UNIVERSITY OF CALIFORNIA, SAN DIEGO

Methods for Integrative Analysis of RNA Binding Proteins

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

in

Bioinformatics and Systems Biology

by

Gabriel Asbury Pratt

Committee in charge:

Professor Gene Yeo, Chair
Professor Sheng Zhong, Co-Chair
Professor Jens Lykke-Andersen
Professor Simpson Joseph
Professor Bing Ren

2018

The dissertation of Gabriel Asbury Pratt is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____
Co-Chair

_____
Chair

University of California, San Diego

2018

DEDICATION

To my parents who taught me to be curious, my mentors who showed me the cool things to be curious about, my friends who kept me balanced through out all of this.

EPIGRAPH

Progress isn't made by early risers. It's made by lazy men trying to find easier ways to do something

—Robert Heinlein

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

To all my co-authors and lab mates I couldn't have done this without you. Specifically I want to call out Mike Lovci, who taught me how to analyze CLIP-seq data. Ive leaned on your tools more than you know. Katannya Kapeli who was the first person to drag me through a full biological story. Finally I wouldnt have finished my PhD with out Eric Van Nostrand, who taught me everything else, but most importantly taught me how to walk carefully through analyses instead of just running ahead. I wouldnt be half the researcher I am without out.

I also want acknowledge the people who put me on this path. My high school Genetics teacher Penny Pagels, who first introduced me to the field. Chuck Murry who took a chance on a first year undergrad who wanted to work in a lab. My two main undergraduate mentors, Jonathan Golob, who exposed me to everything bioinformatics could be and Sharron Paige, who let me help on the coolest projects.

A short acknowledgement, cannot properly recognize everyone who has helped me along the way, or encompass what Ive learned from everyone. I am eternally grateful for all the time and energy you all have invested, and hope to pay it back some day.

Chapter 2, in full, is a reprint of the material as it appears in Nature Communications 2016. Katannya Kapeli, Gabriel A. Pratt, Anthony Q. Vu, Kasey R. Hutt, Fernando J. Martinez, Balaji Sundararaman, Ranjan Batra, Peter Freese, Nicole J. Lambert, Stephanie C. Huelga, Seung J. Chun, Tiffany Y. Liang, Jeremy Chang, John P. Donohue, Lily Shiue, Jiayu Zhang, Haining Zhu, Franca Cambi, Edward Kasarskis, Shawn Hoon, Manuel Ares Jr., Christopher B. Burge, John Ravits, Frank Rigo, Gene W. Yeo Nature Publishing Group, 2016. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 3, in full, has been submitted for publication of the material as it may appear in Nucleic Acids Research, 2018. Gabriel A. Pratt, Eric L. Van Nostrand, Brian A. Yee, Alain Domissy, Steven M. Blue, Chelsea Gelboin-Burkhart, Thai B. Nguyen, Ines Rabano, Ruth Wang, Balaji Sundararaman, Keri Garcia, Rebecca Stanton, Gene W. Yeo. Nucleic Acids Research,

2018. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Molecular Cell 2015. Suzanne R. Lee, Gabriel A. Pratt, Fernando J. Martinez, Gene W. Yeo, Jens Lykke-Andersen. Elsevier, 2015. The dissertation/thesis author was an investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Nature 2016. Stefan Rentas, Nicholas T. Holzapfel, Muluken S. Belew, Gabriel A. Pratt, Veronique Voisin, Brian T. Wilhelm, Gary D. Bader, Gene W. Yeo, Kristin J. Hope. Nature, 2016. The dissertation/thesis author was an investigator and author of this paper.

VITA

2011	B. S. in Computer Science, University of Washington, Seattle
2018	Ph. D. in Bioinformatics and Systems Biology, University of California, San Diego

PUBLICATIONS

Gabriel A. Pratt, Eric L. Van Nostrand, Brian A. Yee, Alain Domissy, Steven M. Blue, Chelsea Gelboin-Burkhart, Thai B. Nguyen, Ines Rabano, Ruth Wang, Balaji Sundararaman, Keri Garcia, Rebecca Stanton, Gene W. Yeo. "Guidelines and Best Practices for enhanced CLIP experiments and analysis". *In Submission*, 2017

Eric L Van Nostrand, Peter Freese, Gabriel A Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M Blue, Daniel Dominguez, Neal A.L. Cody, Sara Olson, Balaji Sundararaman, Lijun Zhan, Cassandra Bazile, Louis Philip Benoit Bouvrette, Jiayu Chen, Michael O Duff, Keri E. Garcia, Chelsea Gelboin-Burkhart, Abigail Hochman, Nicole J Lambert, Hairi Li, Thai B Nguyen, Tsultrim Palden, Ines Rabano, Shashank Sathe, Rebecca Stanton, Julie Bergalet, Bing Zhou, Amanda Su, Ruth Wang, Brian A. Yee, Ashley L Louie, Stefan Aigner, Xiang-dong Fu, Eric Lecuyer, Christopher B. Burge, Brenton R. Graveley, Gene W. Yeo. "A Large-Scale Binding and Functional Map of Human RNA Binding Proteins" *In Submission*, 2017

Eric L Van Nostrand, Chelsea Gelboin-Burkhart, Ruth Wang, Gabriel A Pratt, Steven M Blue, Gene W Yeo. "CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins". *Methods*, 2016

Fernando J Martinez, Gabriel A Pratt, Eric L Van Nostrand, Ranjan Batra, Stephanie C Huelga, Katannya Kapeli, Peter Freese, Seung J Chun, Karen Ling, Chelsea Gelboin-Burkhart, Layla Fijany, Harrison C Wang, Julia K Nussbacher, Sara M Broski, Hong Joo Kim, Rea Lardelli, Balaji Sundararaman, John P Donohue, Ashkan Javaherian, Jens Lykke-Andersen, Steven Finkbeiner, C Frank Bennett, Manuel Ares, Christopher B Burge, J Paul Taylor, Frank Rigo, Gene W Yeo. "Protein-RNA Networks Regulated by Normal and ALS-Associated Mutant HNRNPA2B1 in the Nervous System." *Neuron*, 2016

Kristopher W Brannan, Wenhao Jin, Stephanie C Huelga, Charles AS Banks, Joshua M Gilmore, Laurence Florens, Michael P Washburn, Eric L Van Nostrand, Gabriel A Pratt, Marie K Schwinn, Danette L Daniels, Gene W Yeo. "SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes." *Molecular Cell*, 2016

Katannya Kapeli, Gabriel A. Pratt, Anthony Q. Vu, Kasey R. Hutt, Fernando J. Martinez, Balaji Sundararaman, Ranjan Batra, Peter Freese, Nicole J. Lambert, Stephanie C. Huelga, Seung Chun, Tiffany Y. Liang, Jeremy Chang, John P. Donohue, Lily Shiue, Jiayu Zhang, Haining Zhu,

Franca Cambi, Edward Kasarskis, Manuel Ares Jr., Christopher B. Burge, John Ravits, Frank Rigo, Gene W. Yeo. "Distinct and shared molecular targets and functions of ALS-associated TDP-43, FUS, and TAF15 revealed by comprehensive multi-system integrative analyses." *Nature Communications*, 2016

Stefan Rentas, Nicholas T Holzapfel, Muluken S Belew, Gabriel A Pratt, Veronique Voisin, Brian T Wilhelm, Gary D Bader, Gene W Yeo, Kristin J Hope. "Musashi-2 attenuates AHR signalling to expand human haematopoietic stem cells." *Nature*, 2016

Anne E Conway, Eric L Van Nostrand, Gabriel A Pratt, Stefan Aigner, Melissa L Wilbert, Balaji Sundararaman, Peter Freese, Nicole J Lambert, Shashank Sathe, Tiffany Y Liang, Anthony Essex, Severine Landais, Christopher B Burge, D Leanne Jones, Gene W Yeo. "Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival." *Cell Reports*, 2016

Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, Gene W Yeo. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)." *Nature Methods*, 2016

Balaji Sundararaman, Lijun Zhan, Steven M Blue, Rebecca Stanton, Keri Elkins, Sara Olson, Xintao Wei, Eric L Van Nostrand, Gabriel A Pratt, Stephanie C Huelga, Brendan M Smalec, Xiaofeng Wang, Eurie L Hong, Jean M Davidson, Eric Lcuyer, Brenton R Graveley, Gene W Yeo. "Resources for the comprehensive discovery of functional RNA elements." *Molecular Cell*, 2016

T. Hung, G. A. Pratt, B. Sundararaman, M. J. Townsend, C. Chaivorapol, T. Bhangale, R. R. Graham, W. Ortmann, L. A. Criswell, G. W. Yeo, T. W. Behrens. "The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression." Science, 2015

Suzanne R. Lee, Gabriel Pratt, Fernando Martinez, Gene W. Yeo, Jens Lykke-Andersen. "Target discrimination in nonsense-mediated mRNA decay requires the ATPase activity of Upf1." *Molecular Cell*, 2015

Singh G, Gabriel Pratt, Yeo GW, Moore MJ. "The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion." *Annual Review of Biochemistry*, 2015

Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, Liang TY, Stark TJ, Gehman LT, Hoon S, Massirer KB, Gabriel Pratt, Black DL, Gray JW, Conboy JG, Yeo GW. "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges." *Nature Structural and Molecular Biology*, 2013

Sharon L. Paige, Sean Thomas, Cristi Stoick-Cooper, Hao Wang, Richard Sandstrom4, Lisa Maves, Lil Pabon, Hans Reinecke, Gabriel Pratt, Gordon Keller, Randall T. Moon, John Stamatoyannopoulos, and Charles E. Murry. "A Temporal Chromatin Signature in Human Embryonic Stem Cells Identifies Novel Regulatorsof Cardiovascular Development." *Cell* 2013

