloss, J. A. 2008. The potential and challenges of nanopore sequencing. Nat Biotechnol, 26, 1146-53.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods, 10, 1213-8.

Deffit, S. N., Yee, B. A., Manning, A. C., Rajendren, S., Vadlamani, P., Wheeler, E. C., Domissy, A., Washburn, M. C., Yeo, G. W. & Hundley, H. A. 2017. The C. elegans neural editome reveals an ADAR target mRNA required for proper chemotaxis. Elife, 6.

Einstein, J. M., Perelis, M., Chaim, I. A., Meena, J. K., Nussbacher, J. K., Tankka, A. T., Yee, B. A., Li, H., Madrigal, A. A., Neill, N. J., Shankar, A., Tyagi, S., Westbrook, T. F. & Yeo, G. W. 2021. Inhibition of YTHDF2 triggers proteotoxic cell death in MYC-driven breast cancer. Mol Cell, 81, 3048-3064 e9.

Fu, S., Wang, A. & Au, K. F. 2019. A comparative evaluation of hybrid error correction methods for error-prone long reads. Genome Biol, 20, 26.

Gebauer, F., Schwarzl, T., Valcarcel, J. & Hentze, M. W. 2021. RNA-binding proteins in human genetic disease. Nat Rev Genet, 22, 185-198.

Gerstberger, S., Hafner, M. & Tuschl, T. 2014. A census of human RNA-binding proteins. Nat Rev Genet, 15, 829-45.

Gordon, S. P., Tseng, E., Salamov, A., Zhang, J., Meng, X., Zhao, Z., Kang, D., Underwood, J., Grigoriev, I. V., Figueroa, M., Schilling, J. S., Chen, F. & Wang, Z. 2015. Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule mRNA Sequencing. PLoS One, 10, e0132628.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H. & Glass, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell, 38, 576-89.

Hentze, M. W., Castello, A., Schwarzl, T. & Preiss, T. 2018. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol, 19, 327-341.

Hwang, B., Lee, J. H. & Bang, D. 2018. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med, 50, 96.

Iwamoto, M., Bjorklund, T., Lundberg, C., Kirik, D. & Wandless, T. J. 2010. A general chemical method to regulate protein stability in the mammalian central nervous system. Chem Biol, 17, 981-8.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol, 17, 239.

Li, B. B., Qian, C., Gameiro, P. A., Liu, C. C., Jiang, T., Roberts, T. M., Struhl, K. & Zhao, J. J. 2018. Targeted profiling of RNA translation reveals mTOR-4EBP1/2-independent translation regulation of mRNAs encoding ribosomal proteins. Proc Natl Acad Sci U S A, 115, E9325-E9332.

Li, Q. H., Brown, J. B., Huang, H. Y. & Bickel, P. J. 2011. Measuring Reproducibility of High-Throughput Experiments. Annals of Applied Statistics, 5, 1752-1779.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094-3100.

Li, H., Ruan, J. & Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res, 18, 1851-8.

Lorenz, D. A., Sathe, S., Einstein, J. M. & Yeo, G. W. 2020. Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. RNA, 26, 19-28.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G. & Yeo, G. W. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol, 20, 1434-42.

Love, M. I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 15, 550.

Matthews, M. M., Thomas, J. M., Zheng, Y., Tran, K., Phelps, K. J., Scott, A. I., Havel, J., Fisher, A. J. & Beal, P. A. 2016. Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity. Nat Struct Mol Biol, 23, 426-33.

Mcmahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W. & Rosbash, M. 2016.

TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. Cell, 165, 742-53.

Meyer, K. D. 2019. DART-seq: an antibody-free method for global m(6)A detection. Nat Methods, 16, 1275-1280.

Nguyen, D. T. T., Lu, Y., Chu, K. L., Yang, X., Park, S. M., Choo, Z. N., Chin, C. R., Prieto, C., Schurer, A., Barin, E., Savino, A. M., Gourkanti, S., Patel, P., Vu, L. P., Leslie, C. S. & Kharas, M. G. 2020. HyperTRIBE uncovers increased MUSASHI-2 RNA binding activity and differential regulation in leukemic stem cells. Nat Commun, 11, 2026.

Nussbacher, J. K., Batra, R., Lagier-Tourenne, C. & Yeo, G. W. 2015. RNA-binding proteins in neurodegeneration: Seq and you shall receive. Trends Neurosci, 38, 226-36.

Ponthier, J. L., Schluepen, C., Chen, W., Lersch, R. A., Gee, S. L., Hou, V. C., Lo, A. J., Short, S. A., Chasis, J. A., Winkelmann, J. C. & Conboy, J. G. 2006. Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. J Biol Chem, 281, 12468-74.

Porter, D. F., Miao, W., Yang, X., Goda, G. A., Ji, A. L., Donohue, L. K. H., Aleman, M. M., Dominguez, D. & Khavari, P. A. 2021. easyCLIP analysis of RNA-protein interactions incorporating absolute quantification. Nat Commun, 12, 1569.

Rahman, R., Xu, W., Jin, H. & Rosbash, M. 2018. Identification of RNA-binding protein targets with HyperTRIBE. Nat Protoc, 13, 1829-1849.

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

Rhoads, A. & Au, K. F. 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics, 13, 278-89.

Rosenberg, B. R., Hamilton, C. E., Mwangi, M. M., Dewell, S. & Papavasiliou, F. N. 2011. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat Struct Mol Biol, 18, 230-6.

Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J. & Abate, A. R. 2017. Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. Sci Rep, 7, 44447.

Song, Y., Yang, W., Fu, Q., Wu, L., Zhao, X., Zhang, Y. & Zhang, R. 2020. irCLASH reveals RNA substrates recognized by human ADARs. Nat Struct Mol Biol, 27, 351-362.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R. & Smibert, P. 2017. Simultaneous epitope and transcriptome

measurement in single cells. Nat Methods, 14, 865-868.

Tan, F. E., Sathe, S., Wheeler, E. C., Nussbacher, J. K., Peter, S. & Yeo, G. W. 2019. A Transcriptome-wide Translational Program Defined by LIN28B Expression Level. Mol Cell, 73, 304-313 e3.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B. B., Siddiqui, A., Lao, K. & Surani, M. A. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods, 6, 377-82.

Tegowski, M., Flamand, M. N. & Meyer, K. D. 2022. scDART-seq reveals distinct m(6)A signatures and mRNA methylation heterogeneity in single cells. Mol Cell, 82, 868-878 e10.

Thoreen, C. C., Chantranupong, L., Keys, H. R., Wang, T., Gray, N. S. & Sabatini, D. M. 2012. A unifying model for mTORC1-mediated regulation of mRNA translation. Nature, 485, 109-13.

Van Nostrand, E. L., Gelboin-Burkhart, C., Wang, R., Pratt, G. A., Blue, S. M. & Yeo, G. W. 2017. CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. Methods, 118-119, 50-59.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 13, 508-14.

Wolf, F. A., Angerer, P. & Theis, F. J. 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol, 19, 15.

Xu, W., Rahman, R. & Rosbash, M. 2018. Mechanistic implications of enhanced editing by a HyperTRIBE RNA-binding protein. RNA, 24, 173-182.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X. D. & Gage, F. H. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol, 16, 130-7.

Zhang, P., He, D., Xu, Y., Hou, J., Pan, B. F., Wang, Y., Liu, T., Davis, C. M., Ehli, E. A., Tan, L., Zhou, F., Hu, J., Yu, Y., Chen, X., Nguyen, T. M., Rosen, J. M., Hawke, D. H., Ji, Z. & Chen, Y. 2017. Genome-wide identification and differential analysis of translational initiation. Nat Commun, 8, 1749.

Chapter 3

Transcriptome-wide characterization of the ALS modifier RNA-binding protein ataxin-2

3.1 Abstract

Ataxin-2 is a ubiquitously expressed RNA-binding protein (RBP) determined to be a toxic modifier of TDP-43 in the context of amyotrophic lateral sclerosis (ALS) models. This has been a promising target gene for therapeutic intervention, though little is known about the regulatory role of ataxin-2 in the central nervous system, particularly with respect to RNA metabolism. To address this, we performed RNA sequencing on spinal cords of ataxin-2 knockout mice crossed with either normal or pathogenic TDP-43 mice. To map the RNA-binding landscape, we also performed enhanced CLIP (eCLIP) in mouse spinal cord and brain. We find hundreds of differentially expressed genes affected by ataxin-2 knockout, the majority of which can be explained by direct binding of ataxin-2 to UG-rich sequences in the 3'UTR of target transcripts. We also map discover a subset of genes whose signatures can be reversed with ataxin-2 depletion, corresponding to neuroinflammatory pathways and neuronal processes, though most of these genes are not bound by ataxin-2. Altogether, this study adds another layer of detail to uncovering the regulatory roles of ataxin-2 in spinal cord in both normal and disease contexts, as well as additional support that ataxin-2 might serve as a potentially effective therapeutic target.

3.2 Introduction

Amyotrophic lateral sclerosis (ALS) is a rare and fatal neurodegenerative disease caused by progressive death of both upper and lower motor neurons (MNs) in the brain and spinal cord (Brown and Al-Chalabi, 2017, Taylor et al., 2016, Kiernan et al., 2011). The underlying causes of ALS remain a mystery, with only 10% of ALS cases being explained through genetic inheritance (familial ALS, or fALS) which the remainder of cases occurring sporadically (sALS) (Pasinelli and Brown, 2006, Byrne et al., 2011). However, aberrant RNA-binding protein (RBP) biology is a recurrent molecular hallmark of both fALS and sALS (Neumann et al., 2006, Kwiatkowski et al., 2009, Neumann et al., 2011, Kim et al., 2013, Diaz-Garcia et al., 2021). The nuclear RBP TDP-43 (TARDBP) has been closely linked to ALS pathogenesis, with most patients undergoing cytoplasmic mislocalization and aggregation of TDP-43 protein in MNs of the motor cortex and spinal cord (Neumann et al., 2006). Additionally, pathogenic mutations have been characterized in ALS patients (Sreedharan et al., 2008), highlighting the importance of this protein in disease causation. It is therefore still of great interest to understand the both the loss-of-function (Ling et al., 2015, Klim et al., 2019, Melamed et al., 2019, Ma et al., 2022) and toxic gain-of-function (Neumann et al., 2006, Barmada et al., 2010, Walker et al., 2015) roles that TDP-43 plays in neuropathogenesis.

Due to the essential nature of TDP-43 for organismal development and survival, recent efforts have focused on uncovering alternative target genes for therapeutic intervention. Once such protein, ataxin-2 (ATXN2) was found to be a potent modifier of TDP-43 toxicity in cells (Elden et al., 2010). Ataxin-2 is an RBP most often associated with causing spinocerebellar ataxia type 2 (SCA2), a lethal autosomal dominant disease affecting predominantly the cerebellum, caused by longer CAG-trinucleotide (¿34X) expansion sequence mutations in the ATXN2 gene (Imbert et al., 1996, Pulst et al., 1996, Lastres-Becker et al., 2008). This sequence encodes the amino acid glutamine (Q), and thus often falls into a broader family of neurodegenerative disorders called polyglutamine (polyQ) diseases (Shao and Diamond, 2007). At the same

time, individuals with intermediate expanded polyQ alleles (27-34X) in the ATXN2 gene were at higher risk of ALS (Elden et al., 2010, Yu et al., 2011). Surprisingly, reduction of endogenous levels of ataxin-2, either by genetic depletion (knockout; KO) or antisense-oligonucleotide (ASO) knockdown, was found to be largely protective in an in vivo model of TDP-43 toxicity, significantly expanding lifespan and motor function of transgenic mice (Becker et al., 2017). Seeing that 97% of total sALS and fALS cases undergo some degree of aberrant TDP-43 pathology (Mackenzie et al., 2007), this served as seminal discovery in the field of ALS therapeutics by highlighting a broadly applicable strategy to target a non-lethal gene for ALS treatment.

Given the prominent role that ataxin-2 plays in disease modification and causation, it is of tremendous interest to determine the functional role of ataxin-2 in the central nervous system. Ataxin-2 is a ubiquitously expressed RBP conserved across eukaryotic species from yeast to human, yet its regulatory roles in both healthy and diseased neurons remain largely unknown, as do its part in mediating downstream ALS pathogenesis. Ataxin-2 ablation is seemingly tolerated in mammals, as Atxn2 knockout mouse models prove to viable through adulthood and remain fertile, though mice may suffer from adult-onset obesity through metabolic pathway disruption (Kiehl et al., 2006). Although ataxin-2 may not be essential for proper development and neurogenesis in mouse models, it remains crucial to thoroughly understand the long-term consequences that might result from ataxin-2 depletion. Knowing that RBP dysregulation and aberrant RNA-processing play a critical role in neurological disease pathogeneses, we sought to characterize the transcriptome-wide landscape of ataxin-2 in central nervous system (CNS) tissue in both normal and ALS-associated contexts.

To do so, we performed RNA sequencing on lumbar spinal cord samples harvested from Atxn2 knockout mice that had been crossed with either normal or pathogenic TDP-43 (Tg) backgrounds a pre-symptomatic (postnatal day 15, or P15) and symptomatic (P21) timepoints. Although overall fold changes were subtle on a per-gene basis, we identified several hundred genes (2,328 in wildtype, 923 in TDP-43 mice), that were differentially expressed with Atxn2

knockout at P21, the typical time that transgenic mice lose the ability to walk (Wils et al., 2010, Becker et al., 2017) prior to death at approximately P24. Although genes commonly depleted genes overlapping in these backgrounds are typically responsible for maintaining neuronal processes and development, genes downregulated in TDP-43 background fell along the axis of stress response and apoptotic signaling pathways, indicating a reversal in cell death. Much like the motor phenotype seen in mice, the severity of gene expression changes for many genes were dose-dependent, with *Atxn2* heterozygous mice experiencing intermediate effects in terms of RNA abundance. We also identified a subpanel of 609 genes that undergo directional reversal with ataxin-2 deletion, resulting in overall molecular signatures that resemble wildtype littermates. Though the vast majority (95%) of differentially expressed genes bound by enhanced CLIP (eCLIP) saw an overall decrease in abundance with ataxin-2 depletion, and thus are likely stabilized by the presence of ataxin-2 protein, the reversal of these genes could not be clearly explained through binding. Therefore, we conclude that there are both RNA binding-dependent and -independent molecular signature changes are associated with Atxn2 knockout-mediated rescue of toxic TDP-43. Taken together, these findings further inform the relationship between ataxin-2 and TDP-43 at the level of RNA metabolism, as well as other cellular pathways that may be affected downstream of ataxin-2 depletion.

3.3 Results

3.3.1 Ataxin-2 knockout in toxic TDP-43 background results in restorative gene expression changes

We sought to understand the underlying gene expression changes accompanying genetic depletion of ataxin-2 within the central nervous system. To assess global transcriptome-wide changes occurring with ataxin-2 depletion, we performed RNA-seq on lumbar spinal cord sections of both wildtype (WT) and ataxin-2 knockout (Atxn2 KO) mice in the presence of both normal and pathogenic TDP-43 (TDP-43 transgenic, $TDP\text{-}43^{Tg/Tg}$) at both pre-symptomatic

(postnatal day 15, P15) and post-symptomatic (postnatal day 21, P21) timepoints (Figure 3.1A). This particular mouse model, which expresses super-physiological levels of human TDP-43 under control of the pan-neuronal Thy1-promoter beginning at day P7, results in an aggressive phenotype whereby mice display gross motor dysfunction and death as a result of abnormal TDP-43 aggregation in homozygous backgrounds (Wils et al., 2010). However, crossing with null *Atxn2* mice was found to be neuroprotective, with both homozygous (*TDP-43$^{Tg/Tg}$ Atxn2$^{-/-}$*) and heterozygous (*TDP-43$^{Tg/Tg}$ Atxn2$^{+/-}$*) mice surviving longer than their ataxin-2 expressing littermates (*TDP-43Tg/Tg Atxn2$^{+/+}$*) in a dose-dependent manner (Becker et al., 2017).

Transcriptome-wide analysis of transgenic TDP-43 mice (referred to as TDP-43 samples) revealed widespread bidirectional gene expression changes compared to wildtype controls in a time-dependent manner, whereby noticeable differential expression (FDR < 0.05) of genes could only be observed at post-symptomatic timepoints (2,597 downregulated, 2,532 upregulated) rather pre-symptomatic (163 total genes total) (Figure 3.1B, C). Although these signatures match the motor dysfunction phenotype that typically presents around day 18-21, this was somewhat surprising given the drastic change in disease progression that occurs only days later. However, we did notice that aside from TDP-43 transgene status, timepoint largely drove sample clustering on a global level (Figure 3.2A). The vast majority of genes differentially expressed at P15 could also be detected at P21 (Figure 3.3E). Upregulated genes roughly fell into ontology pathways associated with proinflammatory pathways and negative regulation cell cycle response (Figure 3.2B), while downregulated pathways affected key neuronal pathways such as sterol metabolism and neurotransmitter production, ion transport across transmembrane, and cytoskeletal organization (Figure 3.2C). These gene signatures also match known phenotypes where homozygous transgenic mice experience neuronal loss in layer V of the motor cortex and anterior horn of the spinal cord, as well as some evidence for gliosis following pro-inflammatory response, which have been reported as soon as P18 (Wils et al., 2010). Indeed, curated assessment of cell-type enriched marker genes also shows a relative depletion of many well-characterized neuronal and oligodendrocyte genes, while resulting in gene-specific responses in some reactive

astrocyte genes such as *Gfap* and microglia-enriched genes such as *Cd68* and *Ccl3* (Figure 3.2D).

When comparing full *Atxn2* knockout to wildtype mice in transgenic TDP-43 background, we observe several hundred genes that are differentially expressed (FDR < 0.05) in mouse spinal cord at both early (214 upregulated, 464 downregulated) and late (468 downregulated, 455 downregulated). Interestingly, most of these gene expression changes are subtle with respect to fold-change, with median absolute log2 (fold change) values between 0.3-0.37, and were subject to noticeable filtering with increasing levels of fold-change thresholding (Figure 3.3 A-C). As was the case before, many of these gene expression changes appear to be time-dependent, though both conditions saw very few genes detected at higher fold-change cutoff. We therefore opted to include all genes, regardless of fold-change, in our downstream analysis so long as they reached statistical significance. Unlike or TDP-43 versus WT comparison, however, we noticed relatively fewer overlapping genes between timepoints (44 upregulated and 121 downregulated; Figure 3.3F), especially when compared to concordance within *Atxn2* KO versus WT alone (Figure 2.4D). In addition to *Atxn2*, other genes downregulated include the N-acetyltransferase enzyme *Nat8l*, potassium voltage-gated channel subunit protein *Kcnc3*, kinesin motor proteins *Kif5b* and *Kif5c*, the sodium-gated channel *Scn4b*, and the translation initiation factor *Eif5a2*. These are noteworthy observations, given that each of these genes was not only found in our non-TDP-43 *Atxn2* KO comparison, but found to be downregulated in the spinal cord of a recently generated knock-in model of SCA2 harboring 100 CAG repeats in the *Atxn2* locus (Canet-Pons et al., 2021, Sen et al., 2019), insinuating potential loss-of-function occurring downstream of ataxin-2 depletion that may still be tolerated in this otherwise disease-protected background. Meanwhile, shared upregulated genes include the translation elongation factor *Eef1a1*, which has been shown to activate expression of HSP70 chaperone proteins (Vera et al., 2014, Yu et al., 2021) that may be protective in ALS backgrounds through controlling phase separation properties of TDP-43, as well *Gap43*, a growth-cone localizing protein believed to play important roles in facilitating axonal growth and regeneration following neuronal lesions (Bomze et al., 2001, Caroni et al., 1997). Collectively, differentially expressed genes were also found to be

in concordance with phenotype, as noticeable upregulated pathways included trans-synaptic signaling, neuron projection development (Figure 3.1G), while major downregulated pathways included endoplasmic reticulum stress response, regulation of apoptosis, and autophagy (Figure 3.1H). Overall, these genetic signatures accurately reflect disease progression and closely match the endpoint phenotype of their respective genetic backgrounds.