Golob, J. L., Kumar, R. M., Guenther, M. G., Pabon, L. M., Pratt, G. A., Loring, J. F., Laurent, L. C., Young, R. A., and Murry, C. E. “Evidence That Gene Activation and Silencing during Stem Cell Differentiation Requires a Transcriptionally Paused Intermediate State.” *PLoS ONE*, 2011

ABSTRACT OF THE DISSERTATION

Methods for Integrative Analysis of RNA Binding Proteins

by

Gabriel Asbury Pratt

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California, San Diego, 2018

Professor Gene Yeo, Chair
Professor Sheng Zhong, Co-Chair

Cross-linking immunoprecipitation (CLIP) has been used to profile the binding sites of over 100 RNA binding proteins (RBPs). However computational pipelines, quality control metrics, and downstream analyses needed to process CLIP data at scale have yet to be well defined. Here we describe in detail the characterization of a single RBP, TAF15, which is known to be involved in amyotrophic lateral sclerosis. We detail computational processing techniques, including integration of RNA-seq, microarray splicing, RNA bind-n-seq (RBNS) and stability assays to understand the function TAF15 in mouse and human brains. Next we describe how to scale analyses from one RBP to many. We present our ENCODE eCLIP processing pipeline, enabling

users to go from raw reads to significant, reproducible peaks, that can be directly compared against ENCODE eCLIP experiments. In particular, we discuss processing steps designed to address common artifacts, including quantifying unique RNA fragments bound by both unique genomic- and repetitive element-mapped reads. Using manual quality annotation of 350 ENCODE eCLIP experiments, we develop metrics for quality assessment of eCLIP experiments before and after sequencing, including recommendations for library yield, number of unique fragments in library, binding information, and biological reproducibility. In particular, we quantify the linkage between sequencing depth and peak discovery, and derive methods for estimating sequencing depth based on pre-sequencing metrics. Finally we provide recommendations for the integration of RBP binding and RNA-seq experiments to generate splicing maps. These pipelines and QC metrics enable large-scale processing and analysis of eCLIP data, and enable rigorous and standard analysis of RBP binding data. Finally we describe results from analysis of additional RBPs that illustrate the utility of studying the dynamics of RBP binding in different contexts. Specifically we detail how understanding the location of UPF1 binding lead to a better understanding of the mechanism of action for UPF1 in nonsense medicated decay. We also detail how information on Musahi 2 binding improved understanding of the mechanism of haematopoietic stem cells expansion.

Chapter 1

INTRODUCTION

1.1 OVERVIEW

There are over 1,500 RNA binding proteins (RBPs) in the human genome [GHT14], RNA binding proteins regulate all aspects of RNA metabolism, including splicing, polyadenelation, stability, transport, and translation. Additionally they mediate other important interactions, controlling the function of microRNAs, repetitive elements and other cellular processes [NBLTY15]. Unsurprisingly, RBPs regulate the normal cellular function in many different cell types. Critically playing roles in cellular differentiation and immune response [Kim16, Lad16, KSK14]. Additionally dysfunction of RBPs or mutations in RPBs have been associated with neurodegenerative disease [NBLTY15, BC16] and cancer [PBA17]. In this introduction I will briefly describe links between RNA binding proteins and cellular function and dysfunction. Then review appropriate experimental and computational methods used to understanding the binding dynamics of RNA proteins. Finally I will review computational methods used to integrate hundreds of datasets offering a prospective on how these methods can be used for RBPs, and discuss the importance of quality control methods to underpin the accuracy of these findings. Detailed introductions for each project can be found at the start of each individual chapter.

1.2 OVERVIEW OF RBP FUNCTION

Splicing is one of the most well characterized direct functions of RBPs. The first experiments assessing direct interaction of RBPs with RNA identified that RBPs directly regulate alternative splicing [UJR+03, YCL+09]. RBPs regulate splicing in many ways, briefly RBPs can bind near exons and promote exon inclusion or exclusion, binding or lack of binding in intronic regions can promote cryptic splicing. Changes to 5 and 3 start sites, and intron retention has also been observed. In addition to specific modulation of alternative exons core splicing machinery is are also classified as RBPs, further demonstrating the need to better understand and characterize the function of these proteins. The characterization of all splicing effects and RBPs effecting splicing is a vast field, and has been well summarized in [FA14].

Localization of RNA to correct subcellular locations is also controlled by RBPs. RBPs have been implicated in nuclear and cytoplasmic retention, and shuttling proteins in both directions between the nucleus and the cytoplasm [BB14]. Of special interest are the SR proteins, previously thought to be purely splicing factors, have also been implicated in the transport of RNA from the nucleus to the cytoplasm [Jeo17].

Translation and translational control are important functions for RNA binding proteins. The Ribosome is an RNA-protein complex, which binds to mRNAs, and as such falls in the field of study. Additionally there are RBPs that modulate translation. FMRP is a known translational repressor [DVZ+11], binding and slowing the rate of translation of target mRNAs. Conversely RBPs can also promote translation, eIF3, under stress conditions can work in conjunction with m6a writers bind and promote translation of specific RNAs [MPZ+15].

Stability is regulated by RBPs through both direct interactions, and indirect actors. A well characterized method of regulation is the nonsense mediated decay (NMD) pathway, where UPF1 and co-factors recognize that a transcript hasnt fully been translated, and target those transcripts degradation [KJ12]. Additionally the microRNA processing and decay pathway by definition is

regulated by RNA binding proteins, plays a major role in the degradation of specific regulated targets. [JT06].

Cleavage and Polyadenylation occurs after transcription, and core group of approximately 20 CSPF proteins to process each transcript. In addition to this basic processing there are a number of alternative polyadenylation factors the can fine tune aspects of this process [SM00, EUA13]. It is important to note that both the length of the polyA tail, and the location of polyadenylation can be regulated by RBPs.

RNA editing and modification is just beginning to be explored. Currently two main RNA modifications have been studied. A to I editing which is edited by the ADAR family [Nis16], and m6a modification, which are regulated by a group of m6a readers, writers and erasers [LPK17]. Both these modifications can corporate with other RBPs to regulate RNA metabolism. We have only just begun to scratch the surface of RBPs that read and write RNA modifications. To date over 100 RNA modifications have been discovered, and most of their functions have yet to be described [CCR$^+$11].

1.3 RBPs IN DISEASE

RBPs have been associated with neurodegenerative diseases [NBLTY15]. Of special interest to this dissertation are RBPs involved in amyotrophic lateral sclerosis (ALS). ALS is a deadly neurodegenerative disease characterized by progressive death of motor neurons.

First, mutations in TDP-43 (Tardbp), an RBP, were found to be associated with ALS [AHA$^+$06]. TDP-43 is normally localized to the nucleus and regulates both normal and cryptic splicing [PLTH$^+$11, LPTW15]. Cytoplasmic aggregations of TDP-43 are a hallmark of ALS [BB12]). Because of this dual dysregulation of splicing, and potential cytotoxicity of cytoplasmic aggregations it has yet to be determined if the death of motor neurons is caused by loss of splicing function or gain of toxicity in the cytoplasmic aggregates.

Causative mutations have also been found in FUS/TLS [VRH+09]. Characterization of FUS/TLS revealed that the RBP also in the nucleus, in ALS it is localized into cytoplasmic aggregates, and loss of function of FUS/TLS also causes splicing changes [LTPC10].

FUS/TLS is part of the FET family of proteins, FUS/TLS, EWS and TAF15. As a result it was through that the rest of the FET family might play a role in ALS. TAF15 was discovered as a potential cause of ALS in a yeast screen [CHS+11]. Mutations in TAF15 in ALS patients were also discovered [TVL+11]. The characterization of functional effects of TAF15 have been less conclusive. TAF15 may regulate splicing changes [IMA+13, HLR+11], although analysis in relevant human cell lines has been inconclusive [KPV+16].

It should be noted that the majority of ALS cases are now explained by a hexanucleotide expansion in C9ORF72 [RMW+11, DHMB+11], mutations in this protein still led to cytoplasmic aggregates of RBPs, leaving open the question of loss or gain of function issues for FUS, TAF15 or TDP-43, or some other effect of C9ORF72.

1.4 METHODS TO IDENTIFY RBP BINDING

A low throughput way to identify RBP binding is an EMSA assay. This is good validation, but does not address the issue of identifying binding sites at scale, or in normal folded context. To address the problem of identifying RBP motifs in a high throughput manner array and high throughput sequencing approaches were applied.

The first such approach was RNACompete [RKC+09, RKC+13], which used a microarray hybridization method to identify RBPs that bound preferentially to specific sequences. This approach was used to catalogue the binding specificities of over 205 different RBPs, and helped initially classify common RBP motifs and diversity.

An alternative approach, RNA bind-n-seq (RBNS), was developed to take advantage of high throughput sequencing approaches [LRJ+14]. In this approach an the RNA binding domain

of a protein of interest is cloned, combined with a purification tag, incubated with random RNA sequences, purified to obtain preferentially bound RNA sequences and sequenced. This approach has since been used understand the binding preferences of the IMP family [CVP$^+$16], TAF15 and FUS [KPV$^+$16], hnRNPA2B1 [MPV$^+$16], PPR10 [MRM$^+$17], the Musashi family [KLL$^+$14]. Additionally large-scale integrative analysis of RBNS data has also been performed [DFA$^+$17].

RIP and RIP-seq were initially developed to enable broad understanding of RBP-mRNA interactions. Because there is no fragmentation step in RIP-seq, it is impossible to determine the specific binding sites for a specific RBP, however it is possible to determine general mRNA targets of a specific RBP [TCLK00, KKF06]. A strength of RIP-seq is that that it is semi- quantitative, allowing for the estimation of binding strength of a RBP to a mRNA [LPM$^+$15]. Due to a lack of crosslinking, lysis of RIP-seq experiments may lead to issues with re-association of an RBP to an incorrect RNA [MS04] formaldehyde crosslinking followed by RIP-seq has allowed for an increase in spatial resolution in RIP-seq, allowing more accurate identification of RBP-RNA interactions [HKT$^+$16]. Recently DO-RIP-seq has been developed, which optimized the digestion conditions for RIP-seq allowing for even higher spatial resolution, and accurate quantification of the strength of RBP binding than has been previously reported [NFBK17].

To address the issue of spatial resolution in RIP-seq, and allow for the identification of the specific location of RBP-RNA interaction, CLIP-seq (or HITS-CLIP) was developed [UJR$^+$03]. CLIP-seq UV-crosslinks RBPs to RNAs before fragmentation, removing the issue of post-lysis reassocation, and also allows for an understanding of the specific binding targets of an RBP. CLIP-seq has been widely applied to understand the binding profiles of many RBPs, including RBFOX2 [YCL$^+$09], the hnRNP family [HVA$^+$12], FUS, TAF15 and TDP43 [LTPH$^+$12, KPV$^+$16, PLTH$^+$11], and many others. Although CLIP has been successful in helping understand the genome-wide function of RNA binding proteins, significant technical challenges remained. CLIP suffers from low cross-linking efficiency, and low complexity libraries. Both of which limit the ability for a new biology to be discovered.

PAR-CLIP was developed due to the low crosslinking efficiency of RBPs under standard crosslinking conditions. PAR-CLIP incorporates a nucleoside analog and crosslinks as a different wave length than regular CLIP [HLB$^+$10]. While this approach has significant advantages in terms of crosslinking efficiency, and library recovery, it suffers due to the inclusion of the nucleoside analog, limiting the protocol to use only in in-vitro settings. Additionally the analog itself is an analog to uracil which makes PAR-CLIP only work on RBPs that bind in proximity to a U on the genome.

iCLIP was developed to try to fix the issue of low library complexity in CLIP-seq [KZR$^+$10]. It pioneered two experimental improvements. First, traditional CLIP-seq uses two ligations to attach the 5 and 3 illumina adapters. iCLIP performs circular ligation followed by restriction to reduce the number of ligations steps to one. Additionally to address the sill present issue of low library complexity a unique molecular identifier (UMI) is attached to each read to unambiguously identify the origin of each read sequenced. As a result of these improvements iCLIP reads tend to stop at the direct site where an RBP interacts with RNA. In CLIP reverse transcription is performed after the ligation of the two adapters and the RT tends to fall off at the crosslinking site. Even when successful there are frequently deletions or mutations at the crosslinking site. Although it is estimated RT after CLIP crosslinking yields approximately 15% of initial bound fragments. Because the 5 end of the read in iCLIP should be where the RT falls off (at the adduct site) iCLIP allows for the identification of the exact location of RBP-RNA interactions.

Although iCLIP and PAR-CLIP address the original CLIP protocols shortcomings, there were still issues with library efficiency, amazingly even with iCLIP improvements often times libraries needed to be amplified more than 20 PCR cycles to get an appropriate library concentration. To address this, two new approaches were recently developed. Infared-CLIP (ir-CLIP) [ZFS$^+$16], which increases the speed of the traditional CLIP protocol, while allowing for recovery of a more complex library, and enhanced-CLIP (eCLIP) [VPS$^+$16], which adjusts ligations

conditions to allow for the recovery of a more complex library than previously observed. In addition to increasing library complexity eCLIP introduced a sized matched input, which allows improved normalization and increased signal when identifying peaks compared to CLIP alone.

Interestingly, unlike in ChIP-seq, IgG as an input control works poorly. IgG eCLIP libraries are low complexity, and uninformative with regards to identification of significant peaks [VNGB$^+$17]. Where as a size matched input control significant improves the ability of eCLIP to identify true positive peaks (Van Nostrand et al. 2016)

In addition to improvements in CLIP efficiency, there have been two attempts to adapt CLIP to identify RNA-RNA interactions, CLASH [GFM$^+$14, HKDT13], CLEAR-CLIP [MSL$^+$15]. These methods have succeeded in identifying miRNA-mRNA interacting partners mediated by Ago, however, they have a very low efficiency. Containing only thousands of reads (out of millions sequenced) that support these those observations. Interestingly more recent attempts have identified these same interactions in c elegans using unmodified iCLIP with similar efficiency [BLH$^+$16].

1.5 CLIP COMPUTATIONAL METHODS

Concurrent with the development of experimental methods to better identify RBP-RNA interactions computational methods have also been developed to improve the ability to gain information from CLIP-seq experiments. Here I briefly review key computational improvements to the CLIP field.

1.5.1 UNIQUE MOLECULAR IDENTIFIERS (UMIs)

UMIs for CLIP were initially developed in coordination with iCLIP [KZR$^+$10]. Initial approaches to identifying unique molecules were simple. Collapsing reads if they contained the exact same UMI. This approach contains drawbacks due to the inability to identify reads that are

identical, but have different UMIs due to errors during PCR or during sequencing. Because unique molecules sometimes have more than 1000 reads assigned to a signal molecule, and Illumina sequencing error rates are on average 0.5% this can be a significant issue. To address this attempts were made to model mutation rates and intelligently collapse UMIs that were close to each other, using an expectation maximization approach. [DVZ+11]. More recently simple graph based approaches have been applied to detect close UMIs and collapse them [SHS17]. Although these approaches were first developed to address short comings with CLIP-seq data UMIs have more recently seen broader application to RNA-seq, specifically single-cell sequencing approaches [IZJ+14].

1.5.2 PEAK FINDING TOOLS

In addition to removing technical noise, a main objective in CLIP-seq analysis is to identify sites of enriched binding, known as clusters or peaks. Although similar in method to ChIP-seq peak finding techniques such as MACS or GPS/GEM [ZRS+08, GPA+10], CLIP-seq peak calling approaches need to be customized for the protocol due to, a lack of input controls, strand specificity of the protocol, and additional need to control for variation in coverage between intronic and exon regions of a gene.

Initial CLIP-seq methods contained so few reads that complicated methods were not needed. Indeed two of the first CLIP-seq papers simply counted reads in regions, compared to a random background across the gene, and calculated significance [UJR+03, YCL+09]. A small improvements on this approach were formalized in a general CLIP analysis tool kit [AGVB+11].

For vanilla CLIP it was observed that RT over the crosslinking adduct often times leads base skipping during RT. This observation was used to increase the confidence in regular CLIP-seq experiments by calling peaks on Crosslinking induced mutations (CIMS) [ZD11].

PAR-CLIP also has telltale signs true UV crosslinking occurred, also known as a diagnostic event. In the case of PAR-CLIP the nucleoside used converts T-¿C during sequencing, Areas

where this transition were observed to more frequently than background were assumed to be crosslinking sites [CGM+11]. Further refining this method waveR [SSS+12] was developed. waveR applies a mixture model to regions of potential peaks, to determine the rate of T-¿C mutations is likely to come from a normal mutational process or is indicative of RBP binding. This approach for PAR-CLIP has high spatial resolution, but relies on T-¿C transitions, so is not generally applicable to other CLIP experiments.

miCLIP [WCK+14], again builds upon the mixture model developed in waveR, but generalizes its approach to be generally useful for all CLIP-seq type experiments. It does this by solving the peak detection problem as an hidden markov model, calling regions as either bound or not bound based on their relative enrichment within the genome. Then for each base within a bound region an analysis of mutational frequency can be optionally applied to identify direct sites of RBP-RNA interaction.

Another way to generalize this approach is to combine this idea of detecting diagnostic events Piranah created a more general framework that takes in any type of diagnostic event, and used a zero-truncated negative binomial distribution to identify sites of significant binding [UBSB+12]. Although this approach is general, it creates a background distribution across the entire genome rather than specifically for each RNA family, which, given different gene expression levels of RNAs creates a large number of false positives.

Other groups have developed methods to identify peaks based on edge detection approaches [LKC+12], although this approach has not been widely adapted.

Finally there has been one attempt to identify differential RBP binding between two experiments or conditions. dCLIP [WGB+14], does this by extending the HMM approach used by miCLIP to identify bound regions. It defines an HMM with three states, enriched in experiment 1 over experiment 2, equally enriched, or enriched in experiment 2 over experiment 1.

1.5.3 MOTIF CALLING

A main question after peak calling for CLIP-seq is what sequence does the RBP frequently bind. Unlike peak calling standard approaches are appropriate to apply in this situation. CLIP-seq peaks may be analyzed using HOMER, MEME or other motif finding algorithms [HBS+10, Bai11, Boe16]. Unfortunately RBPs tend to have highly degenerate motifs and short motifs, especially relative to highly specific motifs identified in ChIP-seq [KAS+07, DFA+17].

Although motif finding has been difficult, incorporating structural information into approaches has allowed for a better understanding of RBP binding. For example it was discovered that sequence motifs were more easily detectable from RIP-seq experiments when accessibility of punitive regions was taken into account [LQLM10]. RBPs may frequently bind both sequence and structure motifs simultaneously. For example LIN28 only binds a GGAGA sequence motif embedded within a short hairpin [WHK+12]. Few attempts have been made to identify these structural motifs at a wide scale.