3.3.2 Mapping the binding landscape of Atxn2 in mouse central nervous system using eCLIP

Presently, there have only been two published studies to map the transcriptome-wide landscape of ataxin-2, neither of which were performed in the mammalian central nervous system (Yokoshi et al., 2014, Singh et al., 2021). To identify CNS-specific RNA targets of ataxin-2 *in vivo*, we performed enhanced CLIP (eCLIP) on whole brain and spinal cord of adult mice in replicate (Figure 3.4A, Figure 3.5A). Overlapping of significant peaks ($p < 0.001$, fold enrichment over a sized-matched input > 8) following irreproducible discovery rate (IDR) peak merging between replicates revealed a large consensus of bound genes between brain and spinal cord, with 1,206 targets shared between the two tissues (brain, mBr: 1,529 genes; spinal cord, mSC: 1,707). Region-specific annotation of profiles reveals ataxin-2 reveals an overwhelming preference to interact with 3' UTR sequences, mapping 95.5% and 96.9% of peaks in mSC and mBr, respectively. This matches what has been published previously, where ataxin-2 binds in *cis* to elements in the 3'UTR of target genes to mediate post-transcriptional regulation. Reassuringly, peaks in commonly bound genes looked qualitatively similar (see *Uhmk1* in Figure 2.4E), insinuating functional target conservation across different areas of the central nervous system.

To date, we believe this establishes the first neuronal map of ataxin-2 binding, which is significant given its multiple implications in causing or modifying the development of neuronal diseases. Performing motif analysis on 6-mers reveals an enrichment for UG-rich, with the top motifs enriched in both brain and spinal cord eCLIPs being 'UGUGUG' or 'GUGUGU' (Figure

3.4F). This surprisingly matches the consensus motif for TDP-43 binding preference *in vitro* and *in vivo* (Kuo et al., 2009, Polymenidou et al., 2011, Tollervey et al., 2011), and is very interesting given supposed RNA-dependent nature of TDP-43's interaction with ATXN2 (Elden et al., 2010). Although TDP-43 predominantly binds to distal and proximal intronic sequences (>90%) to influence alternative splicing decisions, a subset or targets lie within 3'UTR (Figure 3.5B) and mediate post-transcriptional processes such as stability, translation, and transport of cognate transcripts (Alami et al., 2014, Krach et al., 2018, Briese et al., 2020). Through re-analysis of previously published iCLIP data in embryonic mouse brain (Rogelj et al., 2012), we find some evidence for co-occupancy of Tdp-43 and Atxn2 on 3'UTR of target transcripts such as *Lhfpl4*, which undergoes expression reversal with ataxin-2 depletion (Figure 3.5C), this mechanistic validation would require further investigation with similar experimental methods and tissue types.

Other motifs that were discovered were the consensus RBFOX-family motif ('UCGAUG') (Yeo et al., 2009, Lovci et al., 2013), as well as the UGAU-motif known to be recognized in part by several 3'UTR binders (Figure 3.4.G). This includes cleavage factor and 3'UTR processing enzyme CFI_m25 (Yang et al., 2010) and the Pumilio-family proteins PUM1/2 (Martinez et al., 2019, Smialek et al., 2021), which recognize a consensus 8-mer sequence ('UGUANAUA') that contains UGUA as its core. Ataxin-2 is a known interactor of RBFOX1, which was originally named ataxin-2 binding protein 1 (A2BP1) after being discovered to directly interact with the C-terminus of ataxin-2 (Shibata et al., 2000), so there is biological reasoning to expect such a motif. Although elements of these sequence motifs have appeared in previous mapping efforts, the ataxin-2 motif was previously identified as a much more generic U- or AU-rich element (Yokoshi et al., 2014, Singh et al., 2021). It is likely these disparities can be explained by technical differences in experiment type, but the high concentration of uridine residues in the 3'UTR forms a consensus binding preference across systems.

3.3.3 Ataxin-2 binding influences transcriptional abundance of target mRNA

To assess the regulatory impact of ataxin-2 knockout independent of pathogenic TDP-43, we began by looking further into our expression data from spinal cords of ataxin-2 knockout mice compared to wildtype at P21. We chose this timepoint because it resulted in the largest number of differentially expressed genes (Figure 3.6A; Figure 3.4A, D) and the mice have been reported to be otherwise healthy throughout development. Much like in the transgenic background, we noticed global, time-dependent changes in gene expression (258 genes at P15; 2,328 genes at P21) that were also sensitive to drastic fold-change thresholding (Figure 3.4A, D). Comparing the knockout data between normal and transgenic backgrounds at P21, we see many more differentially expressed genes in the monogenic context (2,328 versus 918), likely in part due to the complex genetic interaction that occurs between ataxin-2 depletion and pathogenic TDP-43 overexpression (Figure 3.6 B). We do see some shared, directionally matched differentially expressed genes in both backgrounds, with a higher degree of concordance in downregulated genes (170) over upregulated genes (69). This leads us to believe that there are both normal- and disease-specific perturbations that occur downstream of ataxin-2 depletion, likely due to the large differences in transcriptional composition in each separate state. For genes that are shared between both sets, we tend to see an ataxin-2 dose-dependent effect for both downregulated (Figure 3.6D) and upregulated (Figure 3.6E) genes, where heterozygous mice see an intermediate effect upon RNA abundance. These genes include *Ulk1*, a kinase suppressed in Atxn2 knockout and is a critical positive regulator of autophagy downstream of mTOR1 and may play a role in limiting selective autophagy to prolong survival in late stages of ALS (Rudnick et al., 2017). Additionally, suppression of ataxin-2 also led to an increase in levels of *Vegf*, a major vascularizing factor which has been found to be largely protective of motor neurons following ischemic injury to mice (Lambrechts et al., 2003). At the same time, we did notice examples of genes that were haplosufficient in nature, insinuating that these responses are likely secondary

effects of ataxin-2 depletion (Figure 3.7A, B).

We next wished to assess functionality of ataxin-2 interactions by intersecting our eCLIP data with our transcriptome-wide analyses. Interestingly, a large subset of all differentially expressed genes with ataxin-2 knockout could be tied to direct ataxin-2 binding, with 48% and 32% of differentially expressed genes at P15 (Figure 3.7C) and P21 (Figure 3.6F) respectively. This eCLIP binding proportionally enriched over total and unaffected gene populations, suggesting that direct binding may be regulating gene expression. Indeed, the majority of target genes were found to be suppressed with *Atxn2* knockout (94.4% of P15 genes, 94.9% of P21 genes; Figure 3.7C, Fidgure 3.6F). This observation is in line with previous works detailing ataxin-2 as a stabilizer and expression promoter of cognate RNA targets (Yokoshi et al., 2014, Singh et al., 2021). On the other hand, while still enriched in differentially expressed genes over both total and unchanged gene pools, ataxin-2 was responsible for a smaller proportion of differentially expressed genes in the Ataxin-2 knockout/TDP-43 background (Figure 3.7D, Figure 3.6G). More importantly, while ataxin-2 depletion still resulted in target gene downregulation at the pre-symptomatic stage (Figure 3.7D), binding could only account for approximately half of all differentially expressed genes at P21 (Figure 3.6G). While this is likely to be the case for complicated genetic interactions that take place between ataxin-2 and TDP-43, it does make identification of meaningful ataxin-2 target genes in disease context difficult. However, it does support the notion that ataxin-2 functions predominantly as a gene stabilizer in basal neuronal state, though this primary role in mediating RNA expression may become may be largely superseded in later stages of disease.

Knowing the effectiveness of ataxin-2 depletion in preventing TDP-43 aggregation-induced degeneration, we hypothesized that disease associated gene signatures could be reversed as well. To test this, we narrowed down our genes of interest to those that were differentially expressed with the introduction of pathogenic TDP-43 post-symptom onset (Figure 3.1C). We then looked for genes in the transgenic background that underwent differential expression in the opposite direction once crossed with ataxin-2 knockout mice. This yielded a set of 609 genes that

underwent a significant directional reversal with the removal of ataxin-2 (Figure 3.8A). Using these genes we were able to achieve successful grouping of *TDP-43$^{Tg/Tg}$ Atxn2 $^{-/-}$* mice with wildtype mice and apart from *TDP-43$^{Tg/Tg}$ Atxn2 $^{+/+}$* mice through hierarchical clustering. To focus more specifically on genes most likely to be directly affected by ataxin-2, we looked at 'conserved' genes that were not only reversed in TDP-43 pathogenic state with knockout but were also supported by non-diseased ataxin-2 knockout. This yielded a much narrower subset of 87 total genes (Figure 3.8B, C), a much smaller sample compared to the total 609 genes reversed at P21 with knockout, insinuating that there might be disease state-specific binding or expression paradigms that cannot be accounted for in wildtype sampling alone. Of this more conservative gene set, only 24 (19/56 downregulated genes, 5/31 upregulated genes) could be explained by direct binding via eCLIP. Although most bound genes see a decrease in expression as expected, this is a much smaller proportion than had been observed for all differentially expressed genes in non-diseased mice (Fig. 3.8B). Considering that this small subset only accounts for a fraction of the genes reversed with knockout, and that even fewer of these genes are found to be directly bound by ataxin-2, it may be concluded that while restorative gene expression changes can be accounted for with ataxin-2 knockout, most of these corrective effects of ataxin-2 depletion are not directly tied to its regulatory role as a stabilizer of target transcripts.

3.4 Discussion

Ataxin-2 is a broadly expressed RBP associated with several neurodegenerative disease. In addition to driving SCA2 through CAG-repeat expansion mutations, it's also a known modifier of several other late onset neurotological disorders including ALS, frontotemporal dementia (FTD), and Alzheimer's disease, suggesting a close linkage between protein function and disease (Laffita-Mesa et al., 2021). Closely related to its role in disease pathogenesis, ataxin-2 is a toxic modifier of pathogenic proteins such as TDP-43, which mis-localizes to the cytoplasm and aggregates in motor neurons of ALS cases both familial and sporadic in origin. Seeing that

almost all patients undergo some degree of TDP-43 dysregulation, it is of great interest to find a common point of intervention centered around this stage. Ataxin-2 has become a very attractive candidate for gene depletion therapeutics using ASOs, not only due to efficacy in preclinical mouse models of ALS but also the promise of safety due to knockout animal models being perfectly viable. It is therefore of utmost importance to understand potential short- and long-term consequences genetic ablation of ataxin-2 in neurons might cause. Although it is recognized as a clinically significant RBP, there has been little published work to corroborate the regulatory roles of ataxin-2 with respect to RNA metabolism. Through RNA- and eCLIP-seq, we investigated the functional landscape of ataxin-2 in spinal cord of healthy and diseased mice. Although there have been two relatively recent efforts to profile the spinal cord transcriptome downstream of ataxin-2, these were done with knock-in or transgenic models of SCA2 (Scoles et al., 2020, Canet-Pons et al., 2021). Although these studies converge upon some target genes that may shed light on how polyQ expansions contribute to motor neuron loss, neither are able to adequately decipher genetic loss-of-function versus toxic gain-of-function due to the monogenic nature of their models. Additionally, the addition of eCLIP allows us to assess how direct binding more directly to target transcripts might impact downstream gene regulation.

Transcriptomic profiling of mouse spinal cords reveals widespread and multi-tissue gene expression changes following chronic overexpression of human TDP-43 in neurons. While this is not necessarily surprising, these alterations in the transcriptome only became readily apparent at P21, well within the range of time once mice begin developing motor dysfunctions. This homozygous Thy1-TDP-43 model is considered particularly aggressive, resulting in widespread post-translationally modified intranuclear and cytoplasmic aggregation of TDP-43 in the motor cortex, spine, and beyond, and typically results in mice perishing within 4 weeks of birth (Wils et al., 2010). Sampling Atxn2 knockout mice from this pathogenic background and at this same timepoint, we find several hundred, relatively mild alterations in gene expression with gene depletion. Many of these downregulated genes fall into pathways regulating cellular stress response and apoptosis, while upregulated pathways pertain closer to metabolism, cellular

architecture, and trans-synaptic neuronal signaling. For these mice, the overall molecular profile matches their phenotype, affirming the potential of *Atxn2* as a target gene.

It is worth mentioning that of the several hundred transcripts found dysregulated with *Atxn2* knockout regardless of background, many genes (including *Nat8l, Kcnc3, Kif5b Kif5c, Scn4b, Unc13c, Eif5a2*) are all genes that have been described to also be downregulated in CAG100 knock-in models of SCA2, apparently due to loss-of-function (Canet-Pons et al., 2021). Additionally, gene expression changes become more pronounced with time, and it stands to reason that more impactful perturbations might increase as healthy mice proceed past P21 and into adulthood. Although severity of these gene expression changes remains small (Figure 3.3A) and null mice are considered reasonably healthy and viable, long-term neuronal effects of ataxin-2 depletion still remain inconclusive.

In mapping ataxin-2 across the brain and spinal cord of adult mice using eCLIP, we provide the first physical and functional "interactome" of ataxin-2 in the mammalian central nervous system. This can prove to be an incredibly valuable research tool to those interested in identifying target transcripts of Atxn2 both in terms of function and disease. We find that ataxin-2 predominantly binds the 3'UTR of target genes, and that most interaction partners are shared between brain and spinal cord. In characterizing the binding behavior of ataxin-2 further, we find that it prefers to target U- and UG-rich sequences regardless of tissue type. This is somewhat in-line with PAR-CLIP and HyperTRIBE methods of mapping, which found U-rich and UA-rich motifs in HEK293 cells and Drosophila brain, respectively (Singh et al., 2021, Yokoshi et al., 2014). Compared to eCLIP, these methods do offer potential for sequence selection bias (4SU incorporation, ADAR2-targeting preference) which might ultimately color these motif outcomes. The known motifs we obtain make logical sense given what we know about ataxin-2 (known interactor of RBFOX-proteins, 3'UTR interactor that regulates posttranscriptional processes), but the presence of UG-motifs seems particularly intriguing given it matches TDP-43's known binding preference. Future studies characterizing potential co-occupancy of these proteins will no doubt be mechanistically informative for understanding the true RNA-dependent relationship

of these proteins.

While we find that target genes of ataxin-2 are found to be downregulated upon knockout, this effect becomes diluted with introduction of pathogenic TDP-43 (Figure 3.6F, G). Moreover, we find the majority of gene expression changes reversed with Atxn2 knockout-mediated rescue could not be attributed directly to ataxin-2 depletion in normal tissue nor direct binding of ataxin-2 to mRNA. Thus, although ataxin-2 seems to play important regulatory roles in healthy CNS through binding and stabilizing its genetic repertoire, it's highly likely that its modifier effect is largely independent of its normal function in RNA-processing. However, it may still be playing a direct role in regulating neuronal survival in ALS. Ataxin-2 is a multi-functional protein thought to regulate gene expression through modulation of RNA stability, translation, and stress granule formation, though it has also been implicated in non-RNA-centric regulatory systems. For instance, the yeast ortholog of ataxin-2 (Pbp1) has been shown to act as a negative regulator of mTOR pathway through protein-protein interaction and supposed sequestration of mTORC1 complex, implicating it as s direct regulator of cellular growth beyond RNA-binding (Yang et al., 2019). Given the complicated nature that autophagy plays in early versus late stages of ALS progression, as well as the suppression of autophagy-related genes such as *Ulk1, Bin,* and *Bnip3* with knockout, it is possible that regulation of autophagy may give ataxin-2 depleted neurons a selective advantage of survival. Though we still continue to learn more about the mechanistic basis of ataxin-2's modifier effect, clinical studies of the ATXN2 targeting ASO (BIIB105) will remain incredibly interesting from the perspectives of both efficacy and safety.

3.5 Methods and Materials

3.5.1 Mouse breeding and animal care

Mouse breeding crosses, husbandry, care for homozygous TDP-43$^{Tg/Tg}$ mice, phenotyping/endpoint determination were performed as previously described in (Becker et al., 2017).

3.5.2 Mouse tissue collection

For RNA-seq, tissue collection for lumbar spinal cord sections was performed as previously described in Becker et al., 2017. For P15 samples, n=3 mice per genotype were used. For P21 samples, n=4 (2X M, 2X F) per genotype were used with the exception of Atxn2$^{+/-}$TDP-43$^{WT/WT}$ (n=3). For eCLIP experiments, brain and spinal cord sections were harvested as previously described in (Kapeli et al., 2016) and (Martinez et al., 2019), respectively. For each eCLIP experiment, brains and spinal cord sections were sampled in biological replicate from 8-week-old female C57Bl/6 mice.

3.5.3 RNA Extraction and library preparation

Each frozen sample was immediately resuspended in 1ml Trizol and lysed with a Kontes pestle until completely resuspended. RNA was extracted according to manufacturer's instructions, followed by TURBO DNase treatment using DNA-free DNA Removal Kit. RNA was assessed for quality using a Tapestation 2200 (Agilent). With 1μg of total RNA as input material, RNA-seq libraries were prepared using the TruSeq Stranded mRNA kit from Illumina according to manufacturers' instructions. Libraries were pooled and sequenced on a 75bp single-end run on a HiSeq4000 (Illumina). Libraries were sequenced to an average depth of 18M reads for differential gene expression.

3.5.4 RNA-seq processing

Raw RNA-seq reads were adapter-trimmed using cutadapt (v 1.14.0), mapped to repetitive elements (version 18.05) prior to alignment to mm10 using STAR (v 2.4.0i). Aligned reads were sorted and indexed using samtools (v 1.9), and gene features assigned to gencode annotation (mm10 v15) were quantified using Subread featureCounts package (v 1.5.3). Reads normalized and converted to transcripts to millions (TPM) and differential expression performed using DESeq2 (v 1.30.1). Genes were considered differentially expressed with an FDR < 0.05. Sex was used treated as a factor variable in addition to genotype (design = sex + geno). Only genes

with TPM > 0 were considered for pairwise differential expression analysis. A curated list of cell-specific marker genes for neural tissue was collected from the from (Cahoy et al., 2008, Chiu et al., 2013, D'Erchia et al., 2017, Phatnani et al., 2013).

3.5.5 eCLIP-seq library preparation

The standard, paired-end version of eCLIP-procedure was performed for ataxin-2 (Van Nostrand et al., 2016, Van Nostrand et al., 2020) on mouse brain and spinal cord pulverized and UV-cross-linked (400 mJ cm^{-2}, 254 nm) prior to snap freeze. Still-frozen crosslinked tissues were lysed supplemented with 11 ul Murine RNase Inhibitor per ml of lysate. Samples were sonicated for 5 cycles (30s on, 30s off) and treated with RNase I to digest RNA. For each sample, 10 μg Ataxin-2 polyclonal antibody (Proteintech, 21776-1-AP) was pre-coupled to 125 μl M-280 sheep anti-rabbit Dynabeads then added to lysate and incubated for 2 hours at 4^{o}C with rotation. In parallel, 10% input lysate was removed and used for Rb IgG IP control with 1ug antibody for western blot. Following incubation, 2% input was removed and kept for size-matched input sample, while the remainder of the RNP bound material was washed, subjected to dephosphorylation and 3' end ligated with an RNA adapter on bead. Samples (IP, Input, IgG) were run on an SDS-polyacrylamide gel and transferred onto a nitrocellulose membrane. Membrane pieces for IP and sized-matched input (SMInput) were cut from the size of the protein (140 kDa) to 75 kDa above and RNA was extracted using proteinase K digestion and subsequently purified on column. Input samples were 3'-end ligated separately with its own RNA adapter and libraries were subjected to reverse transcription using AffinityScript RT enzyme, and 5'-linker ligated with a DNA adapter. After quantifying with qPCR, libraries were amplified with Q5 PCR master mix using primers containing Illumina indexes and adapters. Libraries were then sequenced on an Illumina Hiseq 4000 on a 55PE run.