RNAContext [KRC+10], was developed to be used with RNA affinity approaches. It combines secondary structure and primary sequence information in an expanded PWM model to identify potential structural motifs.

GraphProt [MLCB14] extended this approach by using CLIP-seq data to help identify structural motifs, instead of just RNA affinity data. GraphProt takes confident CLIP-seq peaks and applies a graphical model to cluster and identify common sequence and structural motifs near the RBP binding site.

Zagros [BSPSU15]. Also attempts to predict sequence and structural motifs, adds in the concept of diagnostic events to further refine its spatial resolution. It uses traditional motif finding algorithms to identify enriched sequences, but extends the idea of a sequence to the structure of the RBP as well.

These approaches have had limited success perhaps due to low quality of input data (all approaches listed were evaluated using CLIP-seq, not eCLIP or irCLIP data), structural motif

identification is still an active field and should continue to be pursued. Although RNA-structure predictions are accurate for small structures, predictions of secondary structures for molecules as large as mRNAs is still inaccurate. The advent of structural sequencing [LMT+11, FZS+16, SLA+14] may further allow for the identification of accurate RBP secondary structure motifs.

1.5.4 SPLICING MAPS

RBP binding information is can also be integrated with other data sources to provide hints as to the mechanism of RBP regulation. Commonly RBP binding or motif information is visualized with splicing changes [WU11]. Original splicing maps were made visualizing motif presence overlaid with an estimate (either via micro arrays or RNA-seq) of alternative splicing changes [UJR+03, YCL+09]. As CLIP-seq methods improved direct CLIP-seq peaks were overlaid with splicing changes for more direct splicing regulation information [KZR+10, HVA+12, KPV+16]. Recently due to additional background normalization techniques we have started to create splicing maps using normalized read density, rather than peaks (Pratt et. al 2017, In Submission).

1.5.5 CLIP PIPELINES

CLIP analysis is complicated, requiring detailed knowledge of adapter trimming, specific mapping parameters, custom PCR duplicate removal and specialized peak calling methods (as reviewed above), to perform correctly. As a result pipelines have been written to take users from raw fastq files to mapped and analyzed peaks with minimal difficulty.

PIPE-CLIP [CYK+14] is an online server that takes in raw fastq data and outputs processed peaks. Although it is inflexible in terms of customization it effectively provides high quality peaks.

There are two other pipelines that must be run locally that also provide similar services.

CLIPSeqTools [MANM16], is a comprehensive, flexable toolkit that allows for automated calling of peaks. In addition it provides some data analysis and visualization options after peaks have been called.

Conversely CLIP Toolkit (CTK) [SQWVZ17], also provides a unified structure to go from fastqs to called peaks, but does not include downstream analysis. It wraps one of the more successful peak calling algorithms (CIMS analysis), and provides a new peak and valley finding algorithm that is accurate on datasets where CIMS cannot be performed.

1.5.6 CLIP-SEQ DATABASES

Finally there have been databases constructed to catalogue and uniformly process all previously published CLIP-seq datasets. Structured archival and re-analysis of large public datasets is sorely needed for many fields, including CLIP. Quick re-analysis of existing data allows for quick checks to see if a user show follow up on a hypothesis, and extensive re-analysis allows for novel biological inferences to be made. For example a re-analysis of all RNA-seq data on GEO identified the theoretical maximum number splice junctions in the human genome [NJF+16]. Structured re-analysis of ChIP-seq data has allowed for the identification of functional segments of the genome, defined by combinations of histone marks and binding of transcription factors [HBW+12, EK12a].

CLIPZ [KRZ11] mixes database archiving of public data, while allowing users to upload and analyze their own results. StarBase [LLZ+14, YLS+11] was primarily developed to understand miRNA-RBP interactions, and support integrated analysis of different types of transcriptome sequencing data such as CLIP-seq, CLASH, and Degradome-seq. Its individual datasets, and data access however enable a wide range of queries to the database.

1.6 INTEGRATIVE INSIGHTS INTERESTING TO THE RBP FIELD

While the information gained by integrating data and bioinformatics approaches from many different RBPs into one study is only now being seen [VGBW$^+$17], we can understand the direction of the field by looking at the ChIP-seq field, which has been performing integrative analysis for many years to understand the effects of transcription factors and histone modifications in development and disease [HSH$^+$07, PTSC$^+$12]. The ChIP-seq field has performed many additional analyses of health and disease, to date the largest integrative analysis has been performed by the ENCODE project [BSD$^+$07, FGG$^+$04, DKA$^+$12].

Of special interest to the RBP field, the ENCODE project initially characterized the binding motifs of 119 transcription factors in human cell lines [WZI$^+$12, AANL12, WWC$^+$12]. These methods to analyze single TFs are generally applicable to the RBP field, and while there has not been rigorous cross-application of these methods yet, we expect a rigorous re-analysis of eCLIP RBPs for motifs to occur soon.

In addition to characterization of single transcription factors the ENCODE project allowed for the characterization of functional groups of histone modifications. Two separate approaches, Segway and chromHMM [HBW$^+$12, EK12b], have been used to perform semi-supervised clustering on histone modifications. Segway did this by training a Bayesian network, while chromHMM trained a hidden markov model to understand different chromatin states. Both approaches are effective, with Segway having slightly higher spatial resolution. Other differences in the two methods can be found in [HEW$^+$13]. These methods have since been widely applied, underpinning many approaches used to impute missing data and reduce the number of ChIP-seq experiments needed to be performed in the future [EK15].

Finally of interest to future efforts for the ENCODE RBP project integration of TF binding and SNP data has been used to predict functional, disease causing variants. The addition of RBP

binding information into these systems may help elucidate the effects of SNPs not previously seen as functional [WK12].

1.7 QUALITY CONTROL

Underpinning all the above methods is stringent quality control metrics employed by the ENCODE project. Without high quality data underpinning everything, the results would have been much harder to interpret. The ENCODE TF and histone group summarized their experimental and computational quality control metrics in a single paper [LMK+12]. On the experimental side they proposed stringent antibody validation.

They also proposed stringent standards for the how replicates should be handled, and the number of experimental replicates (2) needed to have confidence in the results of an experiment.

Computationally they developed 4 metrics to determine if an experiment has failed. First they look at RSC and NSC, metrics to validate that sequence reads are evenly distributed between the positive and negative strands of the genome, especially near RBP binding sites. Next they observe the fraction of reads in peaks (FRiP), effectively confirming that the majority of information is concentrated in punitive peak locations. Finally they use an IDR [LBHB11] based method to access the reproducibility of biological replicates using two metrics, Rescue Ratio, which verifies that two biological replicates have approximately the same set of peaks, and self-consistency ratio, which verifies that two biological replicates have approximately the same number of peaks. If all these metrics pass then an experiment passes ENCODE quality control, and if it fails then it is not submitted for public use or internal analysis. Re-analysis of all published ChIP-seq data using these quality control metrics has indicated that over 20% of public data is of low quality [MKPW14], further enforcing the need for standards in high throughput sequencing datasets.

In addition others have proposed their own analogous ChIP-seq quality control metrics

[KTP08], and importantly have tried to estimate the required number of sequenced reads needed to obtain a high quality ChIP-seq experiment [JLH+14].

The RBP field has proposed and catalogued similar experimental quality control metrics [SZB+16]. However computational quality control metrics are lacking. So far there have only been two papers directly addressing questions of computational in CLIP-seq data. Both papers review all possible approaches to CLIP-seq analysis, but provides no strong recommendations as to the ideal processing pipeline, and neither provide strong pass / fail metrics or an understanding on what makes a high quality CLIP-seq dataset [LZM+15, WXC+15].

Chapter 2

Distinct and shared functions of ALS-associated proteins TDP-43, FUS and TAF15 revealed by multisystem analyses

2.1 ABSTRACT

The RNA binding protein (RBP) TAF15 is implicated in amyotrophic lateral sclerosis (ALS). To compare TAF15 function to that of two ALS-associated RBPs, FUS and TDP-43, we integrate CLIP-seq and RNA Bind-N-Seq technologies, and show that TAF15 binds to 4,900 RNAs enriched for GGUA motifs in adult mouse brains. TAF15 and FUS exhibit similar binding patterns in introns, are enriched in 3 untranslated regions, and alter genes distinct from TDP-43. However, unlike FUS and TDP-43, TAF15 has a minimal role in alternative splicing. In human neural progenitors, TAF15 and FUS affect turnover of their RNA targets. In human stem cell-derived motor neurons, the RNA profile associated with concomitant loss of both TAF15 and FUS resembles that observed in the presence of the ALS-associated mutation FUS R521G, but contrasts with late-stage sporadic ALS patients. Taken together, our findings reveal convergent

and divergent roles for FUS, TAF15 and TDP-43 in RNA metabolism.

2.2 INTRODUCTION

Amyotrophic lateral sclerosis (ALS) is a fatal disease characterized by progressive degeneration of motor neurons in the motor cortex, brainstem, and spinal cord. Although the precise pathogenesis of ALS remains unknown, aberrant RNA processing appears to be an important contributing factor. The RNA binding protein (RBP) TAR DNA-binding protein 43 (TDP-43) was initially recognized as a major constituent of pathological ubiquitinated protein aggregates in brain and spinal cord tissue of patients with sporadic ALS (sALS)[AHA$^+$06, NSK$^+$06]. Dominant mutations in TDP-43 were subsequently identified in ALS patients[CRG$^+$09, DVK$^+$09, DGC$^+$09, KSV$^+$08, LRH$^+$09, RZB$^+$08] with evidence that these mutations were indeed causative of ALS pathogenesis[SBT$^+$08]. Shortly thereafter, mutations in the gene encoding another RBP, fused in sarcoma (FUS, also known as translocated in liposarcoma or TLS), were identified in a subset of patients with familial ALS (fALS) and sALS[VRH$^+$09, KBL$^+$09]. Although mutations in FUS and TDP-43 are present in only a small fraction of ALS cases, abnormal activity of FUS and TDP-43 is observed in a large fraction of ALS cases.

The discovery of mutations in the genes encoding TDP-43 and FUS received much attention as these proteins have strikingly similar protein domain architectures[LTPC10]. This motivated a search for more structurally similar RBPs as candidate ALS genes and as a result, mutations in TATA box binding protein (TBP)-associated factor 15 (TAF15) were identified in patients with sALS and fALS[CHS$^+$11, TVL$^+$11]. FUS and TAF15 belong to the FET family of hnRNP proteins, which includes Ewing sarcoma breakpoint region 1 (EWSR1). As the protein structure of TAF15 is similar to that of FUS and TDP-4313, it was predicted that TAF15 would be functionally similar to these RBPs. Similar to FUS and TDP-43 , TAF15 is predominantly localized to the nucleus but shuttles to and from the cytoplasm, participates in transcription, is

thought to affect alternative splicing, and has been found to form cytoplasmic inclusions in all FUS-FTLD subtypes and in some sALS patient tissues[CHS$^+$11, IMA$^+$13, NBD$^+$11].

Another commonality among TDP-43, FUS, and TAF15 is that the vast majority of ALS-associated mutations identified in the genes encoding these RBPs are found in their C-terminal Gly-rich domains. An emerging hypothesis is that mutations within the Gly-rich region of these RBPs promote their pathological aggregation[VSN$^+$13, GCZ$^+$11]. Aggregation of FUS, TDP-43, and TAF15 proteins is often accompanied by loss of their nuclear localization, yet it is unclear whether protein aggregation or mislocalization to the cytoplasm is the initiating pathogenic event[SCP14]. In efforts to investigate the normal nuclear function of these RBPs, comprehensive RNA binding maps of TDP-43, FUS, and TAF15 in the normal mouse[REB$^+$12, LTPH$^+$12, PLTH$^+$11] or human[IMA$^+$13] central nervous system (CNS) have been determined. These studies revealed global roles for TDP-43, FUS, and TAF15 in alternative splicing and motif specificities for TDP-43 and FUS in the CNS. Furthermore, loss of TDP-43 or FUS expression affects the RNA levels of genes containing long introns[REB$^+$12, LTPH$^+$12, PLTH$^+$11]. Our understanding of FUS, TDP-43, and TAF15 function in RNA processing has primarily come from examining these proteins individually and under different conditions, making comparisons difficult. This approach has limited our understanding of how the activities of these RBPs may converge on common pathways or act in parallel. A systematic comparison of FUS, TDP-43, and TAF15 to determine their shared and unique functions in mature and developing neurons would be valuable in understanding their contribution to development and ultimately disease.

Here we identify 4,873 RNA targets of TAF15 in mouse brain that reveal a TAF15 binding motif. Expanding on our previous studies[LTPH$^+$12, PLTH$^+$11], we find that FUS and TAF15 exhibit similar global RNA interaction profiles in vivo, but affect a strikingly small subset of common genes. Unexpectedly, TAF15 influences a small fraction of alternative splicing events compared to TDP-43 and FUS in the mouse CNS. In human neural progenitor cells, we find that TAF15 and FUS affect the stability of distinct mRNA populations, many of which are bound

by TAF15 and FUS. Depletion of TAF15, FUS, and TDP-43 in human induced pluripotent stem cell (iPSC)-derived motor neurons also affects different genes. Subsets of TAF15 and FUS-regulated mRNAs, including ALS associated genes, are also differentially expressed in spinal cord motor neurons dissected from sALS patients and iPSC-derived motor neurons from ALS patients harboring a R521G mutation in FUS. Taken together, these findings uncover points of functional convergence and divergence of FUS, TAF15 and TDP-43.

2.3 RESULTS

2.3.1 TAF15 binds RNAs enriched for GGUAAGU motifs in vivo

To identify in vivo RNA substrates recognized by TAF15, we performed CLIP (cross-linking immunoprecipitation)-seq in whole brain tissue from adult mice using a commercially available antibody that specifically recognizes the N-terminus of the TAF15 protein. We isolated RNA from low and high molecular weight TAF15 protein-RNA complexes (Fig. 1a, bands A and B, respectively) and converted the RNA into sequencing libraries for transcript identification. No protein-RNA complexes were immunoprecipitated when using non-specific IgG or in the absence of UV crosslinking (Supplementary Fig. 1a). Interactions between TAF15 and FUS have previously been detected[TGE$^+$13, SLQ$^+$15]. Therefore we tested whether TAF15 and FUS interact post cell lysis. Uniquely tagged versions of TAF15 and FUS proteins were expressed separately in HEK293T cells, and upon mixing lysates from these cell lines, we found that V5-tagged TAF15 immunoprecipitates Myc-tagged FUS (Supplementary Fig. 1b, lane 10) and vice versa (Supplementary Fig. 1b, lane 14). This demonstrated that TAF15 and FUS can physically associate post cell lysis. For our TAF15 CLIP-seq experiments, the use of UV crosslinked cells and highly stringent lysis and wash conditions prevented co-immunoprecipitation of FUS (Supplementary Fig. 1a) and TDP-43 (Supplementary Fig. 1c) with TAF15, ensuring that FUS-RNA and TDP-43-RNA complexes were not inadvertently recovered. Given the high overlap in

sequence similarity between the TAF15 target RNAs isolated from the low (band A) and high (band B) molecular weight complexes (Supplementary Fig. 1d), the libraries were combined (Fig. 1b and Supplementary Fig. 1e), resulting in 5.9 million non-redundant sequenced reads that mapped to 13,633 annotated protein-coding pre-mRNAs having more than 10 reads (5,128,815 reads, 85.8%), non-coding genes (139,382 reads, 2.3%), and intergenic regions (706,897 reads, 11.8%) in the mouse genome (mm9).

Using a published cluster-finding algorithm[PLTH$^+$11], we identified 47,138 TAF15 binding clusters in 4,873 genes. We applied the HOMER algorithm to these clusters to discover in vivo TAF15 binding motifs. The consensus motif GGUAAGU was statistically significantly enriched in TAF15 clusters (Fig. 1c, p¡10-535). Interestingly, this motif is similar to the 5 splice site sequence, GURAGU[SS87], however enrichment of this motif in both the coding sequence and the 3UTR provided evidence that we did not inadvertently extract the 5 splice site sequence within introns. Distribution analysis also illustrated that the TAF15 motif is enriched within the center of the CLIP clusters in the transcriptome (Fig. 1d) and also within 3UTRs (Fig. 1e). We searched for the TAF15 motif in clustered reads from published FUS21 and TDP-4322 CLIP-seq experiments and observed that the TAF15 motif was also enriched in transcriptome-wide FUS CLIP clusters and, to a lesser extent, in TDP-43 CLIP clusters residing in 3UTRs (Figs. 1d and 1e). We conclude that TAF15 interacts with binding sites enriched for a GGUAAGU motif within thousands of genes in vivo.

2.3.2 RNA Bind-n-Seq reveals TAF15 binding to GGUA motif in vitro

To characterize the in vitro sequence specificity of TAF15, we applied RNA Bind-n-Seq (RBNS)[LRJ$^+$14] to recombinant TAF15 and as a comparison, to recombinant FUS protein. Briefly, truncated forms of recombinant TAF15 or FUS containing both the RNA recognition motif and zinc finger domain (amino acids 204-415 for TAF15 and amino acids 235-481 for FUS) were incubated with an RNA pool consisting of random 20mer RNAs flanked by short

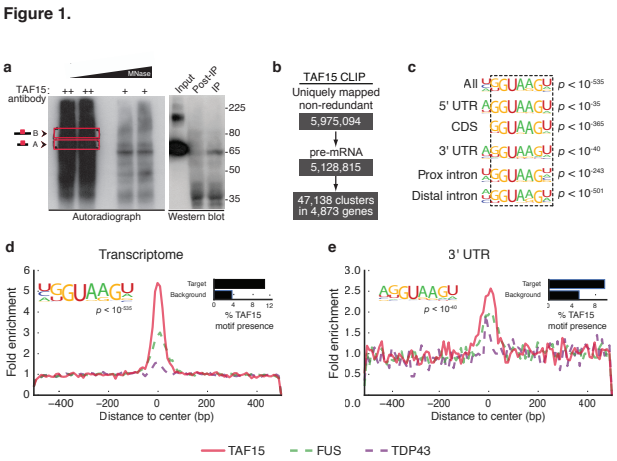

Figure 2.1: **Figure 1. CLIP-seq reveals that TAF15 binds GGUAAGU motifs in the mouse brain** (a) Autoradiograph of TAF15 protein-RNA complexes from the mouse brain immunoprecipitated with an antibody against TAF15 (left panel). RNA residing in the regions outlined by the red boxes was recovered for sequencing. TAF15-RNA complexes migrated at the expected size and were efficiently recovered because little protein remained in post-immunoprecipitation lysate (right panel, middle lane). (b) Flow-chart illustrating CLIP-seq reads analyzed to define TAF15 clusters. (c) De novo sequence motifs enriched above background within the transcriptome or specific genic regions with associated binomial p values. (d) Positional distribution of the TAF15 motif GGUAAGU within TAF15 (red), FUS (green), or TDP-43 (purple) CLIP clusters. Inset graph shows the percent enrichment of the TAF15 motif GGUAAGU within TAF15 targets (Target) or within the transcriptome (Background). (e) Positional distribution analysis of the TAF15 motif GGUAAGU as in (d) but specifically within CLIP clusters residing in 3UTRs. Inset graph shows the percent enrichment of the TAF15 motif GGUAAGU within the 3UTRs of TAF15 targets (Target) or within the 3UTRs of the transcriptome (Background).