3.5.6 eCLIP sequence processing

Raw reads were processed essentially as described in Van Nostrand et al. 2016. Reads were adapter-trimmed and mapped to mouse-specific elements from Repbase (version 18.05) using STAR to remove repeat-mapping reads. Remaining reads were mapped to the mouse genome mm10 once again using STAR. PCR-duplicates were removed using the unique molecular identifier sequences in the 5' adaptor, and remaining reads were retained as "usable reads." Peaks on IP samples were called using CLIPper, and each peak was normalized to its corresponding input sample calculating the fraction of the number of usable reads from the IP sample relative to the usable reads from the SMInput. Each peak was assigned a fold enrichment (relative to SMInput) and p-value (χ^2 test, or Fisher's exact test if the observed or expected read number in eCLIP or SMInput was below 5). All eCLIP processing code is available on GitHub (https://github.com/YeoLab/eclip). Reproducible peak regions were determined across replicates using Irreproducible Discovery Ratio, or IDR (https://github.com/YeoLab/merge_peaks). Consensus IDR peaks were deemed significant at fold enrichment ≥ 8 over input and a $P \leq 0.001$.

3.5.7 Region analysis and enrichment of k-mers

Peaks were assigned genomic regions using mm10 Gencode annotations (https://github.com/byee4/annotator). To analyze sequence preferences in the significant peaks, we performed 6-mer analysis (https://github.com/byee4/clip_analysis), which performs k-mer enrichment analysis and HOMER findMotifs.pl. For 6-mer analysis, each the occurrence of possible sequence was calculated for input peaks and a Z-score was assigned per 6-mer to calculate frequency relative to shuffled background.

3.5.8 iCLIP data reprocessing

Previously published TDP-43 iCLIP from mouse embryonic 18 brain (Rogelj et al., 2012) was re-processed to genome build mm10 using the iCount pipeline (https://github.com/tomazc/iCount). Significant peaks were called on all x-link sites identi-

fied, and subsequent bed files were sorted and merged using bedtools merge.

3.6 Acknowledgements

Chapter 3, in full, is currently being prepared for submission for publication. Marina RJ, Becker LA, Nguyen TB, Gitler AD, Yeo GW. The dissertation author was the primary investigator and author of this material.

3.7 Figures

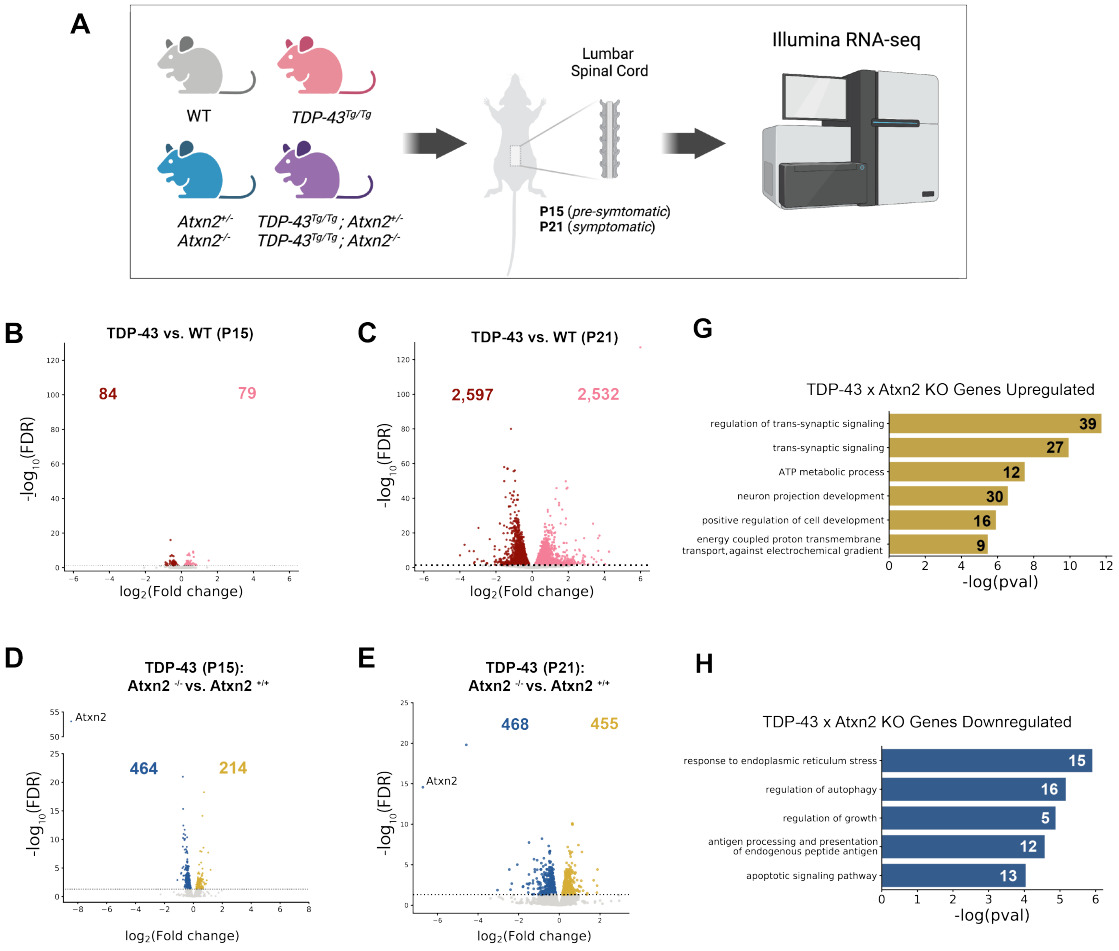
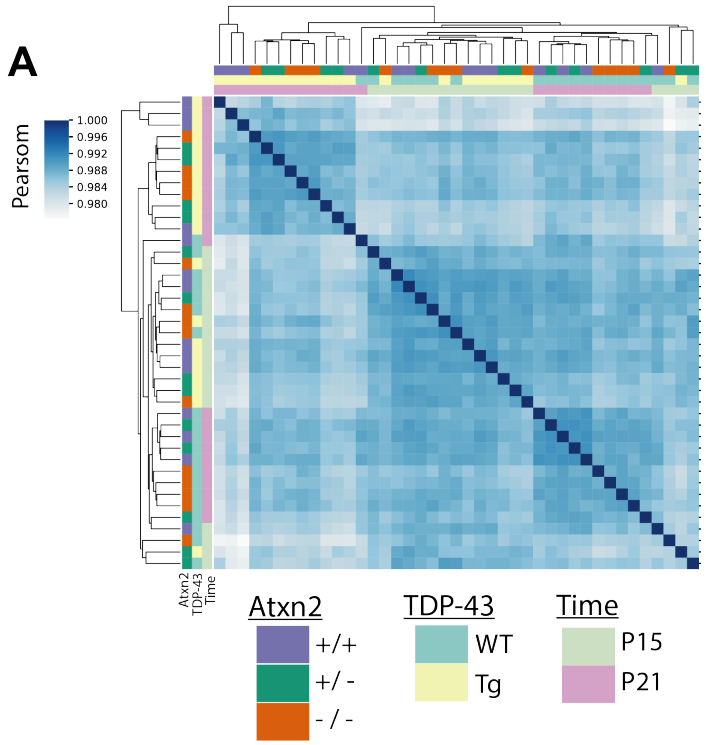
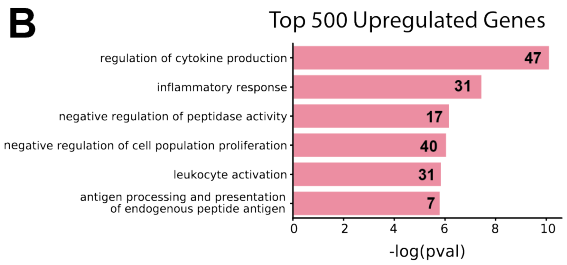
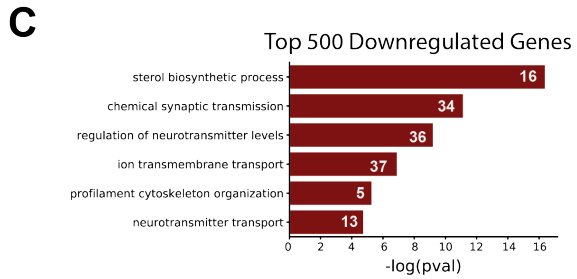


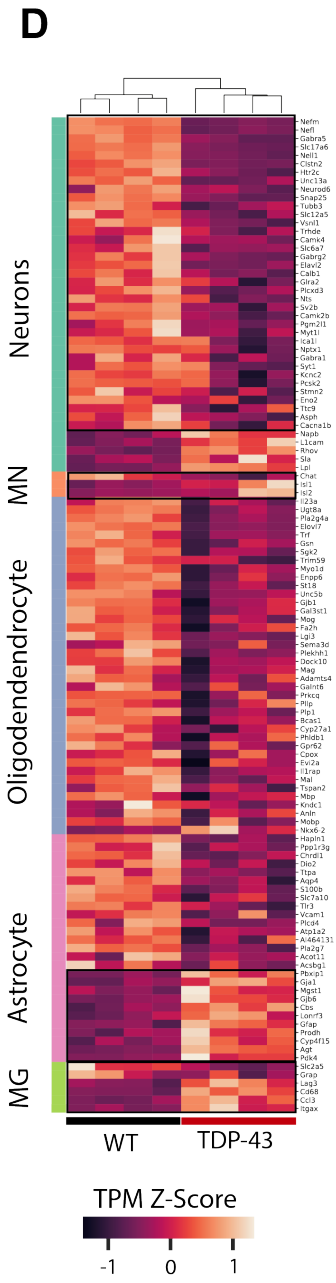
Figure 3.1. Transcriptome-wide signatures match phenotype in Atxn2 WT and KO backgrounds of transgenic TDP-43. A. Schematics for RNA-seq from spinal cord sections taken from pre-symptomatic (P15) and post-symptomatic (P21) timepoints. B. Volcano plots representing differentially expressed genes at P15 or C. at P21 for TDP-43 transgenic mice (FDR $<$ 0.05). D. Volcano plots representing Atxn2 knockout (-/-) mice in TDP-43 background at P15 or E. at P21. G. Gene Ontology terms for upregulated genes in (E). H.) Gene Ontology terms for downregulated genes in (E).

Figure 3.2. Transcriptome-wide analyses in mouse lumbar spinal cord. Supplement to Figure 3.1. A. Heatmap for all samples assayed clustered by Pearson correlation. Samples are labeled according to ataxin-2 knockout status, TDP-43 background, and timepoint assessed. B. Gene Ontology terms for top 500 differentially expressed upregulated genes by fold change of TDP-43 mice at P21 (related to Figure 3.1B). C. Gene Ontology terms for top 500 differentially expressed downregulated genes by fold change of TDP-43 mice at P21 (related to Figure 3.1B). D. Relative expression of tissue-specific marker genes for all neurons, motor neurons, oligodendrocytes, astrocytes, and microglia. Annotations were taken from references found in the Methods and Materials section.


A
Pearsom
1.000
0.996
0.992
0.988
0.984
0.980
Atxn2
TDP-43
Time
Atxn2
+/+
+/ -
- / -
TDP-43
WT
Tg
Time
P15
P21


B
Top 500 Upregulated Genes
regulation of cytokine production
47
inflammatory response
31
negative regulation of peptidase activity
17
negative regulation of cell population proliferation
40
leukocyte activation
31
antigen processing and presentation of endogenous peptide antigen
7
0
2
4
6
8
10
-log(pval)


C
Top 500 Downregulated Genes
sterol biosynthetic process
16
chemical synaptic transmission
34
regulation of neurotransmitter levels
36
ion transmembrane transport
37
profilament cytoskeleton organization
5
neurotransmitter transport
13
0
2
4
6
8
10
12
14
16
-log(pval)


D
Neurons
MN
Oligodendendrocyte
Astrocyte
MG
WT
TDP-43
TPM Z-Score
-1
0
1

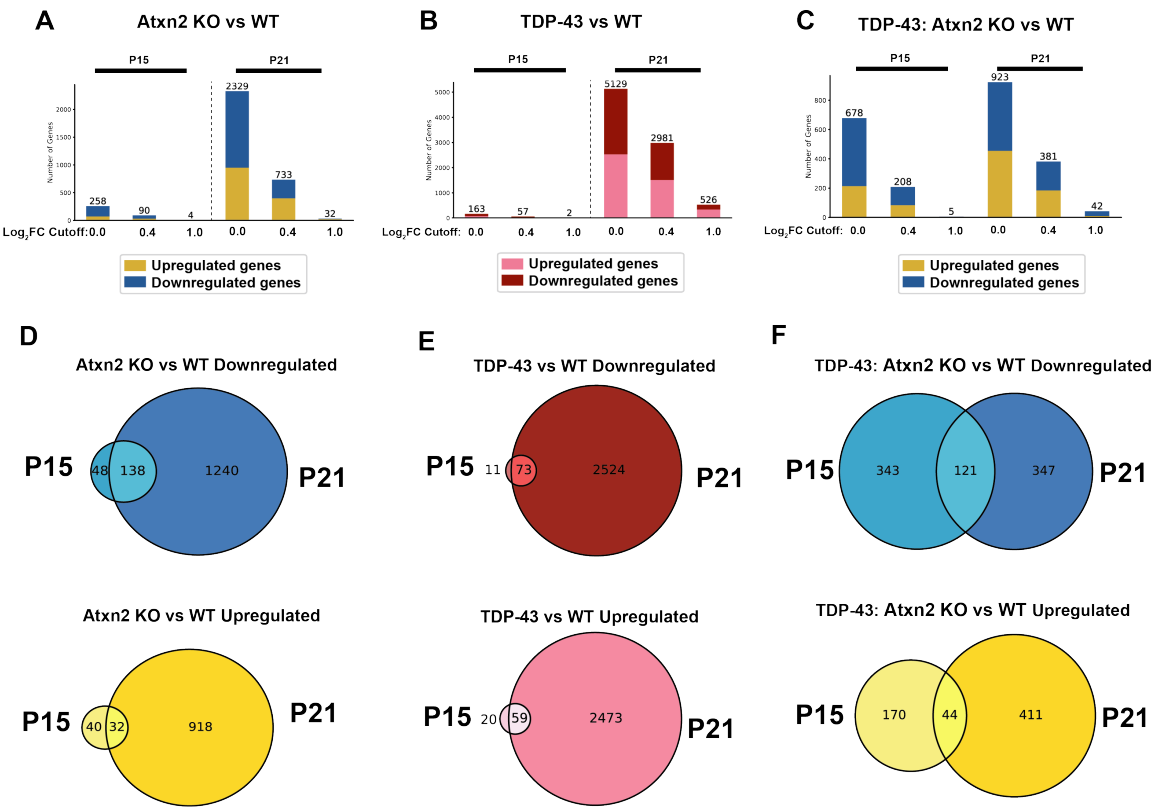
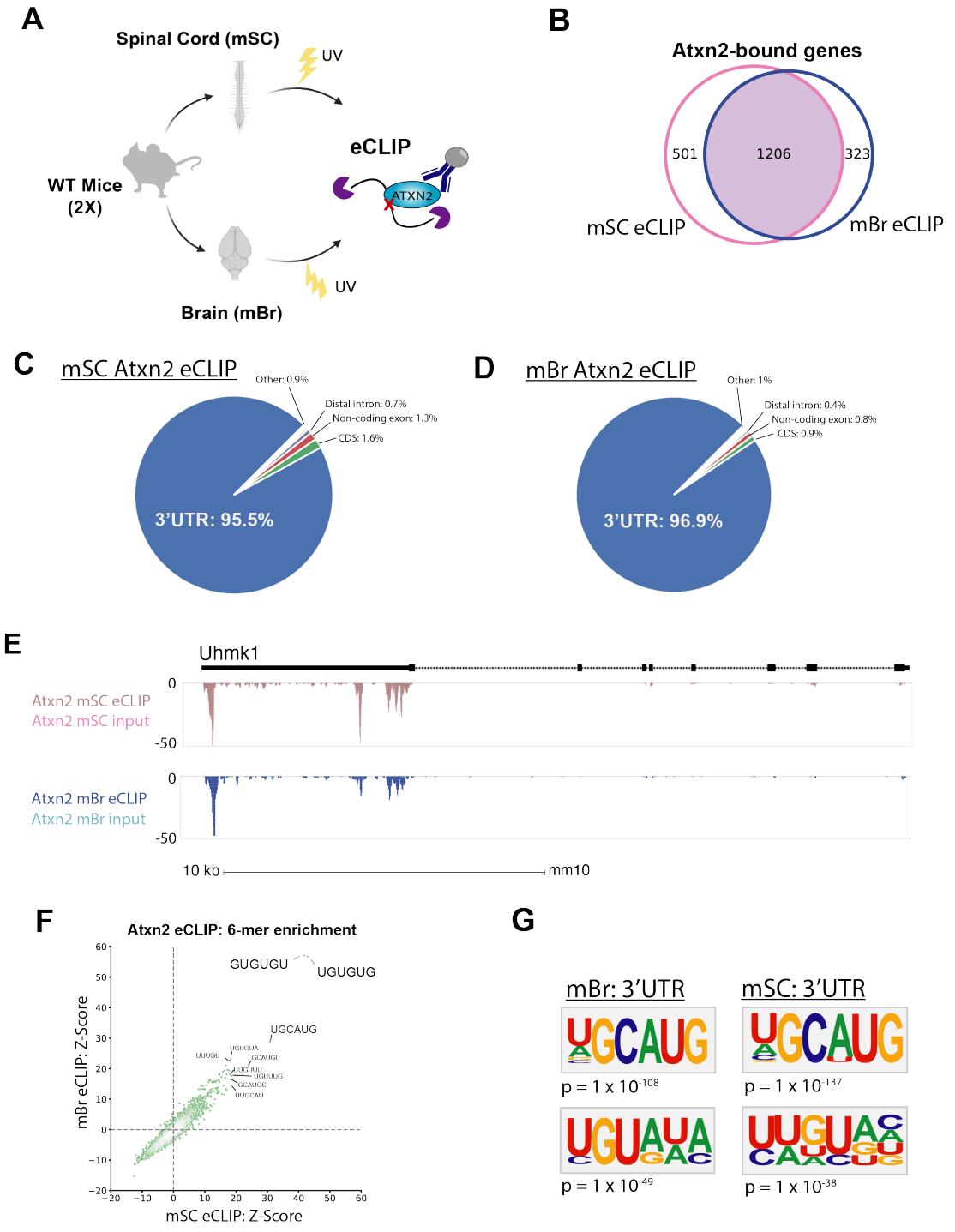


Figure 3.3. Comparative analyses of differential expression data across timepoints. Supplement to Figure 3.1. A. Number of upregulated and downregulated genes passing significant threshold cutoff (FDR < 0.05) for *Atxn2* KO vs. WT comparisons at log2(FC) cutoffs of 0, 0.4, and 1. The same analyses was performed for B. TDP-43 transgenic vs normal and C. *Atxn2* KO vs. WT in TDP-43 background. D-F. Comparison of shared downregulated and upregulated differentially expressed genes across timepoints for D. *Atxn2* KO vs. WT, E. TDP-43 transgenic vs normal, or F. *Atxn2* KO vs. WT in TDP-43 background.

Figure 3.4. Transcriptome-wide mapping of ataxin-2 in adult mouse brain and spinal cord using eCLIP. A. Experimental schematic for ataxin-2 eCLIP in adult mouse brain and spinal cord. Replicate (n=2) whole brain and spinal cord samples were taken from adult (8w, F) mice, UV-crosslinked, and subjected to eCLIP to map RNA-RBP interactions. B. Overlap of target genes significantly bound (IDR peaks with fold change > 8 and a $P < 0.001$) in brain (mBr) and spinal cord (mSC) for ataxin-2 eCLIP. C, D. Region-specific binding profiles of significant peaks in either mSC (C.) or mBr (D.). E. Representative browser tracks of mBr and mSC input and IP reads over the *Uhmk1* gene. Both eCLIP profiles showed heavy 3'UTR preference, as shown here. F. 6-mer enrichment (Z-scores) calculated for significant peaks. The top 6-mers identified (shown here) were all UG-containing. G. Top 2 HOMER consensus motifs for peaks in 3'UTR for both mSC and mBr eCLIP. Accompanying p-value from the motif enrichment analysis shown below.


A
Spinal Cord (mSC)
UV
eCLIP
WT Mice
(2X)
ATXN2
UV
Brain (mBr)
B
Atxn2-bound genes
501
1206
323
mSC eCLIP
mBr eCLIP
C
mSC Atxn2 eCLIP
Other: 0.9%
Distal intron: 0.7%
Non-coding exon: 1.3%
CDS: 1.6%
3'UTR: 95.5%
D
mBr Atxn2 eCLIP
Other: 1%
Distal intron: 0.4%
Non-coding exon: 0.8%
CDS: 0.9%
3'UTR: 96.9%
E
Uhmk1
Atxn2 mSC eCLIP
Atxn2 mSC input
Atxn2 mBr eCLIP
Atxn2 mBr input
0
-50
10 kb
mm10
F
Atxn2 eCLIP: 6-mer enrichment
GUGUGU
UGUGUG
UGCAUG
mBr eCLIP: Z-Score
mSC eCLIP: Z-Score
G
mBr: 3'UTR
mSC: 3'UTR
p = 1 x 10-108
p = 1 x 10-137
p = 1 x 10-49
p = 1 x 10-38

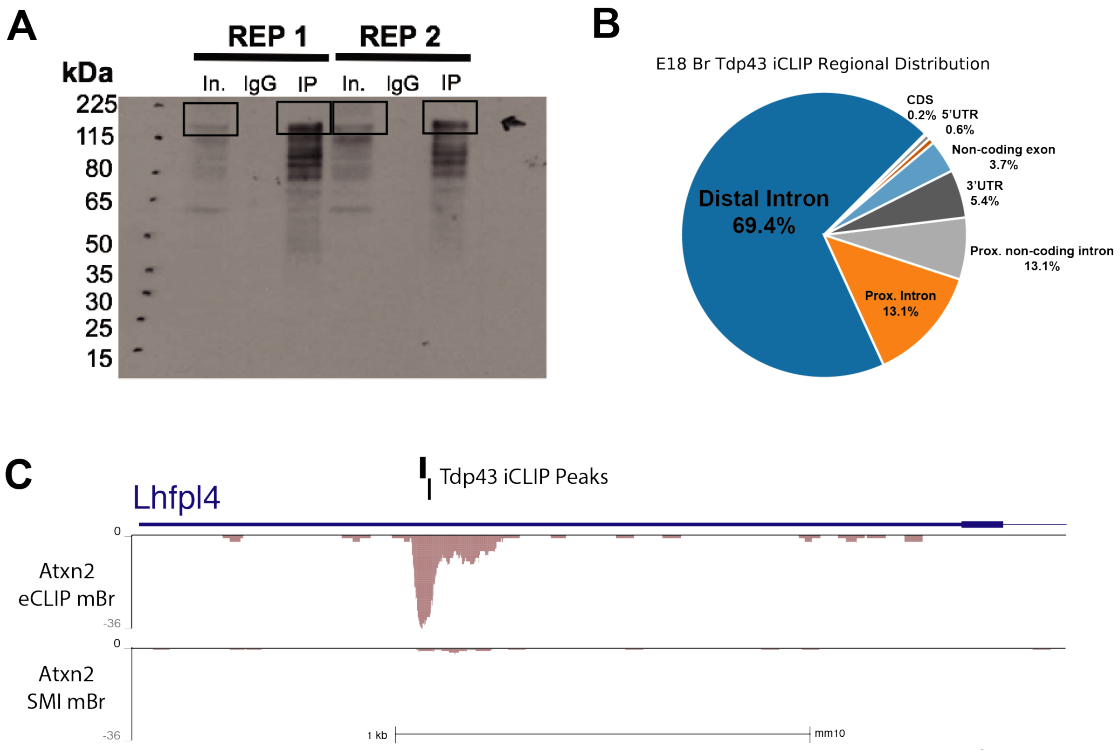
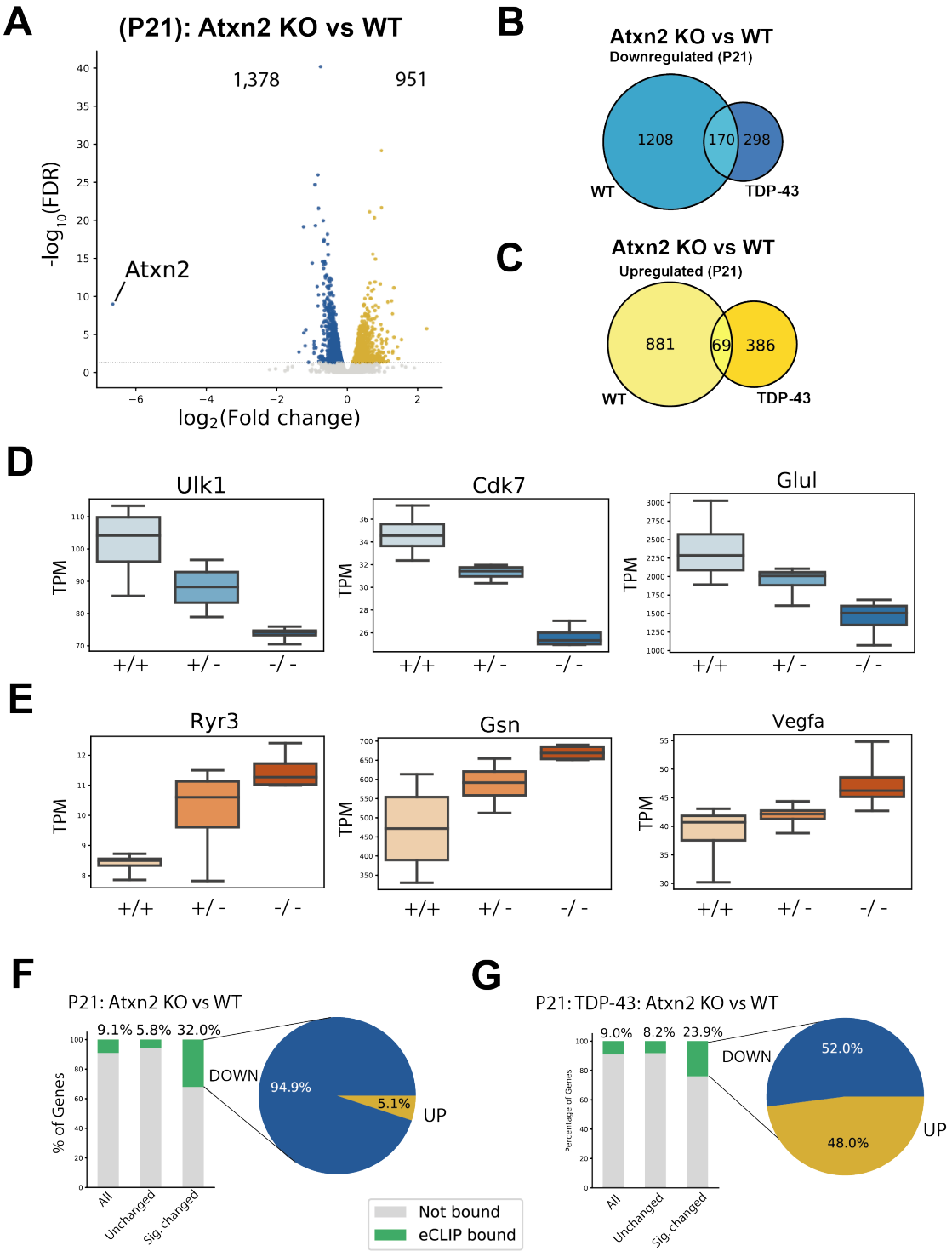


Figure 3.5. Figure 3.5. Isolation of ataxin-2 RNP complex and reanalysis of Tdp-43 iCLIP in mouse brain. Supplement to Figure 3.4. A. Western blot from immunoprecipitation (IP) of ataxin-2 from spinal cord in replicate. A 2% size-matched input and Rb IgG control are included also displayed for each replicate. Boxes indicate approximate cut size on the nitrocellulose membrane. B. Regional distribution of Tdp-43 iCLIP peaks detected in embryonic day 18 mouse brain (from Rogelj et al. 2012). C. Representative browser tracks of mBr ataxin-2 eCLIP IP and sized-matched input (SMI) overlapping Tdp-43 iCLIP peaks (black boxes) in the 3'UTR of *Lhfpl4*.

Figure 3.6. Ataxin-2 expression levels regulate mRNA abundance targets in mouse spinal cord. A. Volcano plots representing differentially expressed genes (FDR < 0.05) for Atxn2 KO mice versus WT. B. Venn diagrams showing directional overlap of Atxn2 KO responsive genes in normal and (WT) and disease (TDP-43) backgrounds at P21. Downregulated genes (top) are represented in blue and upregulated genes (bottom) are regulated in yellow. D, E Ataxin-2 dose-dependent expression effects for differentially expressed genes in Atxn2 KO TDP-43Tg/Tg mice for downregulated genes *Ulk1*, *Cdk7*, and *Glul* (D) and upregulated genes (E) genes *Ryr3*, *Gsn*, and *Vegfa* (WT = +/+, heterozygous knockout = +/-, homozygous knockout = -/-). F. Bar graphs depicting the proportion of bound genes (green) of total, unchanging, or differentially expressed genes in P21 Atxn2 KO mice (left). Pie chart (right) shows the fold change direction for eCLIP bound genes that are differentially expressed (downregulated, blue; upregulated, yellow). G. Bar graphs depicting the proportion of bound genes (green) of total, unchanging, or differentially expressed genes in P21 Atxn2 KO/TDP-43 mice (left). Pie chart (right) shows the fold change direction for differentially expressed genes bound by eCLIP.


A
(P21): Atxn2 KO vs WT
1,378
951
Atxn2
-log10(FDR)
log2(Fold change)
B
Atxn2 KO vs WT
Downregulated (P21)
1208
170
298
WT
TDP-43
C
Atxn2 KO vs WT
Upregulated (P21)
881
69
386
WT
TDP-43
D
Ulk1
Cdk7
Glul
TPM
+/+
+/ -
-/ -
E
Ryr3
Gsn
Vegfa
TPM
+/+
+/ -
-/ -
F
P21: Atxn2 KO vs WT
9.1% 5.8% 32.0%
% of Genes
DOWN
94.9%
5.1%
UP
All
Unchanged
Sig. changed
Not bound
eCLIP bound
G
P21: TDP-43: Atxn2 KO vs WT
9.0% 8.2% 23.9%
Percentage of Genes
DOWN
52.0%
48.0%
UP
All
Unchanged
Sig. changed

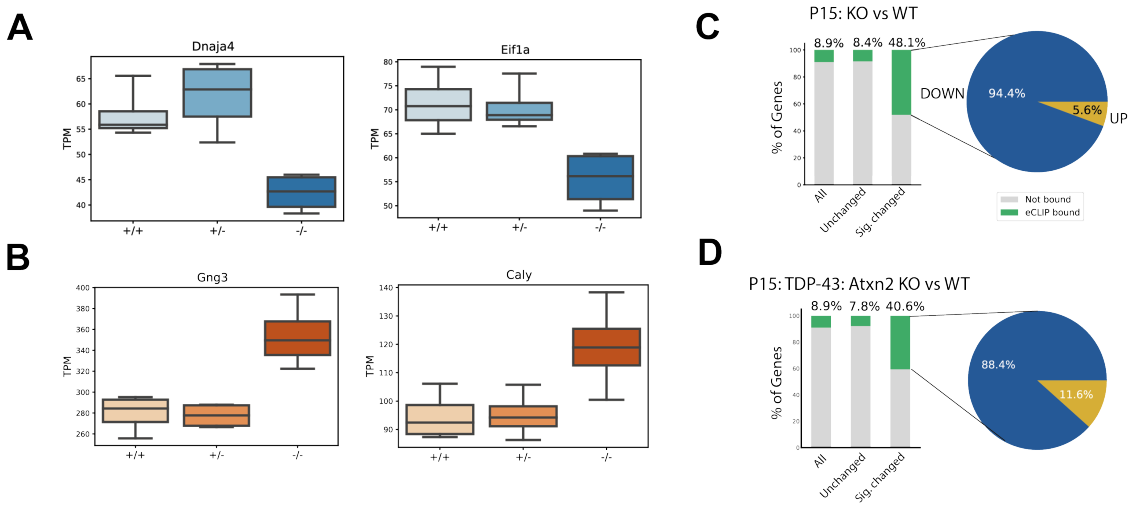


Figure 3.7. Differing effects of ataxin-2 expression on gene expression. Supplement to Figure 3.6. A, B. Examples of haplosufficient effects for differentially expressed genes in Atxn2 KO TDP-43$^{Tg/Tg}$ mice for downregulated genes *Dnaja4* and *Eif1a* (A) and upregulated genes *Gng3* and *Caly* (B). C, D.Bar graphs depicting the proportion of bound genes (green) of total, unchanging, or differentially expressed genes in P15 Atxn2 KO (C) or P15 KO/TDP-43 mice (D) with accompanying chart (right) displaying direction eCLIP bound genes.

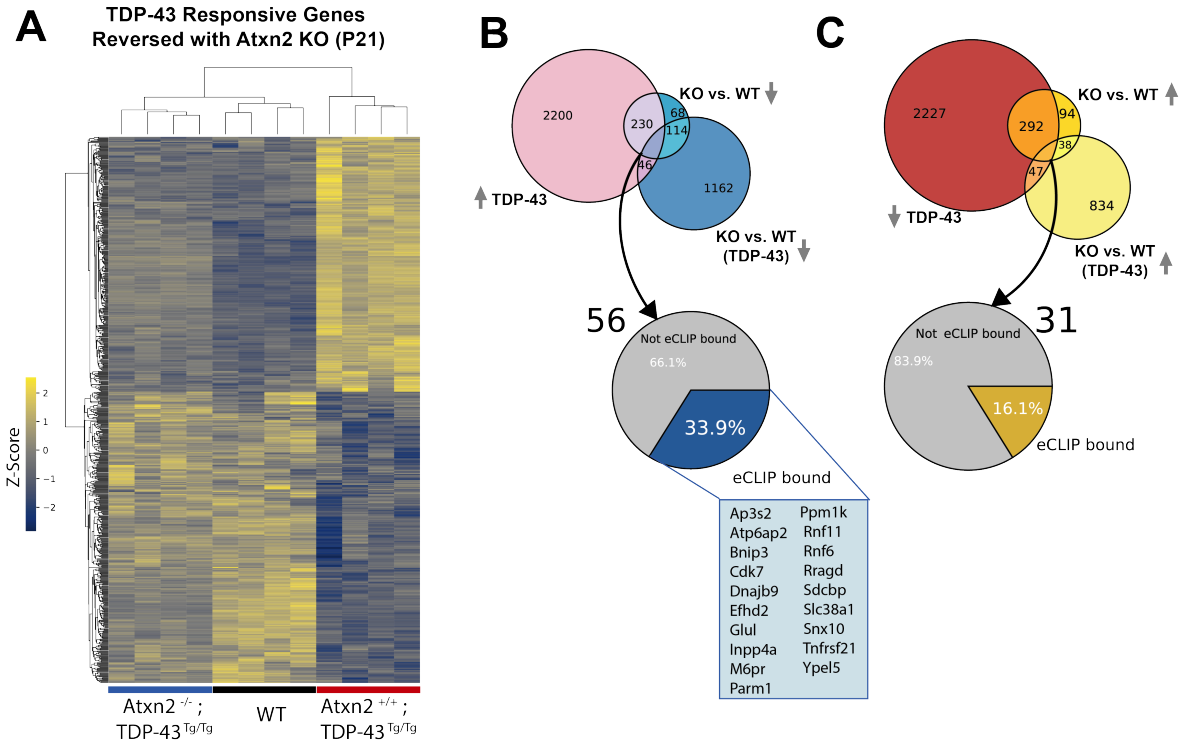


Figure 3.8. Atxn2 knockout restores gene function largely independent of RNA binding. A. Heatmap displaying relative expression (TPM) the 609 genes directionally changed following Atxn2 KO. Atxn2 KO/TDP-43 mice (blue) group with wildtype mice (black) and separate from TDP-43 mice (Tg) via hierarchical clustering. B. Overlap of genes upregulated in TDP-43 that are reversed with Atxn2 KO/TDP-43. (Inset) pie chart displaying common genes also downregulated in Atxn2 KO mice broken down by eCLIP binding status. Bound genes are displayed below. C. Overlap of genes downregulated in TDP-43 that are restored with Atxn2 KO/TDP-43. (Inset) pie chart displaying common genes upregulated in Atxn2 KO mice broken by eCLIP binding status.

References

Alami, N. H., Smith, R. B., Carrasco, M. A., Williams, L. A., Winborn, C. S., Han, S. S. W., Kiskinis, E., Winborn, B., Freibaum, B. D., Kanagaraj, A., Clare, A. J., Badders, N. M., Bilican, B., Chaum, E., Chandran, S., Shaw, C. E., Eggan, K. C., Maniatis, T. & Taylor, J. P. 2014. Axonal transport of TDP-43 mRNA granules is impaired by ALS-causing mutations. Neuron, 81, 536-543.

Barmada, S. J., Skibinski, G., Korb, E., Rao, E. J., Wu, J. Y. & Finkbeiner, S. 2010. Cytoplasmic mislocalization of TDP-43 is toxic to neurons and enhanced by a mutation associated with familial amyotrophic lateral sclerosis. J Neurosci, 30, 639-49.

Becker, L. A., Huang, B., Bieri, G., Ma, R., Knowles, D. A., Jafar-Nejad, P., Messing, J., Kim, H. J., Soriano, A., Auburger, G., Pulst, S. M., Taylor, J. P., Rigo, F. & Gitler, A. D. 2017. Therapeutic reduction of ataxin-2 extends lifespan and reduces pathology in TDP-43 mice. Nature, 544, 367-371.

Bomze, H. M., Bulsara, K. R., Iskandar, B. J., Caroni, P. & Skene, J. H. 2001. Spinal axon regeneration evoked by replacing two growth cone proteins in adult neurons. Nat Neurosci, 4, 38-43.

Briese, M., Saal-Bauernschubert, L., Luningschror, P., Moradi, M., Dombert, B., Surrey, V., Appenzeller, S., Deng, C., Jablonka, S. & Sendtner, M. 2020. Loss of Tdp-43 disrupts the axonal transcriptome of motoneurons accompanied by impaired axonal translation and mitochondria function. Acta Neuropathol Commun, 8, 116.

Brown, R. H. & Al-Chalabi, A. 2017. Amyotrophic Lateral Sclerosis. N Engl J Med, 377, 162-172. Byrne, S., Walsh, C., Lynch, C., Bede, P., Elamin, M., Kenna, K., Mclaughlin, R. & Hardiman, O. 2011. Rate of familial amyotrophic lateral sclerosis: a systematic review and meta-analysis. J Neurol Neurosurg Psychiatry, 82, 623-7.