primers used to add adapters for high-throughput sequencing (Fig. 2a). For FUS, this truncated region was previously shown to exhibit high affinity for RNA[YZK^{+}15]. Complementary to in vivo interactions identified by CLIP-seq, this method evaluates TAF15 and FUS independently of its in vivo complex-interaction with RNA. For TAF15, RBNS discovered degenerate G-rich and GU-rich motifs and notably an (A/G)GGUA motif that resembled the GGUAAGU motif that was identified in vivo by CLIP (Fig. 2b). In fact, the shared GGUA 4mer was significantly enriched in hexamers that were overrepresented in both RBNS and TAF15 CLIP-derived clusters relative to the appropriate control backgrounds (Fig. 2c). Additionally, the same GGUA motif was enriched in the TAF15 clusters located within 3UTRs of target genes (Supplementary Fig. 2a). RBNS applied to FUS domains identified a similar degenerate G-rich motif, a GC-rich

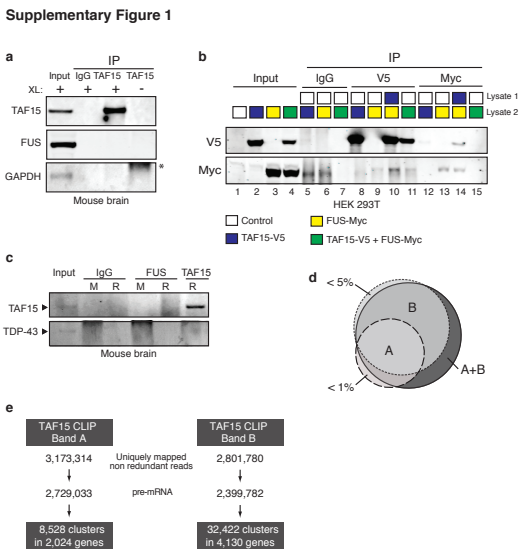

Figure 2.2: **Supplementary Figure 1. CLIP-seq for TAF15 in the mouse brain** (a) Western blot to validate that FUS is not co-immunoprecipitated by TAF15 using an anti-TAF15 antibody from mouse brain tissue that was (+) or was not (-) subjected to UV crosslinking (XL). IgG antibody served as a negative control. The asterisk corresponds to a non-specific band. (b) Post-lysis interaction by ectopically expressed TAF15 and FUS. Cell lysates from HEK293T cells expressing an empty control plasmid (white), V5-tagged TAF15 (blue), Myc-tagged FUS (yellow), or both TAF15-V5 and FUS-Myc (green) were mixed in various paired combinations as indicated. Mixed lysates were used for immunoprecipitation with IgG, anti-V5, or anti-Myc antibodies and immunoprecipitated proteins were analyzed by Western blot for V5 and Myc expression. (c) Co-immunoprecipitation in mouse brain tissue that was subjected to UV crosslinking using two different anti-FUS, an anti-TAF15, or IgG control antibodies. Antibodies were either produced in mouse (M) or rabbit (R). (d) Venn diagram showing the overlap in CLIP peaks between the A and B TAF15-RNA complexes. (e) Flow-chart showing the breakdown of sequence reads for the A and B TAF15-RNA complexes.

motif, and a GGUGG motif (bottom motif in Fig. 2d) that resembled motifs identified in published in vivo CLIP studies[REB$^+$12, LTPH$^+$12]. A similar evaluation of the GUGG 4mer (or GGUG, not shown) confirmed enrichment within FUS in vivo CLIP-seq-derived clusters in the transcriptome (Fig. 2e) and in 3UTRs (Supplementary Fig. 2b). Interestingly, we found that the distribution of the GUGG 4mer was also enriched in the hexamers derived from the TAF15 RBNS experiment (Supplementary Fig. 2c). Similarly, the GGUA 4mer was enriched in the FUS hexamers (Supplementary Fig. 2d). Although both of these motifs were found at a lower level of significance in the hexamers derived from experiments interrogating the other

protein, our results suggest that TAF15 and FUS share some affinity with each others motifs. It is noteworthy that the affinities of TAF15 and FUS to k-mers containing GGUA and GGUG, although significantly different from background, is relatively weak compared to previously studied RBPs such as RBFOX2, MBNL1, and CELF126. We conclude that TAF15 interacts with a previously undiscovered GGUA core motif within significantly enriched clusters in vivo. Importantly, our results demonstrate that the interactions of FUS and TAF15 with their RNA binding sites can occur independently of co-factor associations.

2.3.3 TAF15 interacts with many FUS RNA targets

Like FUS and TDP-43, TAF15 clusters were predominantly found within introns (Supplementary Fig. 3a), consistent with previously published results in HEK293 cells[HLR$^+$11], mouse neurons, and human brain tissue[IMA$^+$13]. As intronic regions account for a substantial proportion of nucleotides in transcribed RNA, this distribution was expected and similar to other predominantly nuclear-localized RBPs such as RBFOX1 and NOVA1 (Supplementary Fig. 3a). Unlike TDP-43, we found that TAF15 and FUS binding was significantly enriched in the 3UTR, akin to RBFOX1 and NOVA1 that also have proposed 3 end formation roles[WSL$^+$08, LMF$^+$08] (Fig. 3a). To illustrate this, the 3UTR of the neurobeachin (Nbea) transcript, encoding a protein involved in synaptic function and autism[ORJ$^+$12], is enriched for TAF15 and FUS binding (Fig. 3b). FUS had a similar binding profile as TAF15, whereas TDP-43 bound an intronic region upstream of the penultimate exon, but with no cluster in the 3UTR (Fig. 3b).

We found that targets of RBFOX1 and NOVA1 do not overlap with TAF15 target genes (Supplementary Fig. 3b). In contrast, the majority of FUS (98%) and TDP-43 (86%) target RNAs were also TAF15 targets (Fig. 3c). For genes that were targets of both TAF15 and FUS, 38% had at least one binding site that overlapped between TAF15 and FUS (Supplementary Fig. 3c). Our results indicate that TAF15 and FUS bind to the same genes with close proximity, consistent with our findings that the GUGG motif preferred by FUS was enriched in TAF15 CLIP clusters and the

GGUA motif preferred by TAF15 was also enriched in FUS CLIP clusters (Supplementary Figs. 2c and 2d). TAF15 also exhibited a saw-tooth like pattern of deposition within genes containing long introns, such as the glutamate receptor delta-1 subunit precursor gene (Grid1), similar to FUS, but dissimilar to TDP-43[LTPH$^+$12] (Fig. 3d). We conclude that TAF15 and FUS binding are enriched in the 3UTRs of target genes and both harbor the same saw-tooth like profiles in long introns.

2.3.4 Distinct roles of TAF15, FUS, and TDP-43 on gene expression

To identify TAF15-regulated RNAs, single-stranded antisense oligonucleotides (ASOs) complementary to TAF15 RNA or non-targeting control ASOs (Control) were delivered into the adult mouse striatum. TAF15 mRNA and protein were depleted by at least 90% in mice treated with TAF15-targeting ASOs (Fig. 3e). RNA extracted from striata of three mice was subjected to strand-specific RNA-seq library generation and sequencing. On average, 24.8 million reads were obtained for each library, of which 86% mapped to the mouse genome (mm9). We identified 194 and 91 genes (Supplementary Data 1) that were significantly (p¡0.05) downregulated (Fig. 3f) and upregulated (Fig. 3g), respectively, when TAF15 protein was depleted. To examine overlapping and unique effects of TAF15, FUS, and TDP-43 on RNA expression, we re-analyzed RNA-seq datasets in which FUS and TDP-43 were depleted from the mouse striatum in the same manner as TAF15[LTPH$^+$12, PLTH$^+$11]. Although Fus and Tdp-43 expression remained unchanged upon TAF15 depletion, Taf15 mRNA level appears to be slightly increased upon FUS depletion (Supplementary Fig. 3d). Similar to FUS and TDP-4321, we found that genes downregulated by loss of TAF15 exhibited exceptionally long introns (Supplementary Fig. 3e). Despite this similar trend in regulation, there was a poor overlap between the differentially regulated genes such that by our conservative re-analysis, only 8 genes (including Park2, Nrxn1, and Kcnip4) were commonly downregulated (Fig. 3f) and no genes were commonly upregulated (Fig. 3g). To distinguish between direct and indirect effects of RBP binding on gene expression, we measured

the fraction of affected genes that were directly bound as determined by CLIP-seq. A significantly higher proportion of genes downregulated upon TAF15 or TDP-43 loss were direct targets of that RBP (Fig. 3h). Closer examination revealed that this association remains significant for the subset of downregulated genes that exhibited TAF15 (and to some extent TDP-43) binding in the 3UTR (Supplementary Fig. 3f). In support of this result, genes that were downregulated upon TAF15 loss were more likely to contain the TAF15 GGUAA motif in their 3UTRs or introns (Supplementary Fig. 3g). Genes that were upregulated or were unaffected upon loss of FUS, TAF15, or TDP-43 were generally not binding targets (Figs. 3i and 3j). Thus, we conclude that although FUS, TAF15, and TDP-43 bind many of the same targets, only a small fraction of genes are similarly affected by loss of each of the three RBPs.