Cahoy, J. D., Emery, B., Kaushal, A., Foo, L. C., Zamanian, J. L., Christopherson, K. S., Xing, Y., Lubischer, J. L., Krieg, P. A., Krupenko, S. A., Thompson, W. J. & Barres, B. A. 2008. A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function. J Neurosci, 28, 264-78.

Canet-Pons, J., Sen, N. E., Arsovic, A., Almaguer-Mederos, L. E., Halbach, M. V., Key, J., Doring, C., Kerksiek, A., Picchiarelli, G., Cassel, R., Rene, F., Dieterle, S., Fuchs, N. V., Konig, R., Dupuis, L., Lutjohann, D., Gispert, S. & Auburger, G. 2021. Atxn2-CAG100-KnockIn mouse spinal cord shows progressive TDP43 pathology associated with cholesterol biosynthesis suppression. Neurobiol Dis, 152, 105289.

Caroni, P., Aigner, L. & Schneider, C. 1997. Intrinsic neuronal determinants locally regulate

extrasynaptic and synaptic growth at the adult neuromuscular junction. J Cell Biol, 136, 679-92. Chiu, I. M., Morimoto, E. T., Goodarzi, H., Liao, J. T., O'keeffe, S., Phatnani, H. P., Muratet, M., Carroll, M. C., Levy, S., Tavazoie, S., Myers, R. M. & Maniatis, T. 2013. A neurodegeneration-specific gene-expression signature of acutely isolated microglia from an amyotrophic lateral sclerosis mouse model. Cell Rep, 4, 385-401.

D'erchia, A. M., Gallo, A., Manzari, C., Raho, S., Horner, D. S., Chiara, M., Valletti, A., Aiello, I., Mastropasqua, F., Ciaccia, L., Locatelli, F., Pisani, F., Nicchia, G. P., Svelto, M., Pesole, G. & Picardi, E. 2017. Massive transcriptome sequencing of human spinal cord tissues provides new insights into motor neuron degeneration in ALS. Sci Rep, 7, 10046.

Diaz-Garcia, S., Ko, V. I., Vazquez-Sanchez, S., Chia, R., Arogundade, O. A., Rodriguez, M. J., Traynor, B. J., Cleveland, D. & Ravits, J. 2021. Nuclear depletion of RNA-binding protein ELAVL3 (HuC) in sporadic and familial amyotrophic lateral sclerosis. Acta Neuropathol, 142, 985-1001. Elden, A. C., Kim, H. J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M. M., Padmanabhan, A., Clay-Falcone, D., Mccluskey, L., Elman, L., Juhr, D., Gruber, P. J., Rub, U., Auburger, G., Trojanowski, J. Q., Lee, V. M., Van Deerlin, V. M., Bonini, N. M. & Gitler, A. D. 2010. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature, 466, 1069-75.

Imbert, G., Saudou, F., Yvert, G., Devys, D., Trottier, Y., Garnier, J. M., Weber, C., Mandel, J. L., Cancel, G., Abbas, N., Durr, A., Didierjean, O., Stevanin, G., Agid, Y. & Brice, A. 1996. Cloning of the gene for spinocerebellar ataxia 2 reveals a locus with high sensitivity to expanded CAG/glutamine repeats. Nat Genet, 14, 285-91.

Kapeli, K., Pratt, G. A., Vu, A. Q., Hutt, K. R., Martinez, F. J., Sundararaman, B., Batra, R., Freese, P., Lambert, N. J., Huelga, S. C., Chun, S. J., Liang, T. Y., Chang, J., Donohue, J. P., Shiue, L., Zhang, J., Zhu, H., Cambi, F., Kasarskis, E., Hoon, S., Ares, M., Jr., Burge, C. B., Ravits, J., Rigo, F. & Yeo, G. W. 2016. Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses. Nat Commun, 7, 12143.

Kiehl, T. R., Nechiporuk, A., Figueroa, K. P., Keating, M. T., Huynh, D. P. & Pulst, S. M. 2006. Generation and characterization of Sca2 (ataxin-2) knockout mice. Biochem Biophys Res Commun, 339, 17-24.

Kiernan, M. C., Vucic, S., Cheah, B. C., Turner, M. R., Eisen, A., Hardiman, O., Burrell, J. R. & Zoing, M. C. 2011. Amyotrophic lateral sclerosis. Lancet, 377, 942-55.

Kim, H. J., Kim, N. C., Wang, Y. D., Scarborough, E. A., Moore, J., Diaz, Z., Maclea, K. S., Freibaum, B., Li, S., Molliex, A., Kanagaraj, A. P., Carter, R., Boylan, K. B., Wojtas, A. M., Rademakers, R., Pinkus, J. L., Greenberg, S. A., Trojanowski, J. Q., Traynor, B. J., Smith, B. N., Topp, S., Gkazi, A. S., Miller, J., Shaw, C. E., Kottlors, M., Kirschner, J., Pestronk, A., Li, Y. R., Ford, A. F., Gitler, A. D., Benatar, M., King, O. D., Kimonis, V. E., Ross, E. D., Weihl, C. C., Shorter, J. & Taylor, J. P. 2013. Mutations in prion-like domains in hnRNPA2B1 and hnRNPA1

cause multisystem proteinopathy and ALS. Nature, 495, 467-73.

Klim, J. R., Williams, L. A., Limone, F., Guerra San Juan, I., Davis-Dusenbery, B. N., Mordes, D. A., Burberry, A., Steinbaugh, M. J., Gamage, K. K., Kirchner, R., Moccia, R., Cassel, S. H., Chen, K., Wainger, B. J., Woolf, C. J. & Eggan, K. 2019. ALS-implicated protein TDP-43 sustains levels of STMN2, a mediator of motor neuron growth and repair. Nat Neurosci, 22, 167-179.

Krach, F., Batra, R., Wheeler, E. C., Vu, A. Q., Wang, R., Hutt, K., Rabin, S. J., Baughn, M. W., Libby, R. T., Diaz-Garcia, S., Stauffer, J., Pirie, E., Saberi, S., Rodriguez, M., Madrigal, A. A., Kohl, Z., Winner, B., Yeo, G. W. & Ravits, J. 2018. Transcriptome-pathology correlation identifies interplay between TDP-43 and the expression of its kinase CK1E in sporadic ALS. Acta Neuropathol, 136, 405-423.

Kuo, P. H., Doudeva, L. G., Wang, Y. T., Shen, C. K. & Yuan, H. S. 2009. Structural insights into TDP-43 in nucleic-acid binding and domain interactions. Nucleic Acids Res, 37, 1799-808.

Kwiatkowski, T. J., Jr., Bosco, D. A., Leclerc, A. L., Tamrazian, E., Vanderburg, C. R., Russ, C., Davis, A., Gilchrist, J., Kasarskis, E. J., Munsat, T., Valdmanis, P., Rouleau, G. A., Hosler, B. A., Cortelli, P., De Jong, P. J., Yoshinaga, Y., Haines, J. L., Pericak-Vance, M. A., Yan, J., Ticozzi, N., Siddique, T., Mckenna-Yasek, D., Sapp, P. C., Horvitz, H. R., Landers, J. E. & Brown, R. H., Jr. 2009. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. Science, 323, 1205-8.

Laffita-Mesa, J. M., Paucar, M. & Svenningsson, P. 2021. Ataxin-2 gene: a powerful modulator of neurological disorders. Curr Opin Neurol, 34, 578-588.

Lambrechts, D., Storkebaum, E., Morimoto, M., Del-Favero, J., Desmet, F., Marklund, S. L., Wyns, S., Thijs, V., Andersson, J., Van Marion, I., Al-Chalabi, A., Bornes, S., Musson, R., Hansen, V., Beckman, L., Adolfsson, R., Pall, H. S., Prats, H., Vermeire, S., Rutgeerts, P., Katayama, S., Awata, T., Leigh, N., Lang-Lazdunski, L., Dewerchin, M., Shaw, C., Moons, L., Vlietinck, R., Morrison, K. E., Robberecht, W., Van Broeckhoven, C., Collen, D., Andersen, P. M. & Carmeliet, P. 2003. VEGF is a modifier of amyotrophic lateral sclerosis in mice and humans and protects motoneurons against ischemic death. Nat Genet, 34, 383-94.

Lastres-Becker, I., Rub, U. & Auburger, G. 2008. Spinocerebellar ataxia 2 (SCA2). Cerebellum, 7, 115-24.

Ling, J. P., Pletnikova, O., Troncoso, J. C. & Wong, P. C. 2015. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. Science, 349, 650-5.

Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W., Conboy, J. G. & Yeo, G. W. 2013. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol, 20, 1434-42.

Ma, X. R., Prudencio, M., Koike, Y., Vatsavayai, S. C., Kim, G., Harbinski, F., Briner, A., Rodriguez, C. M., Guo, C., Akiyama, T., Schmidt, H. B., Cummings, B. B., Wyatt, D. W., Kurylo, K., Miller, G., Mekhoubad, S., Sallee, N., Mekonnen, G., Ganser, L., Rubien, J. D., Jansen-West, K., Cook, C. N., Pickles, S., Oskarsson, B., Graff-Radford, N. R., Boeve, B. F., Knopman, D. S., Petersen, R. C., Dickson, D. W., Shorter, J., Myong, S., Green, E. M., Seeley, W. W., Petrucelli, L. & Gitler, A. D. 2022. TDP-43 represses cryptic exon inclusion in the FTD-ALS gene UNC13A. Nature, 603, 124-130.

Mackenzie, I. R., Bigio, E. H., Ince, P. G., Geser, F., Neumann, M., Cairns, N. J., Kwong, L. K., Forman, M. S., Ravits, J., Stewart, H., Eisen, A., Mcclusky, L., Kretzschmar, H. A., Monoranu, C. M., Highley, J. R., Kirby, J., Siddique, T., Shaw, P. J., Lee, V. M. & Trojanowski, J. Q. 2007. Pathological TDP-43 distinguishes sporadic amyotrophic lateral sclerosis from amyotrophic lateral sclerosis with SOD1 mutations. Ann Neurol, 61, 427-34.

Martinez, J. C., Randolph, L. K., Iascone, D. M., Pernice, H. F., Polleux, F. & Hengst, U. 2019. Pum2 Shapes the Transcriptome in Developing Axons through Retention of Target mRNAs in the Cell Body. Neuron, 104, 931-946 e5.

Melamed, Z., Lopez-Erauskin, J., Baughn, M. W., Zhang, O., Drenner, K., Sun, Y., Freyermuth, F., Mcmahon, M. A., Beccari, M. S., Artates, J. W., Ohkubo, T., Rodriguez, M., Lin, N., Wu, D., Bennett, C. F., Rigo, F., Da Cruz, S., Ravits, J., Lagier-Tourenne, C. & Cleveland, D. W. 2019. Premature polyadenylation-mediated loss of stathmin-2 is a hallmark of TDP-43-dependent neurodegeneration. Nat Neurosci, 22, 180-190.

Neumann, M., Bentmann, E., Dormann, D., Jawaid, A., Dejesus-Hernandez, M., Ansorge, O., Roeber, S., Kretzschmar, H. A., Munoz, D. G., Kusaka, H., Yokota, O., Ang, L. C., Bilbao, J., Rademakers, R., Haass, C. & Mackenzie, I. R. 2011. FET proteins TAF15 and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. Brain, 134, 2595-609.

Neumann, M., Sampathu, D. M., Kwong, L. K., Truax, A. C., Micsenyi, M. C., Chou, T. T., Bruce, J., Schuck, T., Grossman, M., Clark, C. M., Mccluskey, L. F., Miller, B. L., Masliah, E., Mackenzie, I. R., Feldman, H., Feiden, W., Kretzschmar, H. A., Trojanowski, J. Q. & Lee, V. M. 2006. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. Science, 314, 130-3.

Pasinelli, P. & Brown, R. H. 2006. Molecular biology of amyotrophic lateral sclerosis: insights from genetics. Nat Rev Neurosci, 7, 710-23.

Phatnani, H. P., Guarnieri, P., Friedman, B. A., Carrasco, M. A., Muratet, M., O'keeffe, S., Nwakeze, C., Pauli-Behn, F., Newberry, K. M., Meadows, S. K., Tapia, J. C., Myers, R. M. & Maniatis, T. 2013. Intricate interplay between astrocytes and motor neurons in ALS. Proc Natl Acad Sci U S A, 110, E756-65.

Polymenidou, M., Lagier-Tourenne, C., Hutt, K. R., Huelga, S. C., Moran, J., Liang, T. Y., Ling, S. C., Sun, E., Wancewicz, E., Mazur, C., Kordasiewicz, H., Sedaghat, Y., Donohue, J. P., Shiue, L., Bennett, C. F., Yeo, G. W. & Cleveland, D. W. 2011. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci, 14, 459-68.

Pulst, S. M., Nechiporuk, A., Nechiporuk, T., Gispert, S., Chen, X. N., Lopes-Cendes, I., Pearlman, S., Starkman, S., Orozco-Diaz, G., Lunkes, A., Dejong, P., Rouleau, G. A., Auburger, G., Korenberg, J. R., Figueroa, C. & Sahba, S. 1996. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. Nat Genet, 14, 269-76.

Rogelj, B., Easton, L. E., Bogu, G. K., Stanton, L. W., Rot, G., Curk, T., Zupan, B., Sugimoto, Y., Modic, M., Haberman, N., Tollervey, J., Fujii, R., Takumi, T., Shaw, C. E. & Ule, J. 2012. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. Sci Rep, 2, 603.

Rudnick, N. D., Griffey, C. J., Guarnieri, P., Gerbino, V., Wang, X., Piersaint, J. A., Tapia, J. C., Rich, M. M. & Maniatis, T. 2017. Distinct roles for motor neuron autophagy early and late in the SOD1(G93A) mouse model of ALS. Proc Natl Acad Sci U S A, 114, E8294-E8303.

Scoles, D. R., Dansithong, W., Pflieger, L. T., Paul, S., Gandelman, M., Figueroa, K. P., Rigo, F., Bennett, C. F. & Pulst, S. M. 2020. ALS-associated genes in SCA2 mouse spinal cord transcriptomes. Hum Mol Genet, 29, 1658-1672.

Sen, N. E., Canet-Pons, J., Halbach, M. V., Arsovic, A., Pilatus, U., Chae, W. H., Kaya, Z. E., Seidel, K., Rollmann, E., Mittelbronn, M., Meierhofer, D., De Zeeuw, C. I., Bosman, L. W. J., Gispert, S. & Auburger, G. 2019. Generation of an Atxn2-CAG100 knock-in mouse reveals N-acetylaspartate production deficit due to early Nat8l dysregulation. Neurobiol Dis, 132, 104559.

Shao, J. & Diamond, M. I. 2007. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. Hum Mol Genet, 16 Spec No. 2, R115-23.

Shibata, H., Huynh, D. P. & Pulst, S. M. 2000. A novel protein with RNA-binding motifs interacts with ataxin-2. Hum Mol Genet, 9, 1303-13.

Singh, A., Hulsmeier, J., Kandi, A. R., Pothapragada, S. S., Hillebrand, J., Petrauskas, A., Agrawal, K., Rt, K., Thiagarajan, D., Jayaprakashappa, D., Vijayraghavan, K., Ramaswami, M. & Bakthavachalu, B. 2021. Antagonistic roles for Ataxin-2 structured and disordered domains in RNP condensation. Elife, 10.

Smialek, M. J., Ilaslan, E., Sajek, M. P. & Jaruzelska, J. 2021. Role of PUM RNA-Binding Proteins in Cancer. Cancers (Basel), 13.

Sreedharan, J., Blair, I. P., Tripathi, V. B., Hu, X., Vance, C., Rogelj, B., Ackerley, S., Durnall, J. C., Williams, K. L., Buratti, E., Baralle, F., De Belleroche, J., Mitchell, J. D., Leigh, P. N., Al-Chalabi, A., Miller, C. C., Nicholson, G. & Shaw, C. E. 2008. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. Science, 319, 1668-72.

Taylor, J. P., Brown, R. H., Jr. & Cleveland, D. W. 2016. Decoding ALS: from genes to mechanism. Nature, 539, 197-206.

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., Konig, J., Hortobagyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E. & Ule, J. 2011. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. Nat Neurosci, 14, 452-8.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 13, 508-14.

Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X. D., Graveley, B. R. & Yeo, G. W. 2020. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol, 21, 90.

Vera, M., Pani, B., Griffiths, L. A., Muchardt, C., Abbott, C. M., Singer, R. H. & Nudler, E. 2014. The translation elongation factor eEF1A1 couples transcription to translation during heat shock response. Elife, 3, e03164.

Walker, A. K., Spiller, K. J., Ge, G., Zheng, A., Xu, Y., Zhou, M., Tripathy, K., Kwong, L. K., Trojanowski, J. Q. & Lee, V. M. 2015. Functional recovery in new mouse models of ALS/FTLD after clearance of pathological cytoplasmic TDP-43. Acta Neuropathol, 130, 643-60.

Wils, H., Kleinberger, G., Janssens, J., Pereson, S., Joris, G., Cuijt, I., Smits, V., Ceuterick-De Groote, C., Van Broeckhoven, C. & Kumar-Singh, S. 2010. TDP-43 transgenic mice develop spastic paralysis and neuronal inclusions characteristic of ALS and frontotemporal lobar degeneration. Proc Natl Acad Sci U S A, 107, 3858-63.

Yang, Q., Gilmartin, G. M. & Doublie, S. 2010. Structural basis of UGUA recognition by the Nudix protein CFI(m)25 and implications for a regulatory role in mRNA 3' processing. Proc Natl Acad Sci U S A, 107, 10062-7.

Yang, Y. S., Kato, M., Wu, X., Litsios, A., Sutter, B. M., Wang, Y., Hsu, C. H., Wood, N. E., Lemoff, A., Mirzaei, H., Heinemann, M. & Tu, B. P. 2019. Yeast Ataxin-2 Forms an Intracellular Condensate Required for the Inhibition of TORC1 Signaling during Respiratory Growth. Cell, 177, 697-710 e17.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X. D. & Gage, F. H. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. Nat Struct Mol Biol, 16, 130-7.

Yokoshi, M., Li, Q., Yamamoto, M., Okada, H., Suzuki, Y. & Kawahara, Y. 2014. Direct binding of Ataxin-2 to distinct elements in 3' UTRs promotes mRNA stability and protein expression. Mol Cell, 55, 186-98.

Yu, H., Lu, S., Gasior, K., Singh, D., Vazquez-Sanchez, S., Tapia, O., Toprani, D., Beccari, M. S., Yates, J. R., 3rd, Da Cruz, S., Newby, J. M., Lafarga, M., Gladfelter, A. S., Villa, E. & Cleveland, D. W. 2021. HSP70 chaperones RNA-free TDP-43 into anisotropic intranuclear liquid spherical shells. Science, 371.

Yu, Z., Zhu, Y., Chen-Plotkin, A. S., Clay-Falcone, D., Mccluskey, L., Elman, L., Kalb, R. G., Trojanowski, J. Q., Lee, V. M., Van Deerlin, V. M., Gitler, A. D. & Bonini, N. M. 2011. PolyQ repeat expansions in ATXN2 associated with ALS are CAA interrupted repeats. PLoS One, 6, e17951.

Chapter 4

Human models to investigate ataxin-2 function in motor neurons

4.1 Abstract

Ataxin-2 (ATXN2) is a widely expressed RBPs implicated in many neuronal diseases including ALS. Although ataxin-2 has shown some promise as a potential therapeutic target in preclinical mouse models, efforts to study its function in human neurons are still missing. Here, we briefly discuss efforts to develop resources to further investigate the role of human ataxin-2 in motor neurons. Using eCLIP, we identify over 5,000 genes bound by ATXN2 in human induced pluripotent stem cell (iPSC)-derived motor neurons. This profile matched our previous mouse CNS data, featuring prominent 3'UTR binding to UG-rich target sequences. We also generated several means of ATXN2 depletion using CRISPR/Cas9 genomic engineering and short hairpin-RNA knockdown, finding 77 commonly dysregulated genes that are largely independent of RNA-binding. Some of these candidate genes are involved in the p53 pathway and may play a role in suppressing cellular apoptosis in later stages of disease. Lastly, we detail an isogenic cell modeling strategy to study ataxin-2 function in the context of specific lengths of polyQ mutations. Overall, we believe this compendium of resources will be valuable in identifying novel disease-relevant pathways that potentially serve as targetable avenues for therapy and provide a broadly applicable disease modeling strategy to investigate the direct genetic influence of other repeat expansion mutations in future work.

4.2 Introduction

ATXN2 is an RBP that was first characterized as an ALS-associated gene through a genetic screen for modifiers of TDP-43 toxicity revealing that ataxin-2 expression promotes cell death (Elden et al., 2010). This synergistic relationship is conserved across multiple lifeforms, as studies have shown that co-expression of ataxin-2 with TDP-43 results in dose-dependent neuronal degeneration (Elden et al., 2010, Kim et al., 2014, Becker et al., 2017). Due to the prominence of this deleterious interaction, and because CAG trinucleotide expansion mutations within the polyQ tract of the ATXN2 gene have long been known to cause Purkinje cell degenerative in spinocerebellar ataxia 2 (SCA2), associations between CAG-repeat length and ALS have also been investigated. While CAG length for unaffected individuals is between 22 and 23, several large, internationally sampled ALS patient cohorts have revealed a significant correlation between intermediate length polyQ repeats (27-34 CAG) and ALS onset, with the most significant association being observed with a lower threshold of 29-30 CAGs (Elden et al., 2010, Lee et al., 2011, Chen et al., 2011, Van Damme et al., 2011, Tavares de Andrade et al., 2018). In addition to genetic linkage, ataxin-2 has been shown to directly or indirectly interact with major ALS-associated proteins like TDP-43 (Elden et al., 2010, Hart and Gitler, 2012), FUS (Farg et al., 2013), and C9orf79 (Sellier et al., 2016). Furthermore, intermediate-length mutations strengthen these interactions and result in deleterious effects such as increased pTDP-43 aggregation, caspase 3 cleavage, and cellular toxicity, further supported by histopathological data from ALS patient motor neurons revealing distinct pathology of cytoplasmic pTDP-43 aggregates in ALS cases of intermediate-length polyQ mutations (Hart and Gitler, 2012, Hart et al., 2012). Although these models indicate that it may act as simply a modifier gene to exacerbate ALS pathogenicity, it is unknown whether mutant ataxin-2 itself affects neuronal function and contributes to disease in through a more direct mechanism. It is therefore highly desirable to investigate the direct role of ATXN2 with genetically modifiable human models. Here, we detail efforts to characterize the functional landscape of ATXN2 in lower motor neurons, as well as

mimic disease status through introduction of ALS-associated polyQ mutations. Here, we detail efforts to characterize the functional landscape of ATXN2 in lower motor neurons, as well as mimic disease status through introduction of ALS-associated polyQ mutations.

4.3 Results

4.3.1 Uncovering ATXN2 RNA binding sites in human iPSC-MNs

To establish the first human transcriptome-wide binding map in neurons, we differentiated mature motor neurons (MN) from human induced pluripotent stem cells (iPSC) and performed ATXN2 eCLIP in replicate (Figure 4.1A). This yielded 19,484 reproducible peaks across 5,474 genes enriched over size-matched input, with the majority of peaks (81.5%) falling within 3'UTR of target genes (Figure 4.1B, C). Although not quite as prominent as we had seen in mouse brain and spinal cord, with approximately 13% of peaks falling into CDS as well, this matched our general expectation. We also see strong preference for UG-rich sequences in our significant peaks, with 'UGUANA' emerging as our top enriched motif in all and 3'UTR focused peaks (Figure 4.1D). This also falls in line with our mouse data, which observed strong UG-bias, as well as an enrichment for the 'UGUA' sequence motif. We were also able to the RBFOX consensus 'UGCAUG' in the 3'UTR region of target genes, though enrichment was not as strong as seen in mouse tissues. As a comparison to bulk tissue, we actually see a fair amount of overlap between human motor neurons eCLIP targets and those found in mouse spinal cord, where approximately 57% of target genes overlapping with MN eCLIP data (Figure 4.1E-G). Thus we are able to highlight not only the ability to recover motor-neuron enriched binding events in our bulk mouse tissue, but also an additional level of overlap between our model systems.

4.3.2 Orthogonal methods of ATXN2 depletion reveals consensus gene perturbations in motor neurons

We next wished to determine downstream target genes affected by ataxin-2 depletion in motor neurons. Previous personal attempts to perturb ataxin-2 in cell culture, including antisense

oligonucleotides (ASOs) or CRISPR interference (CRISPRi) proved to be inefficient for reliable ataxin-2 knockout in motor neurons. Seeing that deletion of ataxin-2 leads to viable and relatively healthy mice (Kiehl et al., 2006, Becker et al., 2017), we decided to delete ataxin-2 in iPSCs using CRISPR/Cas9 prior to differentiation into motor neurons and performing RNA-seq (Figure 4.1A, top). To discern bona fide gene expression changes from potential off-target changes in differentiation potential that knockout might cause, we also performed short-hairpin RNA (shRNA)-mediated knockdown of ATXN2 in mature motor neurons in late stages (d30+) of differentiation (Figure 4.1A, bottom). Both methods were efficient in achieving target-specific depletion of ataxin-2, as validated by western blot at the motor neuron stage (Figure 4.2B, C). Comparative RNA-seq to each respective control condition reveals several hundred differentially expressed genes ($|\text{log2FC} > 1|$ and $\text{FDR} < 0.05$, as determined by DESeq2). We noticed considerably more differentially expressed genes in our CRISPR knockout lines compared to controls (Figure 4.2D; 426 upregulated, 795 downregulated) than we did with shRNA knockdown (Figure 4.2E; 203 upregulated, 206 downregulated). Though this is likely due to secondary effects that accumulate downstream of gene knockout throughout the course of differentiation. Indeed, we find 77 common, high-confidence genes between the two depletion strategies (Figure 4.2F), a subset of which are involved in the p53 transcrtiptional program, including p53 (TP53) itself. The p53 signaling pathway is typically activated downstream of genotoxic stress and mediator of cellular apoptosis and was recently found to be a central regulator driving neuronal death in models of C9orf72 ALS (Maor-Nof et al., 2021). There has been increasing interest in studying the causal link between pathogenic TDP-43 and p53-mediated cellular apoptosis (Vogt et al., 2018, Mitra et al., 2019), and so investigating the suppression of these targets will be of considerable interest. Additionally, and perhaps non-intuitively, we noticed the consistent downregulation of the chaperone protein *BAG3* as well as the heat-shock response protein *HSPB1*, both of which are typically associated with counteracting protein aggregation proteins through selective autophagy (Casarotto et al., 2022) and promoting TDP-43 disassembly (Lu et al., 2021), respectively. Whether these genes are dysregulated as a sign of downstream cellular vulnerability

or an indication of decreased stressed levels in neurons remains unclear. Thus, additional studies will be focused on dissecting the interaction between ATXN2 and these relevant genes.

4.3.3 Modeling specific-length polyQ expansion mutations in isogenic human stem cell models

Although there has been much effort dedicated to investigating the role that ATXN2 depletion plays on neuronal vulnerability, there is little known about the role of polyQ mutations on disease state. The ATXN2 gene encodes a polyQ tract that normally contains between 22-23 copies of CAG/CAA trinucleotide repeats. Sequence expansion mutations in this repetitive region are associated with the diseases spinocerebellar ataxia (SCA) 2 and ALS, with intermediate length (27-34X) CAG-repeats correlating with disease in 1-5% of ALS cases. However, since ATXN2 intermediate polyQ expansions often co-occur with ALS onset, it is difficult to delineate causative effects from mere correlation. We therefore sought to "mimic" polyQ mutations as they manifest in ALS in an isogenic fashion. To do so, we generated a strategy to introduce specific lengths of CAG repeats into the endogenous ATXN2 locus using CRISPR/Cas9 genomic engineering. By cloning custom-lengths of CAG sequences into a DNA donor vector comprised of homology arms matching the sequence of exon 1 of the ATXN2 gene, we could co-deliver this vector with pre-assembled Cas9 protein paired with a gRNA targeting an upstream sequence, this would initiate localized double-strand DNA cuts, which could then be "repaired" using homology-directed recombination using the donor vector as a template. This would effectively "knock-in" our desired length of CAG expansions into the ATXN2 locus in place of the normal polyQ sequence (Figure 4.3). To model intermediate-length repeat expansions that occur in ALS patients, we specifically chose 31Q donor vectors comprised of CAG sequences interrupted by CAA sequences (Figure 4.3). Both of these codons are translated to glutamine, and only these orientations were found to be present in ATXN2-expanded ALS patients (Yu et al., 2011), whereas longer, uninterrupted CAG sequences are typical with SCA2. Using the healthy donor CV-B iPSC background, we were able to successfully generate heterozygous 31Q knock-in

mutations into the endogenous ATXN2 locus without repeat-length loss or shrinkage in a lesion-free manner (Figure 4.4A). Importantly, we were able to isolate clones that expressed both short and long length alleles at the RNA level (Figure 4.4B) without seemingly impairing protein synthesis throughout differentiation into motor neurons (Figure 4.4C). Excitingly, this strategy also seems to be broadly applicable to differing lengths of CAG-expansions, as we were also able to use this strategy to knock-in uninterrupted 39Q expansion lengths that could be found in SCA2 patients (Figure 4.4D) (Lastres-Becker et al., 2008). Although still in relatively early phases, these isogenic models offer tremendous potential to investigate the individual, disease-associated contributions of specific-lengths of ATXN2 polyQ expansions in a common genetic background.

4.4 Materials and Methods

4.4.1 Induced pluripotent stem cell (iPSC) culture

The CV-B iPSC line, originating from Craig Venter, is publicly available and has been previously described (Gore et al.). Cells were cultured on Matrigel-coated plates using mTeSR1 or mTESR plus at 37^{o}C with 5% CO2. The media was exchanged every 24 hours. When the iPSC reach about 70-90% confluency, the cells were split using Versene or ReLeSR and plated in a 1:10 ratio with mTeSR1 or mTESR plus. Cells were monitored for pluripotency as well as normal karyotype, and regularly tested for mycoplasma contamination.

4.4.2 Motor neuron (MN) differentiation

A dual-SMAD inhibition-based protocol (Chambers et al., 2009) was used to generate MNs described previously (Markmiller et al., 2018). Briefly, iPSC were dissociated into single cells using Accutase. After cell counting, 0.208 x 10^6 cells per cm2 were plated on a Matrigel-coated well with mTESR1 or mTESR plus supplemented with 5μM RI. When cells reached 95% confluency, cells were treated with a basal-media, further referred as N2B27 (DMEM/F12, 0.5xN2, 0.5x B27, 100μM ascorbic acid, 1x Pen/Strep), was used for culture and dilution of

compounds. From day one to day six, the cells were incubated with N2B27 supplemented with 1μM Dorsomorphin (Dorso), 10μM SB431542 (SB), 3μM CHIR99021 (CHIR). The media was exchanged on daily basis. From day six to day 15 the cells were incubated with N2B27 with Dorso, SB, RI and, in addition, 1.5μM retinoic acid (RA) and 200nM of Smoothed agonist (SAG). When the cells reached the stage of MN progenitors (MNPs, day 15 of differentiation), the MNPs were dissociated into single cells using Accutase and cells were either frozen, differentiation was continued directly, or cells were optionally expanded in motor neuron progenitor (MNP) medium as described in (Du et al., 2015) for up to 5 passages. Following dissociation 1x10^6 cells/ml were plated on PDL/Lam-coated plates in N2B27 with RA, SAG, RI, and 2ng/ml of brain-derived neurotrophic factor (BDNF), glial-derived neurotrophic factor (GDNF), and ciliary neurotrophic factor (CNTF), respectively. For further culture, the media containing RA, SAG, 2μM RI and NFs was exchanged completely every other day until day 22. On day 22 of MN differentiation, the media was switched to N2B17 with NFs, 2μM RI and the γ-secretase inhibitor DAPT (2μM). On day 24, the cells were fed with the same media again. From day 25 on the cell were cultured until day 30 solely with N2B27 supplemented with NFs and 2μM RI. Mature motor neurons were harvested for eCLIP and RNA-seq at this time.

4.4.3 Lentivirus production and infection of iPSC-derived MNs

Lentivirus was produced by transfecting a confluent 15cm^2 dish of HEK293T cells with 8 μg pLKO.1 shRNA containing either a scrambled control sequence (CAACAAGATGAAGAG-CACCAACTCNAGTTGGTGCTCTTCATCTTGTTG) or an ATXN2-targeting sequence (GC-CTCAGTCTACGATTTCTTTCTCGAGAAAGAAATCGTAGACTGAGGC), 6 μg PSPAX2, and 4 μg pMD2.G plasmids using Lipofectamine 3000 (Thermo Fisher) in OPTI-MEM. The transfection mix was added to cell culture medium (DMEM + 10% FBS) directly and media was replaced after 6 hours. Virus-containing media was collected 48 and 72 hours following transfection, and was concentrated 100X using lenti-X concentrator solution (Takara). The viral pellet was resuspended in DMEM/F12 base media and aliquoted and stored at -80^{o}C.

To infect of motor neurons, wildtype CV-B motor neurons were split with Accutase on day 27, 2 days following DAPT withdrawal. 3 million cells were seeded per well onto a laminin-coated 12 well plate (4 wells per shRNA). On day 28, each well was infected with 1:100 shRNA lentivirus, with media being exchanged after 24 hours. After 72 hours, the cells were hit with a second dose of lentivirus and incubated for an additional 72 hours. At day 35, approximately 7 days after initial treatment, cells were harvested with either with Trizol for RNA extraction (3X) or RIPA lysis buffer for Western Blot validation of knockdown.

4.4.4 Generation of ATXN2 knockout iPSC

Genetic knockouts were performed in CV-B line. Cells were dissociated with Accutase and passed through a 40 μm strainer to achieve single-cell dissociation. Cells were spun down at 200xg for 5 minutes and resuspended in mTeSR plus supplemented with 10uM ROCK-inhibitor (RI). Cells were then counted and 1x105 cells per knockout were taken and pelleted once more. To make Cas9-gRNA RNA complex, 2 μl of 20 μM of Cas9 (Synthego) was mixed with 3 μl of 100 μM modified gRNA (Synthego) and allowed to incubate at room temperature for 10 minutes. One of two gRNAs was delivered per transfection, and sequences were designed to target exon 5 in the ATXN2 gene (g1: AAAGUACAGAAUCCAGUUCG; g2: GAAAAGUACAGAAUCCAGUU). Cells were resuspended in supplemented P3 Primary Cell 4D-Nucleofector (Lonza) buffer and mixed with the RNP complex. Cells were nucleofected on the X-Unit of the 4D-Nucleofector system (Lonza) using pulse code CA-137. Cells were resuspended and in mTeSR plus supplemented with RI, and seeded onto one well of a Matrigel coated 6-well plate. Cells were allowed to recover and expand before being dissociated once more with Accutase and seeded on a 10cm2 coated dish in single cell colonies. Individual colonies were picked, expanded, and assessed for ATXN2 knockout using Western Blot.

4.4.5 Generation of polyQ donor vectors

ATXN2 arms of homology were obtained by PCR amplification of a 4.2 kb locus flanking ATXN2 exon 1 (F Primer: 5'-AAGTTCGCACATTGTTTTGAGGT-3', R Primer: 5'- GCCAGGTGTGGTAGCACGAACC-3'), which was then cloned into a DNA vector using Gibson assembly. Repeat sequences were introduced through inverse PCR F Primer: 5'- (CAG/CAA X N)-CCGCCGCCCGCGGCTGCCAATG-3', R Primer: 5'-CCTCACCATGTCGCTGAAGCCCC-3', where the 5' tail contained the specific length of CAG expansions desired. Blunt ends of the PCR product were ligated together using T4 Ligase and transformed into competent E. coli for selection and sequence validation. To prevent re-sectioning of DNA with CRISPR/Cas9, a PAM mutation proximal to the gRNA recognition site was introduced using inverse PCR without changing the amino acid sequence.

4.4.6 Generation of ATXN2 knock-in iPSC

Genetic polyQ knock-ins were performed in CV-B line. Cells were dissociated with Accutase and passed through a 40 μm strainer. Cells were spun down at 200xg for 5 minutes and resuspended in mTeSR plus supplemented with 10μM ROCK-inhibitor (RI). Following counting, 5x105 cells per transfection were taken and pelleted once more. Cas9-gRNA RNA complex was made by mixing 2 μl of 20 μM of Cas9 (Synthego) with 3 μl of 100 μM modified gRNA (Synthego) and allowed to incubate at room temperature for 10 minutes. The knock-in gRNA targets a sequence in exon 1 of the ATXN2 gene proximal to the polyQ tract (Knock-in gRNA: CGGCGUGCGAGCCGGUGUAU). Cells were resuspended in supplemented P3 Primary Cell 4D-Nucleofector (Lonza) buffer and mixed with the RNP complex, as well as with 3μg polyQ donor plasmid and 1μg pCE-mP53DD plasmid (Addgene: 41856) in inhibit DNA breakage-associated apoptosis. Cells were nucleofected on the X-Unit of the 4D-Nucleofector system (Lonza) using pulse code CA-137. Cells were then collected in mTeSR plus supplemented with RI and seeded onto a Matrigel coated 10cm^2 dish. Single colonies were allowed to expand, were

picked, and assessed for ATXN2 knock-in using DNA genotyping assessment on an agarose gel (F Primer: 5'-CGTGCGAGCCGGTGTATG-3', R Primer: 5'- CGACGCTAGAAGGCCGCTG-3'). Clones that passed this initial screening were expanded and subjected to secondary genotyping by isolating RNA and performing RT-PCR using primers that anneal to exon 1 and exon 2, respectively (F Primer: 5'-CCGCCCGGCGTGCGAGCCGGTGTATGG-3', R Primer: 5'-GCAGTCCTTTGTTACTGTTTCGACCT-3'). PCR products were then isolated, subjected to TOPO cloning and confirmed via Sanger sequencing.

4.4.7 RNA Extraction and library preparation

RNA was extracted using Trizol and Direct-zol RNA miniprep columns (Zymo) according to manufacturer's instructions, complete with on-column DNase treatment. With 1μg of total RNA as input material, RNA-seq libraries were prepared using the TruSeq-based Stranded mRNA Prep kit from Illumina according to manufacturers' instructions. Libraries were pooled and sequenced on a 100bp paired-end run on a NovaSeq6000 (Illumina).

4.4.8 RNA-seq processing

Raw RNA-seq reads were adapter-trimmed using cutadapt (v 1.14.0), mapped to repetitive elements (version 18.05) prior to alignment to mm10 using STAR (v 2.5.2b). Aligned reads were sorted and indexed using samtools (v 1.9), and gene features assigned to gencode annotation (mm10 v15) were quantified using Subread featureCounts package (v 1.5.3). Reads normalized and converted to transcripts to millions (TPM) and differential expression performed using DESeq2 (v 1.30.1). Only genes with TPM ¿ 0 were considered for pairwise differential expression analysis. Genes were considered differentially expressed with an FDR < 0.05 and —log2(FC)— > 1.