2.3.5 TAF15 has a marginal role in alternative splicing

Using splicing-sensitive microarrays, we detected 182 alternative splicing (AS) events that were altered upon TAF15 depletion in mouse striatum (Fig. 4a). Although TAF15, TDP-43, and FUS proteins were reduced to similar levels, we observed fewer TAF15-dependent splicing events (n=187) compared with the number of splicing events altered by loss of FUS (n=327) or TDP-43 (n=690) (Supplementary Data 2). There was little overlap in AS events altered by loss of TAF15, FUS, or TDP-43 (Fig. 4b and Supplementary Fig. 4a). This suggests that, despite their high similarity in domain architecture and documented interactions with splicing factors[JPV$^+$09, YCY$^+$12], FUS and TAF15 have distinct impacts on AS.

An AS event altered by TAF15 is exon 5 of the glycerophosphocholine phosphodiesterase 1 gene (GPCPD1; Fig. 4c), which was included upon TAF15 knockdown and harbors binding sites for TAF15 and FUS downstream of the 5 flanking exon (indicated by an arrow). Another example, exon 24 in the calcium-activated potassium channel subunit alpha-1 gene (Kcnma1; Fig. 4d), was also included upon TAF15 loss and contained nearby TAF15 and FUS binding sites (indicated by an arrow). Although TDP-43 binding sites were present near these exons, they were

distinct from TAF15 and FUS binding sites (Figs. 4c and 4d). A previous study reported that knockdown of TAF15 promoted the exclusion of exon 19 of the N-Methyl-D-Aspartate Receptor Subunit NR1 (Grin1) gene in mouse neurons15. TAF15 and FUS (but not TDP-43) binding sites were present proximal to this exon, but depletion of TAF15 did not cause a statistically significant change (p=0.106) in the exclusion of exon 19 of Grin1 in the mouse striatum (Supplementary Fig. 4b). This does not appear to be due to differences in tissue specificity as depletion of TAF15 in the mouse brain or spinal cord (Supplementary Fig. 4c) also had no significant effect on Grin1 exon 19 splicing (Supplementary Fig. 4d). We conclude that in contrast to our and others previous findings with FUS and TDP-43[REB$^+$12, LTPH$^+$12, PLTH$^+$11, TCR$^+$11] and a study regarding TAF1515, TAF15 alters the splicing of a relatively small subset of genes, of which the majority (70%) are distinct from those regulated by either FUS or TDP-43.

2.3.6 TAF15 and FUS affect mRNA stability in neural progenitors

To evaluate the role of TAF15 in early neuronal development, we used human neural progenitor cells (NPCs) differentiated from iPSCs in which TAF15 protein levels and, for comparison, FUS protein levels, were individually depleted by lentiviral shRNAs (Figs. 5a and 5b). As TAF15 has a relatively minor role in AS, we investigated a potential role for TAF15 in RNA stability. NPCs were treated with the transcriptional inhibitor Actinomycin D for varying times after which total RNA was collected and prepared for RNA-seq libraries (Fig. 5a). Half-life measurements were determined from a regression-based analyses of gene expression assuming first order decay kinetics. For each shRNA treatment, the median value for the coefficient of determination (R2) describing the log-linear fit across the time course for each gene, across all genes, was 0.54; this value was significantly higher (p 0, by Kolgomorov-Smirnov two-tailed test) than the value (0.12) obtained by randomly shuffling the expression values for each time point within each gene (Fig. 5c, Supplementary Fig. 5a and 5b). To minimize false positives, we evaluated genes for which the R2 value was greater than 0.6. We identified 299 and 330

genes that were highly stabilized (increased half-life), as well as 132 and 44 genes that were highly destabilized (decreased half-life) upon loss of TAF15 and FUS, respectively (Fig. 5d, Supplementary Data 3).

We arbitrarily selected mRNAs that were most stabilized or destabilized by loss of TAF15 (marked by asterisks in Fig. 5e) and performed RNA immunoprecipitation followed by quantitative RT-PCR to determine whether these mRNAs were directly bound by TAF15 and FUS. TAF15 bound to most mRNAs (4 of 5) that were stabilized upon loss of TAF15 (URB1, SNX9, CLN8, SMURF2; Fig. 5f), of which URB1, CLN8 and SMURF2 also exhibited FUS interactions. Additionally, TAF15 bound to most mRNAs (4 of 5) that were destabilized upon loss of TAF15 (ATXN7L3B, PRKRIR, RAPGEF1, CGGBP1; Fig. 5g). Notably, FUS bound to all these mRNAs (including TCERG1), but FUS depletion did not appear to have an effect on mRNA stability of these transcripts (Fig. 5e, Supplementary Data 4). ANAX2 and TIAL1, whose mRNAs were unaltered by TAF15 loss, were also examined for TAF15 and FUS binding (Supplementary Fig. 5c). A gene ontology (GO) analysis of genes affected at the mRNA stability level upon TAF15 knockdown revealed statistical enrichment for genes implicated in DNA-dependent transcription control (p¡10-26) (Supplementary Data 5). An example of TAF15 mRNA turnover target involved in transcriptional control and also neurological diseases is the CGG binding protein 1 (CGGBP1), which binds to CGG repeats in the promoter of the fragile X mental retardation 1 (FMR1) gene resulting in reduced expression[MHDN$^+$00]. We conclude that TAF15 and FUS control mRNA turnover in NPCs of distinct mRNA substrates.

2.3.7 TAF15 and FUS affect different genes in human motor neurons

To discover the molecular events modulated by loss of TAF15, FUS and TDP-43 in an ALS-relevant cell-type, we generated motor neurons (MNs) from wild-type human iPSCs using a directed differentiation protocol[CFP$^+$09]. Briefly, a combination of SMAD signaling inhibitors, Noggin, and the ALK5 inhibitor SB431542 was used to yield a population of cells enriched

for HB9, ISLET1, and TUJ1 (neuron-specific class III)-positive MNs with a minor fraction of OLIG2-positive oligodendrocytes (Figs. 6a and 6b). We subjected the MNs to lentivirus-packaged shRNAs targeting TAF15, FUS, or TDP-43. As our in vivo findings indicated that TAF15 and FUS binds to similar RNA substrates, we also simultaneously depleted FUS and TAF15. Mature RNA and protein levels (Figs. 6c, 6d and Supplementary Fig. 6a) of the targeted RBPs were significantly reduced and TAF15 and FUS protein levels did not exhibit reproducible changes (either up or down) in FUS and TAF15 depletions, respectively (Supplementary Fig. 6a). Reduction of TAF15, FUS, or TDP-43 alone or in combination (TAF15 and FUS) in iPSC-derived MNs did not cause noticeable changes in cell morphology or death. We generated RNA-seq data from these cells, obtaining an average of 32.4 million uniquely mapped reads.

Similar to our in vivo depletion studies, we observed a minor overlap in the genes down-regulated (61 genes) or upregulated (6 genes) upon loss of all three RBPs (Fig. 6e, Supplementary Data 6). In contrast to our findings in the adult mouse striatum, introns within downregulated genes affected by loss of TAF15 and FUS in MNs were not significantly longer than upregulated or unaffected genes (data not shown). Expectedly, 76% and 85% of the genes in the FUS-only and TAF15-only knockdown experiments were also downregulated in the double knockdown. However, we found that a subset of genes (n=144) were downregulated only upon combined loss of TAF15 and FUS in human MNs (Fig. 6f), indicating a potential redundancy between TAF15 and FUS in controlling gene expression. These genes that were downregulated upon combined TAF15 and FUS loss were enriched for GO terms reflecting extracellular cellular matrix composition, cell proliferation, wound healing, and cytokine activity.

2.3.8 Genes affected by RBP loss are similar to ALS-linked FUS mutant

To investigate if the molecular changes observed upon loss of FUS, TAF15, or both proteins were relevant to ALS pathogenesis, we obtained fibroblasts from two ALS patients with the causative R521G mutation in FUS. The fibroblasts were reprogrammed into iPSCs and

subjected to cellular, molecular, and genetic characterization to confirm that they are pluripotent (Supplementary Fig. 6b), exhibit a normal karyotype, and harbor the presence or absence of the mutation at nucleotide position 1561 (Supplementary Fig. 6c). Three individual clones from two FUS R521G patient-derived iPSC lines (two clones were from one line) and two control iPSCs (from healthy, age-matched non-mutant individuals) were directly differentiated to MNs. RNA isolated from these cells were subjected to RNA-seq library preparation and sequencing to obtain an average of 20 million reads, of which 90% mapped uniquely to the human genome (hg19). To ensure that the differentiation process yielded MNs at similar stages of differentiation and similar subtypes of cells, we compared expression of a panel consisting of genes representing housekeeping, astrocyte, oligodendrocytes, neural precursor, and neuronal subtypes. The similarities in expression profiles among the MN cell lines confirmed that differentiation of the iPSC lines were consistent and hence enabled downstream comparative analysis (Supplementary Fig. 6d). We identified 901 downregulated and 805 upregulated (Supplementary Data 7) genes that were differentially affected in the FUS R521G MNs compared to wild-type control MNs. Interestingly, although the majority of mutant-dependent gene expression changes were unique, there existed statistically significant overlaps in the genes downregulated in the FUS R521G MNs (relative to control) with genes downregulated upon loss of FUS (p¡10-9) or TAF15 (p¡10-3) (Fig. 6g). Importantly, this overlap increased in number when we compared the genes affected by simultaneous depletion of both FUS and TAF15 (p¡10-22) (Fig. 6g). In contrast, we observed no significant overlap in genes upregulated by any condition (Fig. 6g). Overall, these findings are consistent with our observations that FUS and TAF15 are redundant in their effects on molecular targets and implies a partial loss of molecular function by the FUS R521G mutation.