4.4.9 eCLIP-seq library preparation

The standard, paired-end version of eCLIP-procedure was performed for ataxin-2 (Van Nostrand et al., 2016, Van Nostrand et al., 2020). On day 30 of differentiation, cell dishes were placed on a cold metallic block and UV-cross-linked (400 mJ cm^{-2}, 254 nm) prior rinshing with PBS and snap frozen. Still-frozen crosslinked cell pellets were lysed supplemented with 11 μl Murine RNase Inhibitor per ml of lysate. Samples were sonicated for 5 cycles (30s on, 30s off) and treated with RNase I to digest RNA. For each sample, 10 μg Ataxin-2 polyclonal antibody (Proteintech, 21776-1-AP) was pre-coupled to 125 μl M-280 sheep anti-rabbit Dynabeads then added to lysate and incubated for 2 hours at 4^{o}C with rotation. In parallel, 10% input lysate was removed and used for Rb IgG IP control with 1μg antibody for western blot. Following incubation, 2% input was removed and kept for size-matched input sample, while the remainder of the RNP bound material was washed, subjected to dephosphorylation and 3' end ligated with an RNA adapter on bead. Samples (IP, Input, IgG) were run on an SDS-polyacrylamide gel and transferred onto a nitrocellulose membrane. Membrane pieces for IP and sized-matched input (SMInput) were cut from the size of the protein (140 kDa) to 75 kDa above and RNA was extracted using proteinase K digestion and subsequently purified on column. Input samples were 3'-end ligated separately with its own RNA adapter and libraries were subjected to reverse transcription using AffinityScript RT enzyme, and 5'-linker ligated with a DNA adapter. After quantifying with qPCR, libraries were amplified with Q5 PCR master mix using primers containing Illumina indexes and adapters. Libraries were then sequenced on an Illumina Hiseq 4000 on a 55PE run.

4.4.10 eCLIP sequence processing

Raw reads were processed essentially as described in Van Nostrand et al. 2016. Reads were adapter-trimmed and mapped to mouse-specific elements from Repbase (version 18.05) using STAR to remove repeat-mapping reads. Remaining reads were mapped to the human

genome hg19 once again using STAR. PCR-duplicates were removed using the unique molecular identifier sequences in the 5' adaptor, and remaining reads were retained as "usable reads." Peaks on IP samples were called using CLIPper, and each peak was normalized to its corresponding input sample calculating the fraction of the number of usable reads from the IP sample relative to the usable reads from the SMInput. Each peak was assigned a fold enrichment (relative to SMInput) and p-value (χ^2 test, or Fisher's exact test if the observed or expected read number in eCLIP or SMInput was below 5). All eCLIP processing code is available on GitHub (https://github.com/YeoLab/eclip). Reproducible peak regions were determined across replicates using Irreproducible Discovery Ratio, or IDR (https://github.com/YeoLab/merge_peaks). Consensus IDR peaks were deemed significant at fold change > 8 and a $P < 0.001$.

4.4.11 Region analysis and enrichment of k-mers

Peaks were assigned genomic regions using hg19 Gencode annotations (https://github.com/byee4/annotator). To analyze sequence preferences in the significant peaks, we performed 6-mer analysis (https://github.com/byee4/clip_analysis), which performs k-mer enrichment analysis and HOMER findMotifs.pl.

4.5 Figures

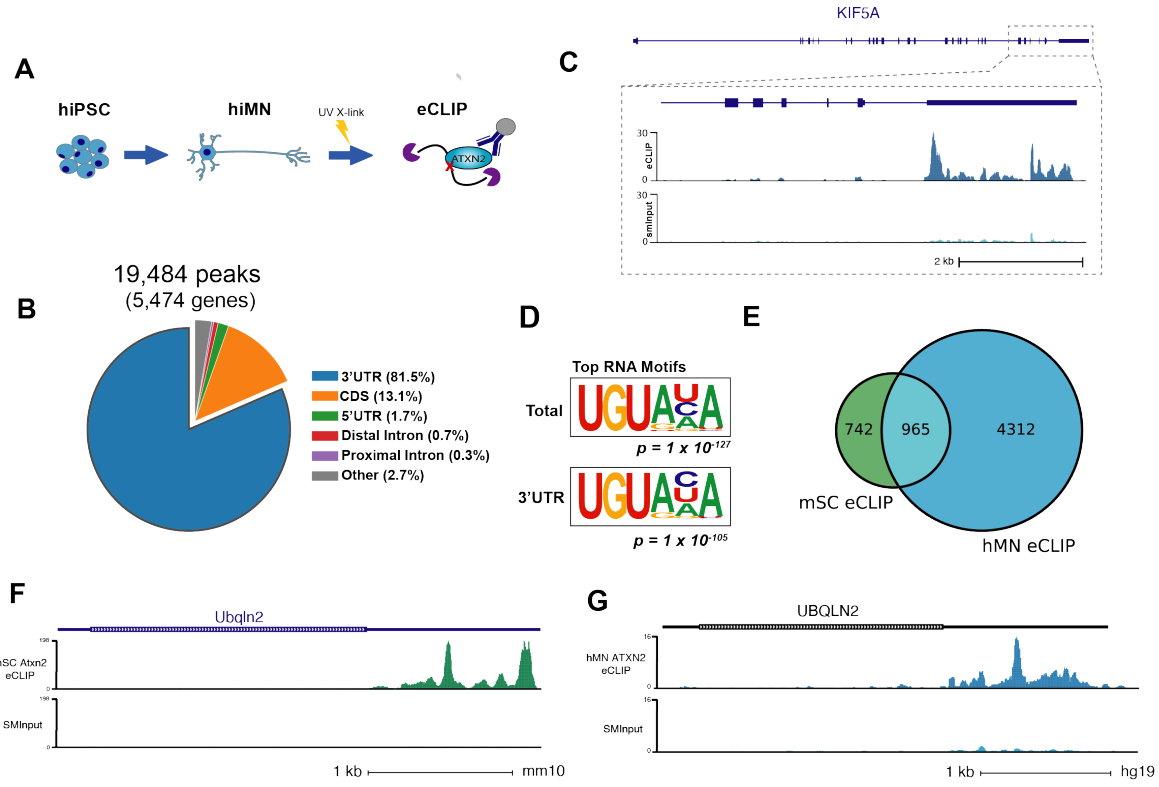


Figure 4.1. Mapping RNA-binding landscape of ATXN2 in iPSC-MNs. A. Schematic representation of differentiation of iPSCs into motor neurons followed by eCLIP. B. Regional distribution of reproducible (by IDR) and significant (FC $>$ 8, pvalue $<$ 0.001) peaks. This revealed 19,484 peaks across 5,474 RNAs with a heavy preference for 3'UTR sequences. C. Representative browser tracks of ATXN2 eCLIP (top) and size-matched input (bottom) to the 3'UTR of *KIF5A*. D. Motif analysis of all (top) or 3'UTR-enriched (bottom) ATXN2 peaks via HOMER. E. Overlap of target genes found in mouse spinal cord Atxn2 eCLIP and human iPSC-MN ATXN2 eCLIP. F, G. Comparative binding of ataxin-2 for a shared gene ubiquitin-2 in mouse spinal cord (F) and iPSC-MN (G).

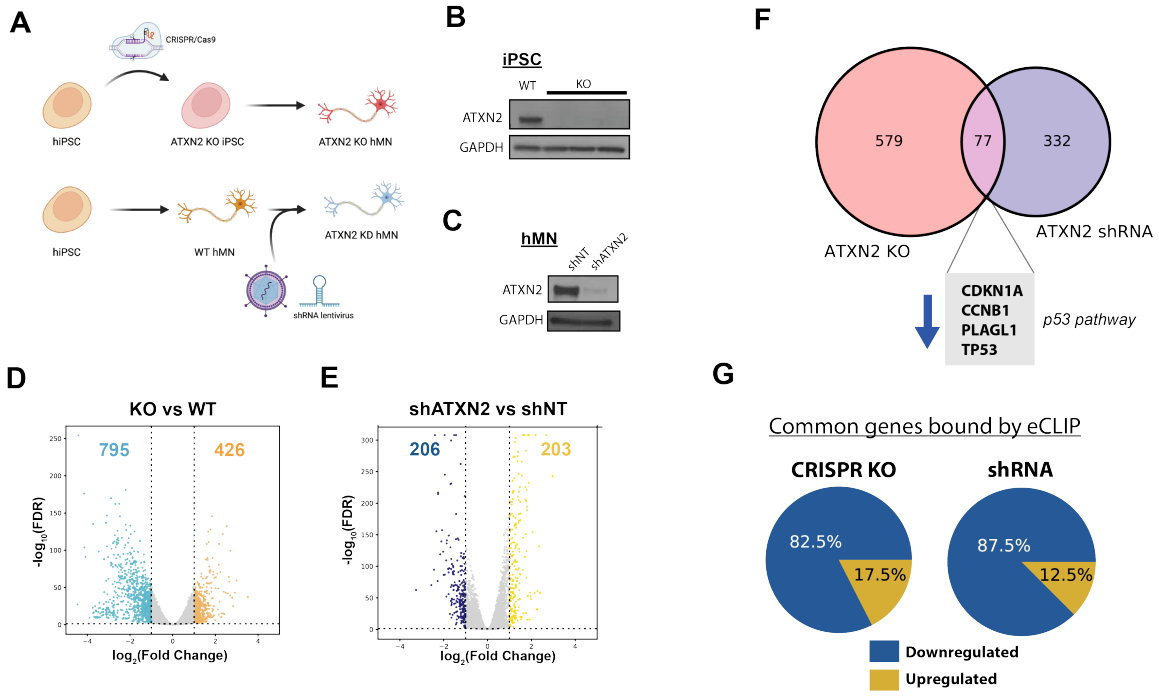


Figure 4.2. Orthogonal methods of ATXN2 depletion reveals suppression of pro-apoptotic genes. A. Schematic detailing methods of ATXN2 ablation in iPSC-MNs. ATXN2 can either be knocked out using CRISPR/Cas9 in the iPSC stage prior to differentiation into motor neurons or knocked down using shRNAs in terminal stages of MN differentiation. B, C. Validation of parallel depletion strategies for knockout (B) or shRNA knockdown (C) using Western Blot. GAPDH is presented as a loading control. D, E. Volcano plots representing differentially expressed genes (—log2(FC)— $>$ 1 and FDR $<$ 0.05) in motor neurons for knockout (D) or knockdown (E) lines in comparison to their respective controls. F. Venn diagram depicting shared genes detected as being differentially expressed in both depletion strategies. Notable genes included those of p53 signaling pathway, responsible for regulating and initiating cellular apoptosis. G) Directional change of common target genes from (F) supported by eCLIP binding data.

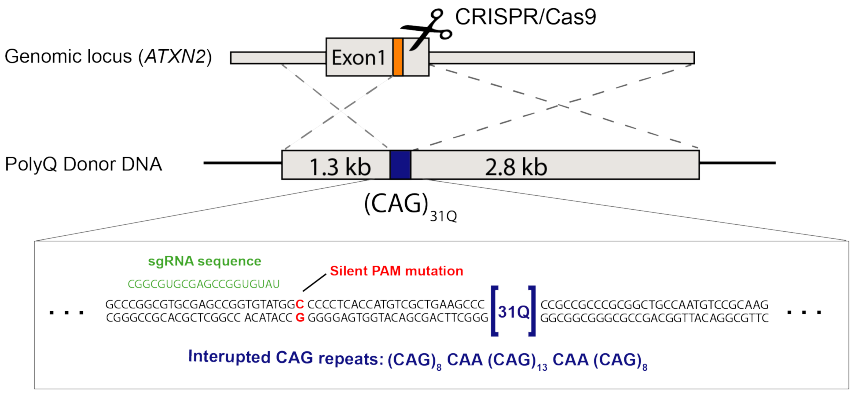


Figure 4.3. Generation of isogenic polyQ knock-in cell lines using CRISPR/Cas9. Schematic representing a CRISPR-based knock-in strategy to introduce polyQ repeats into the endogenous locus of ATXN2. A guide RNA will cut proximal to the polyQ tract in exon 1 of ATXN2, allowing for homology-directed repair (HDR) fix the lesion of the cut allele using a donor construct bearing the desired CAG sequence. A CAA-interrupted, intermediate-length (31Q) sequence that is associated with ALS vulnerability is portrayed. A synonymous PAM mutation is also included to prevent donor cutting or re-cutting by Cas9.

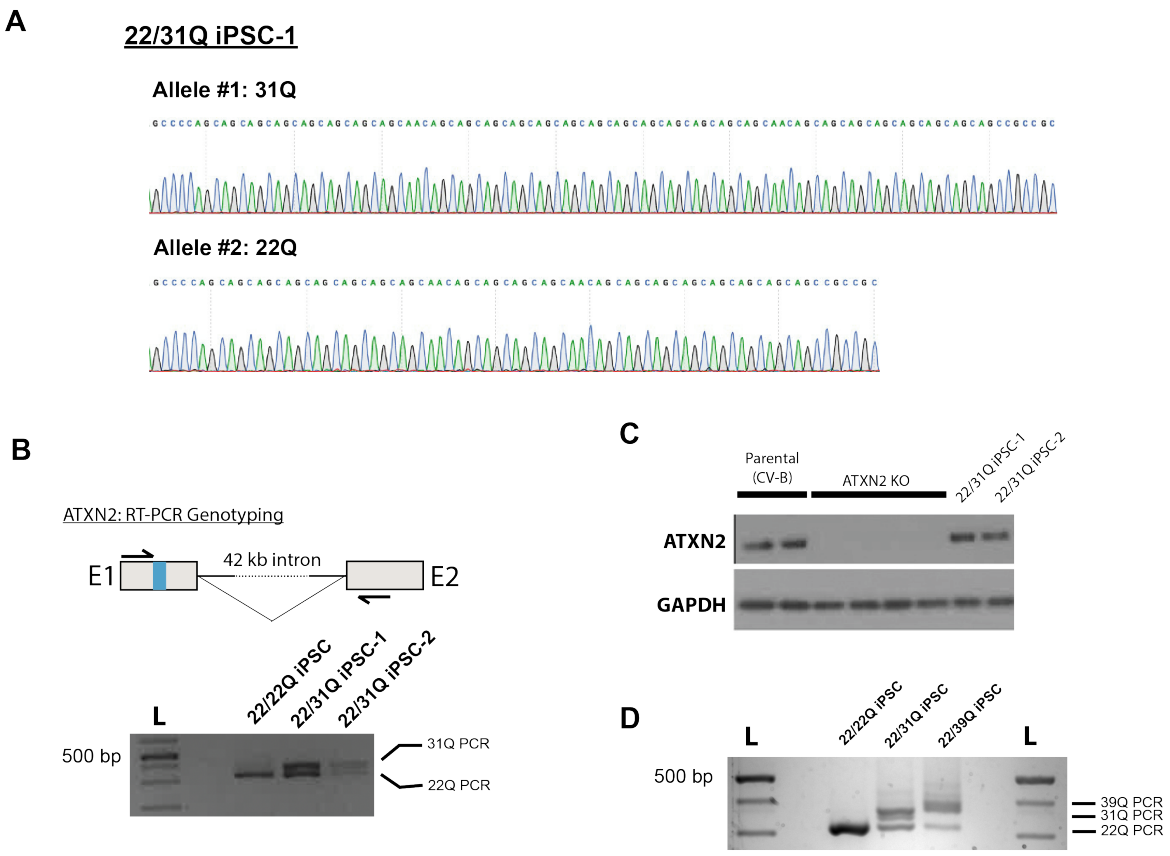


Figure 4.4. Successful establishment of isogenic ATXN2 polyQ iPSC lines A. Sanger sequencing traces for expanded (31Q) and normal (22Q) respective alleles found in a characterized heterozygous knock-in iPSC clone. Sequences show expected length, preserved intervals of CAG/CAA, and no local lesions. B. RT-PCR genotyping of 22/31Q clones demonstrating expression and detectable size differences between 22Q and 31Q alleles. Primers annealed to exon 1 and exon 2 sequences respectively, ensuring that the amplicon was from RNA/cDNA. C. Western Blot in motor neurons confirming ATXN2 expression preserved for knock-in lines. Parental CV-B lines and ATXN2 knockout lines were used as expression controls. GAPDH is presented as a loading control. D. RT-PCR genotyping of 22/31Q and 22/39Q iPSC lines, demonstrating successful editing of longer polyQ expansions into the *ATXN2* locus.

References

Becker, L. A., Huang, B., Bieri, G., Ma, R., Knowles, D. A., Jafar-Nejad, P., Messing, J., Kim, H. J., Soriano, A., Auburger, G., Pulst, S. M., Taylor, J. P., Rigo, F. & Gitler, A. D. 2017. Therapeutic reduction of ataxin-2 extends lifespan and reduces pathology in TDP-43 mice. Nature, 544, 367-371.

Casarotto, E., Sproviero, D., Corridori, E., Gagliani, M. C., Cozzi, M., Chierichetti, M., Cristofani, R., Ferrari, V., Galbiati, M., Mina, F., Piccolella, M., Rusmini, P., Tedesco, B., Gagliardi, S., Cortese, K., Cereda, C., Poletti, A. & Crippa, V. 2022. Neurodegenerative Disease-Associated TDP-43 Fragments Are Extracellularly Secreted with CASA Complex Proteins. Cells, 11.

Chambers, S. M., Fasano, C. A., Papapetrou, E. P., Tomishima, M., Sadelain, M. & Studer, L. 2009. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. Nat Biotechnol, 27, 275-80.

Chen, Y., Huang, R., Yang, Y., Chen, K., Song, W., Pan, P., Li, J. & Shang, H. F. 2011. Ataxin-2 intermediate-length polyglutamine: a possible risk factor for Chinese patients with amyotrophic lateral sclerosis. Neurobiol Aging, 32, 1925 e1-5.

Du, Z. W., Chen, H., Liu, H., Lu, J., Qian, K., Huang, C. L., Zhong, X., Fan, F. & Zhang, S. C. 2015. Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells. Nat Commun, 6, 6626.

Elden, A. C., Kim, H. J., Hart, M. P., Chen-Plotkin, A. S., Johnson, B. S., Fang, X., Armakola, M., Geser, F., Greene, R., Lu, M. M., Padmanabhan, A., Clay-Falcone, D., Mccluskey, L., Elman, L., Juhr, D., Gruber, P. J., Rub, U., Auburger, G., Trojanowski, J. Q., Lee, V. M., Van Deerlin, V. M., Bonini, N. M. & Gitler, A. D. 2010. Ataxin-2 intermediate-length polyglutamine expansions are associated with increased risk for ALS. Nature, 466, 1069-75.

Farg, M. A., Soo, K. Y., Warraich, S. T., Sundaramoorthy, V., Blair, I. P. & Atkin, J. D. 2013. Ataxin-2 interacts with FUS and intermediate-length polyglutamine expansions enhance FUS-related pathology in amyotrophic lateral sclerosis. Hum Mol Genet, 22, 717-28.

Gore, A., Li, Z., Fung, H. L., Young, J. E., Agarwal, S., Antosiewicz-Bourget, J., Canto, I., Giorgetti, A., Israel, M. A., Kiskinis, E., Lee, J. H., Loh, Y. H., Manos, P. D., Montserrat, N., Panopoulos, A. D., Ruiz, S., Wilbert, M. L., Yu, J., Kirkness, E. F., Izpisua Belmonte, J. C., Rossi, D. J., Thomson, J. A., Eggan, K., Daley, G. Q., Goldstein, L. S. & Zhang, K. 2011. Somatic coding mutations in human induced pluripotent stem cells. Nature, 471, 63-7.

Hart, M. P., Brettschneider, J., Lee, V. M., Trojanowski, J. Q. & Gitler, A. D. 2012. Distinct TDP-43 pathology in ALS patients with ataxin 2 intermediate-length polyQ expansions. Acta

Neuropathol, 124, 221-30.

Hart, M. P. & Gitler, A. D. 2012. ALS-associated ataxin 2 polyQ expansions enhance stress-induced caspase 3 activation and increase TDP-43 pathological modifications. J Neurosci, 32, 9133-42.

Kiehl, T. R., Nechiporuk, A., Figueroa, K. P., Keating, M. T., Huynh, D. P. & Pulst, S. M. 2006. Generation and characterization of Sca2 (ataxin-2) knockout mice. Biochem Biophys Res Commun, 339, 17-24.

Kim, H. J., Raphael, A. R., Ladow, E. S., Mcgurk, L., Weber, R. A., Trojanowski, J. Q., Lee, V. M., Finkbeiner, S., Gitler, A. D. & Bonini, N. M. 2014. Therapeutic modulation of eIF2alpha phosphorylation rescues TDP-43 toxicity in amyotrophic lateral sclerosis disease models. Nat Genet, 46, 152-60.

Lastres-Becker, I., Rub, U. & Auburger, G. 2008. Spinocerebellar ataxia 2 (SCA2). Cerebellum, 7, 115-24.

Lee, T., Li, Y. R., Ingre, C., Weber, M., Grehl, T., Gredal, O., De Carvalho, M., Meyer, T., Tysnes, O. B., Auburger, G., Gispert, S., Bonini, N. M., Andersen, P. M. & Gitler, A. D. 2011. Ataxin-2 intermediate-length polyglutamine expansions in European ALS patients. Hum Mol Genet, 20, 1697-700.

Lu, S., Hu, J., Aladesuyi, B., Goginashvili, A., Vazquez-Sanchez, S., Diedrich, J., Gu, J., Blum, J., Oung, S., Yu, H., Ravits, J., Liu, C., Yates, J. & Cleveland, D. W. 2021. Heat shock chaperone HSPB1 regulates cytoplasmic TDP-43 phase separation and liquid-to-gel transition. bioRxiv.

Maor-Nof, M., Shipony, Z., Lopez-Gonzalez, R., Nakayama, L., Zhang, Y. J., Couthouis, J., Blum, J. A., Castruita, P. A., Linares, G. R., Ruan, K., Ramaswami, G., Simon, D. J., Nof, A., Santana, M., Han, K., Sinnott-Armstrong, N., Bassik, M. C., Geschwind, D. H., Tessier-Lavigne, M., Attardi, L. D., Lloyd, T. E., Ichida, J. K., Gao, F. B., Greenleaf, W. J., Yokoyama, J. S., Petrucelli, L. & Gitler, A. D. 2021. p53 is a central regulator driving neurodegeneration caused by C9orf72 poly(PR). Cell, 184, 689-708 e20.

Markmiller, S., Soltanieh, S., Server, K. L., Mak, R., Jin, W., Fang, M. Y., Luo, E. C., Krach, F., Yang, D., Sen, A., Fulzele, A., Wozniak, J. M., Gonzalez, D. J., Kankel, M. W., Gao, F. B., Bennett, E. J., Lecuyer, E. & Yeo, G. W. 2018. Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. Cell, 172, 590-604 e13.

Mitra, J., Guerrero, E. N., Hegde, P. M., Liachko, N. F., Wang, H., Vasquez, V., Gao, J., Pandey, A., Taylor, J. P., Kraemer, B. C., Wu, P., Boldogh, I., Garruto, R. M., Mitra, S., Rao, K. S. & Hegde, M. L. 2019. 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.

Sellier, C., Campanari, M. L., Julie Corbier, C., Gaucherot, A., Kolb-Cheynel, I., Oulad-Abdelghani, M., Ruffenach, F., Page, A., Ciura, S., Kabashi, E. & Charlet-Berguerand, N. 2016. Loss of C9ORF72 impairs autophagy and synergizes with polyQ Ataxin-2 to induce motor neuron dysfunction and cell death. EMBO J, 35, 1276-97.

Tavares De Andrade, H. M., Cintra, V. P., De Albuquerque, M., Piccinin, C. C., Bonadia, L. C., Duarte Couteiro, R. E., Sabino De Oliveira, D., Claudino, R., Magno Goncalves, M. V., Dourado, M. E. T., Jr., Cruz De Souza, L., Teixeira, A. L., De Godoy Rousseff Prado, L., Tumas, V., Bulle Oliveira, A. S., Nucci, A., Lopes-Cendes, I., Marques, W., Jr. & Franca, M. C., Jr. 2018. Intermediate-length CAG repeat in ATXN2 is associated with increased risk for amyotrophic lateral sclerosis in Brazilian patients. Neurobiol Aging.

Van Damme, P., Veldink, J. H., Van Blitterswijk, M., Corveleyn, A., Van Vught, P. W., Thijs, V., Dubois, B., Matthijs, G., Van Den Berg, L. H. & Robberecht, W. 2011. Expanded ATXN2 CAG repeat size in ALS identifies genetic overlap between ALS and SCA2. Neurology, 76, 2066-72.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods, 13, 508-14.

Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X. D., Graveley, B. R. & Yeo, G. W. 2020. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biol, 21, 90.

Vogt, M. A., Ehsaei, Z., Knuckles, P., Higginbottom, A., Helmbrecht, M. S., Kunath, T., Eggan, K., Williams, L. A., Shaw, P. J., Wurst, W., Floss, T., Huber, A. B. & Taylor, V. 2018. TDP-43 induces p53-mediated cell death of cortical progenitors and immature neurons. Sci Rep, 8, 8097.

Yu, Z., Zhu, Y., Chen-Plotkin, A. S., Clay-Falcone, D., Mccluskey, L., Elman, L., Kalb, R. G., Trojanowski, J. Q., Lee, V. M., Van Deerlin, V. M., Gitler, A. D. & Bonini, N. M. 2011. PolyQ repeat exp

Chapter 5

Conclusions

5.1 Abstract

The human transcriptome remains largely uncharacterized and thoroughly captivating to biologists at large. The coordinated effort through which diverse classes of RBPs regulate every stage of post-transcriptional gene expression is both incredibly exciting and daunting at the same time. However, in the era of functional genomics, these challenges are now more approachable than ever, allowing basic researchers and clinicians alike to understand how RNA processing affects human health and disease at both protein-specific and global levels. This ability to fundamentally understand the transcriptome also paves the way the rapid advancement of molecular tools designed to target and manipulate RNA, which collectively offer tremendous potential as effective and non-permanent ways to perform genomic engineering. These two principles of discovery and application have largely guided my thesis work and believe have given me a greater appreciation and understanding for the wide world of RNA.

5.2 Perspectives on RNA-targeting Cas-ADAR project

RNA base editors are naturally occurring RBPs that tightly regulate post-transcriptional gene expression by changing the sequence of RNA molecules to differ from their DNA template. As such, have long served as promising means to achieve non-permanent and reversible means to achieve successful gene editing without having to modify DNA. This is a pressing issue, as there

are thousands of causative pathogenic variants that can be corrected with single base alterations. However, due to the complexity and structural diversity of the transcriptome when compared to chromatin, base editing strategies have a much higher potential for introduction of undesired off-target edits. Therefore, and a fundamental assessment of available RNA editing technologies is highly desirable. In this project, we generated a new, catalytically-dead Cas9-ADAR2 based A-to-I editing platform to modify coding and non-coding sequences of mRNA in mammalian cells. We also, through modification and extension of the accompanying gRNA molecule, described a novel targeting strategy no longer dependent upon the spacer RNA sequence previously fount to be necessary for targeting to both DNA and RNA. When this project first began, we had hoped to use it as a means to usher in a new age of Cas-based RNA editing strategies, as we were on the cutting edge of brand-new technologies. However, as the field rapidly expanded and newly characterized Cas-based RNA-targeting systems became more common, it became apparent to us that no one had yet performed parallel on-target and off-target analyses of these available systems. By uncovering similar on-target efficiency, as well as comparable abundant off-target capacity, across three of the major Cas-based platforms (Cas19, Cas13b, Cas13d), my work on this project allowed us to evaluate the ultimate benefits and limitations RNA-targeting editing systems.

5.3 Future directions of RNA-targeting Cas-ADAR project

One of our main takeaways was that while somewhat effective, exogenous expression of ADAR2 protein domains introduces many off-target edits across the transcriptome regardless of Cas effector, often exceeding the degree of our on-target site. Therefore, strategies that rely on recruitment of endogenously expressed ADAR proteins within cells to achieve on-target editing seem much more feasible for therapeutic intervention post-transcriptional genomic editing. RNA-targeting tools developed in this project and by others in the lab can perhaps be adapted for this strategy, which would prove to be a much simpler and safer method going forward.

5.4 Perspectives on STAMP project

While we found that the robustness of exogenously expressed RNA-base editors might prove to be excessive for precise genetic engineering, remain promising as tools for discovery of RNA-RBP interactions. In the past few years, the sequencing-based method TRIBE has taken advantage of this inherent 'noisiness' of ADAR proteins by fusing them to RBPs of interest and using them to mark sites of target gene interactions using A-to-G mismatches found through RNA-seq. This method was genetically encodable and thus had the capacity to performed *in vitro* and in *Drosophila* models using small amounts of input material. However, authors note that resolution of their method remains quite low, with edits often being located several kilobases away due to double-stranded RNA preferences of the attached ADAR moieties. While this might not an issue for quantifying bound versus unbound genes for an RBP of interest, this lacks the ability to map precise sites of RBP-binding in the same way that CLIP-seq might. We thus took advantage of a broader, single-stranded RNA-recognizing base editing protein APOBEC1, which mediates C-to-U deamination of proximal RNA molecules. Our work show that this alteration captured more precise and robust RBP-RNA interaction information, which we validated with eCLIP data, even successfully recapitulating known RNA motifs of specific RBPs using edit signatures alone. Additionally, the innovation of using APOBEC1-fused ribosomal proteins made it possible to analyze transcript abundance and ribosomal association of genes from a single RNA-seq experiment. Lastly, the ability to couple RBP- and Ribo-STAMP with cutting-edge, transcriptomic-based techniques opens new avenues in functional genomics with respect to RBP and translational biology.

5.5 Future directions of STAMP

As a novel genomic technique, I think STAMP has great potential for growth and innovation. Aside from the obvious use of STAMP to investigate RBP- or ribosomal interactions in cell type-specific contexts, there are many back-end improvements or points of innovation

that are occurring in the meantime. In addition to improved analytical methods to differentiate signal more clearly from noise, technical modifications can easily be made to improve quality of information. Others in the lab are currently in the process of iteratively testing other base editors, natural and synthetically evolved, to assess efficiency and specific in sites across the transcriptome. By expanding our search beyond C-to-U editors, we have the potential to not only identify more faithful base editors less prone to off-target editing, but also leave open the possibility of leveraging different editing signatures for the multiplexing purposes. Such strategies would allow us to assess the binding landscape of multiple RBPs in one experiment through use of different enzymes specific to each RBP-STAMP fusion.

As I alluded to in my discussion, more tunable means of STAMP expression will also increase the sensitivity through which we can study short timescale phenomenon such as changes in translational regulation. For instance, it takes several hours to express our fusion protein under our current dox-inducible platform, but translational bursting may happen on the scale of mere minutes. Having the capacity to more tightly control STAMP expression post-translationally, either through means of a small-molecule stabilized destabilization domain or a light-sensitive photocage module, would be tremendously powerful in dissecting expression changes over small windows of time.

5.6 Perspectives on ATXN2 projects

As I began in graduate school, I was extremely keen on investigating RBPs in the context of neurological disease, encouraged by the many exciting publications coming from our lab at the time (see Kapeli et al. 2016, Martinez et al. 2016). I gravitated towards studying ataxin-2 in the central nervous system not only because of its seemingly significant role in disease pathogenesis and vulnerability, but because there was still very little that was known about it from a regulatory or RNA-processing perspective. My goal was to characterize the functional role of ataxin-2 in neurons similar to what had been one with TDP-43, TAF15, and HNRNPA2B1 previously.

I have found this project, or what has been become projects, to have been an extremely informative, though extremely challenging undertaking that I think has provided some good learning opportunities. From an analytical and bioinformatics perspective, it has opened the door to analyzing rather large and complicated datasets with much less hesitation than I had previously. Entering graduate school with very little coding experience, this project exposed me to several sequencing techniques and analytical methods practical to most areas of genetics and genomics research. Having the responsibility of generating, processing, and analyzing my own high throughput data helped me build a stronger foundation in the areas of bioinformatics and computational biology, and much more familiar with data quality and experimental considerations necessary for publication of successful scientific projects. I will continue to hone and expand these areas of expertise in my postdoc and beyond in order to become confident and proficient bioinformatician capable of training others in a similar manner.

I have also had the opportunity to work with new model systems including human stem cell models, which I was previously unfamiliar with. Doing so allowed me to become more practiced and proficient in stem cell biology, from maintenance to differentiation, that will undoubtedly be useful to me should I continue in disease modeling. Importantly, it has also allowed me to gain expertise in genomic engineering techniques using CRISPR/Cas9, which is a valuable skill in the fields of molecular biology and disease modeling. It has also given me practical experience as to the multiple considerations that need to be met before engaging in a laborious gene editing strategy, such as the inclusion of necessary controls (e.g. "mock edited" controls to process in parallel) consideration of genetic background variation (e.g. multiple lines from different donors), the ability to detect specific cellular phenotypes (cell death, protein aggregation), and necessary scale (number of lines, clones per line, etc.). Additionally, having the ability to compare your findings to disease diagnosis-verified patient iPSC lines is something that should not be easily overlooked.

5.7 Future directions for ATXN2 projects

In the near future, I hope to be able to submit the findings that I have for review, as well as work with others in lab to follow-up on some of the more exciting observations that I have not been able to explore further.

For the mouse project, I would like to see my ataxin-2 binding/expression data by followed-up by a comparative analysis mapping Tdp-43 in the mouse spinal cord using eCLIP. Due to the small changes in gene expression seen in response to ataxin-2 deletion, it stands to reason that there are many changes occurring greater than those that can be explained by Atxn2-RNA interactions alone. Seeing the UG-rich sequence preference of Atxn2 and knowing that Tdp-43 also abides by similar targeting requirements, it would be worthwhile mapping the Tdp-43 landscape in adult mice, which has not been done to this date. Doing so will provide us with an additional later of functional binding data and will hopefully color our perspective regarding the interaction between Atxn2 and Tdp-43 in ALS pathogenesis. For the human models, I hope to work with others in the lab to further investigate from of the targets identified with ataxin-2 depletion in motor neurons. The basal downregulation of p53 and its downstream transcriptional targets is very enticing, as it may help to bridge the gap between proteotoxic or genotoxic stress conferred by aberrant TDP-43 and neuronal apoptosis in late stages of disease. There is already a link between ALS vulnerability and p53-response pathway, however that area remains underexplored. It will therefore be important to see in motor neurons if these molecular pathways are differentially regulated in ALS-based contexts, such as mis-localized TDP-43 or cellular stress, and whether they can be reversed with ataxin-2 depletion. This will not only allow us to compare shared expression pathways observed between our mouse and human models but will also provide us with a means to see if these same genes are important for maintaining neuronal integrity in the presence of toxic TDP-43.

Lastly, it will be interesting to see if expanded polyQ lines can confer some degree of disease vulnerability with respect to CAG length. Although intermediate length polyQ mutations

may not be sufficient to cause disease, long polyQ repeats, or cellular stress such as puromycin or TDP-43 forced mislocalizaation, might lead to added toxicity in motor neurons beyond that seen in healthy genetic backgrounds. As I transition to my next steps as a postdoc, I will be working with others to see these molecular and cellular characterizations through, as well as provide fellow researchers with access to patient-derived, intermediate-expansion length iPSC lines that I have received from Cedars-Sinai Stem Cell Core. It is my hope that these models, in conjunction with the isogenic lines generating during my graduate work, will help to shed light on the underlying cause of ATXN2-mediated ALS vulnerability.

5.8 Potential impact of my thesis work

When I joined Dr. Gene Yeo's lab to begin my doctoral research in 2016, it was my goal to investigate interesting RNA-processing paradigms contributing to human disease. At the same time, I was very interested in the lab's capacity to develop RNA-targeting technologies that might alleviate disease burden by modulating genetic information at the post-transcriptional level. In that respect, I believe that my time has been productive and perhaps provide some benefits to the RNA community as a whole.

My contributions to RNA-targeting technologies have helped to yield some very impactful breakthroughs with respect to therapeutic treatments of repetitive expansion diseases. Early efforts to harness these technologies resulted in several successful preclinical results whereby RNA-targeting Cas9 and Cas13 could be used to target and destroy trinucleotide RNA aggregates in *vitro* and *in vivo*, work that continues a to produce promising therapeutic avenues at LocanaBio here in San Diego. Although not the major focus of my thesis work, this did serve as an entry point into the field of RNA-targeting research, where I quickly helped develop an RNA-editing platform capable of editing transcripts in cells. Although this work ultimately yielded Cas-driven RNA editor that might be limited in its capacity to reverse pathogenic mutations due to the inherently high off-target effects caused by exogenous expression of ADAR proteins, this effort

helped to inform the field about the 'boons and banes' of current RNA-editing strategies. In a field that is extremely promising and fast-moving, we are seeing a gradual shift towards the recruitment of endogenously expressed RNA editors to in order to mediate site-directed sequence changes, as is the current strategy for Shape Therapeutics. Thus, benchmarking papers such as this can be important in informing the general scientific community about the benefits and potential limitations of available technologies.

My RNA editing work with Dr. Kristopher Brannan helped to inform the next phase of RNA editors in the lab–that of STAMP. STAMP is a new sequencing-based method where we actually rely on the robust editing of a particular RNA editor, in this case APOBEC1, to deposit large amounts of detectable edits to track localization of specific RNA-binding proteins fused to APOBEC1 *in situ*. Using these editing 'footprints' that these protein fusions leave behind, we now have a quick and simplified means of being able to track RNA-RBP interactions just from sequencing. Although others have proposed similar methods using RNA editors before us, our collective team effort has pushed the boundary of RNA functional genomics; by pairing STAMP with commonly available sequencing techniques, such as single-cell profiling, we have a straightforward means to probe complicated regulatory processes such as translation more deeply and at increasingly granular levels. I am excited to see what comes next on this front as we begin to share these resources with community at large to uncover new truths with respect to the underpinnings of development and disease.

My work on ataxin-2 will hopefully be informative to a broader community of RNA researchers and neurobiologists who seek to further understand the linkage between aberrant RNA processing and disease vulnerability. Ataxin-2 has long been a mysterious RBP with multiple implications in neurodegeneration. Recent efforts have focused on means to deplete ataxin-2, either directly or indirectly, to potentially provide therapeutic benefit in the face of pathogenic protein aggregation. At the same time, secondary consequences of this depletion strategy are not currently known, nor are mechanistic basis of this protective effect. Although a complex question, I believe that my graduate work will be informative in filling in some of these gaps of

knowledge. By profiling the functional, transcriptome-side landscape of ataxin-2 in mouse and human neurons, we can provide a valuable resource to investigators who seek to understand the relationship between ataxin-2 and its RNA targets on a site-specific, as well as a global, basis. Additionally, the isogenic polyQ modeling strategies that I have generated are widely applicable to many other types of repeat expansion diseases beyond that of ATXN2. These methods have already proven useful to others in the Chan Zuckerberg Initiative Neurodegeneration Challenge Network, who have consulted with our group to generate isogenic iPSC lines for Huntington's Disease, Myotonic Dystrophy type 1, and various forms of Spinocerebellar Ataxia.

Overall, I am grateful for the diverse number of projects that I had the privilege of contributing to, and I hope the results of these efforts help to inform the public on how best to investigate RNA metabolism in health and human disease.