2.3.9 Downregulated genes correlate with a sALS RNA signature

To obtain insights into whether the genes affected by loss of ALS-associated RBPs resemble disease-specific RNA signatures, we turned to a RNA-seq dataset generated from

laser-capture microdissected (LCM) spinal cord samples from sALS patients who had bulbar or arm onset of disease that was caudally progressing and thus had abundant residual MNs in the lumbar region at the time of death[BHV^{+}16]. The RNA-seq dataset consisted of samples from 13 sALS and 9 control patients. 3,876 genes were significantly differentially regulated, of which 71% and 29% were upregulated and downregulated, respectively, in the sALS patient compared to normal samples (Supplementary Data 8). The differentially expressed genes were effectively able to separate the disease patients from the control patients (Supplementary Fig. 6e). Next, we tested the hypothesis that ALS RBP-mediated RNA changes resemble the RNA signature that distinguishes sALS and normal samples. For all the comparisons performed, we observed a significant overlap (p¡0.05, hypergeometric test) in genes that were upregulated in sALS samples and downregulated in FUS, TAF15, TDP-43, or FUS and TAF15 double knockdowns (Supplementary Fig. 6f). We also observed a significant inverse correlation of significantly changing genes between sALS samples and FUS, TDP-43, or FUS and TAF15 double knockdowns (Fig. 6h, R2 between -0.14 and -0.32, p¡0.05), but not between sALS samples and the FUS R521G mutant MNs (Supplementary Fig. 6g). Despite the divergent sets of regulated genes whose mRNA levels are dependent on ALS-associated RBPs, we found that 2,747 genes were upregulated in sALS patient samples. Unlike in vitro differentiated MNs, the sALS patient samples represent more mature MNs at a late stage of disease progression. Our findings indicate that in late stage sALS patient samples with TDP-43 pathology37, a subset of genes that are separable from those found in ALS iPSC-derived FUS R521G MNs, are abnormally higher compared to control patients. Among the commonly differentially regulated genes between knockdown and sALS samples GO terms for extracellular space and matrix organization were statistically enriched (p¡0.01).

2.4 DISCUSSION

Genetic and clinical evidence strongly supports causative roles for FUS, TDP-43 and TAF15 in ALS. Here, we identify common and unique pathways normally controlled by these proteins utilizing diverse in vitro and in vivo neuronal systems (Supplementary Fig. 7). In the adult mouse brain, we identified TAF15 binding sites within 4,900 RNA substrates, and a GGUAAGU TAF15 binding motif not reported in previous studies[IMA$^+$13, HLR$^+$11]. We used RNA Bind-n-Seq technology to confirm a GGUA motif that was enriched within in vivo TAF15 binding sites. Together, we conclude that TAF15 and FUS can interact with their RNA motifs within in vivo RNA substrates without requiring complex co-factor associations. Overall, the RNA binding pattern of TAF15 resembled that of FUS, but was distinct from TDP-43, even when all three RBPs targeted the same genes. TAF15 and FUS exhibited sawtooth-like binding patterns on long introns, a pattern reminiscent of co-transcriptional splicing[AZH$^+$11]. Genes downregulated upon loss of either TAF15 or FUS contained exceptionally longer introns. Additionally, TAF15 and FUS binding sites were also over-represented within 3UTRs, possibly reflecting 3 end processing functions such as RNA turnover, transport, and translation. Upregulation of genes upon FUS and TAF15 loss is likely a secondary effect as these genes are generally not targets. Lastly, unlike TDP-43 and FUS, loss of TAF15 appeared to have a minor impact on alternative splicing in the adult mouse brain.

In models of early human neuronal development, we identified that TAF15 and FUS affected the mRNA turnover of distinct subsets of RNA targets in human neuronal progenitor cells. Furthermore, loss of FUS, TAF15 or TDP-43 in human MNs derived from the same cells resulted in distinct changes in gene expression for each RBP. Additionally, simultaneous depletion of FUS and TAF15 resulted in the downregulation of hundreds of additional genes. FUS and TAF15 have been shown to interact with each other[TGE$^+$13, SLQ$^+$15] (Supplementary Fig. 1b) as well as other common proteins such as RNA Pol II[SEP$^+$12, KKX$^+$13], spliceosome

machinery[SLQ$^+$15, JPV$^+$09, YCY$^+$12], and transcription factors[BLH$^+$96]. One possibility is that if FUS is unable to recruit regulatory factors to an RNA target, this function may be compensated for by TAF15.

To gain insight into disease, we compared the results of our loss-of-function studies to two models of ALS. The first model is MNs from ALS patients carrying the pathogenic FUS R521G mutation. Expression of FUS R521G from the mouse MAPT locus has been recently reported to cause neuronal toxicity in neurons of mice[SLL$^+$16]. Previously, FUS R521G was also associated with a partial loss-of-function in RNA regulation in mouse spinal cords[STK$^+$14]. We did observe a small yet significant overlap in genes downregulated upon loss of TAF15, FUS, or both proteins and genes altered by FUS R521G. This overlapping set of genes may reflect the partial loss-of-function properties of FUS R521G43. As mRNAs downregulated upon loss of these RBPs in mouse brain are often direct binding targets of those RBPs, we speculate that the FUS R521G mutation, which causes cytoplasmic FUS mislocalization, resembles a partial loss-of-function of the RBPs in a model of early development. Nevertheless, the majority of expression changes caused by FUS R521G were mutant-specific such that they did not overlap with genes altered by loss of TAF15, FUS, or both proteins. One interpretation is that these FUS R521G-specific gene changes may contribute the pathological, gain-of-function activities of mutant FUS that was observed to cause motor neuron dysfunction in mice[SLL$^+$16].

To model late stage ALS disease we utilized RNA-seq data obtained from spinal cord samples collected post-mortem by laser-capture microdissection from sALS patients. These samples harbored ubiquinated TDP-43 cytoplasmic inclusions and were from patients with no mutations in known ALS causative genes, including FUS, TDP-43 or TAF15. Intriguingly, our comparisons of RNA signatures revealed an inverse correlation in a separate set of genes that were upregulated in the sALS samples but were downregulated upon loss of ALS-associated RBPs in in vitro derived MNs. This indicates that these genes, whose levels are normally dependent and maintained by FUS, TAF15, and TDP-43, are aberrantly higher in late-stage ALS. A possible

mechanism for gene upregulation is the breakdown of negative feedback loops as is observed for the effect of TDP-43 on its own expression[BB11, PLTH$^+$12]. We did not, however, observe a difference in TDP-43 mRNA levels between sALS and control neurons. Another plausible scenario is that in late stages of the disease, cytoplasmic inclusions of TDP-43 lead to stabilization of trapped, cytoplasmic RNA targets. Future studies to identify the mislocalized RNA targets in cytoplasmic bodies that are protected from degradation, such as stress granules, may yield further insight into disease-relevant targets at late stages in the disease.

In summary, our study delineates convergent and divergent RNA processing functions of ALS-associated FUS, TAF15, and TDP-43 in normal and disease settings. Our comprehensive results shed light on multiple and distinct pathways by which these RBPs regulate gene expression in diverse neuronal systems and provide a framework for how they relate to ALS and other neurodegenerative diseases.

2.5 METHODS

2.5.1 Injections of ASO in mice

Sterotaxtic injections of ASO complementary to TAF15 were performed in eight-week old female C57Bl/b mice to deplete TAF15. ASOs were delivered specifically to the striatum or brain/spinal cord by intrastriatal (12.5 g) or intracerebroventricular injection (300 g), respectively, as described previously[LTPH$^+$12, RCN$^+$14]. Female mice were regularly monitored for 14 days until sacrificed and the tissues were harvested and frozen in TRIzol (Invitrogen). Control mice received a control ASO without any known target in the mouse genome under the same conditions. The ASOs were phosphorothioate gapmers with sequences as follows (capitalized nucleotides containing 2-O-(2-methoxy)ethyl modifications): GGTCTcctccatagcTGCCT (TAF15; brain and striatum), TGGCAatattttacaACGCA (TAF15; spinal cord), CTCAGTAACATTGACACCAC (Control). All procedures were performed using a protocol approved by the Institutional Animal

Care and Use Committee of Ionis Pharmaceuticals and the University of California at San Diego.

2.5.2 Generation of neural precursor cells and motor neurons

Human induced pluripotent stem cells (iPSCs) derived from dermal fibroblasts cells of a healthy individual (RRN08) were induced into neural precursor cells using a pan-neuronal protocol as previously described21. Briefly, stem cells were grown on Matrigel-coated plates (BD) in mTeSR1 growth media (Stem Cell Technologies). Stem cell colonies were grown on ultra low-attachment plates in DMEM/F12 + GlutaMAX supplemented with N2 and FGF-2 (20 ng/ml). After one week, neural rosettes were manually picked, replated, and maintained in DMEM/F12 + GlutaMAX supplemented with N2, B27, and FGF-2 (20 ng/ml).

2.5.3 Generation of human motor neurons

Human motor neurons used in the shRNA knockdown experiments were differentiated from iPSCs (CVB) using a protocol modified from Chambers et al.36. Briefly, human iPSCs were maintained in hEB Media (Knockout D-MEM + 10% Knockout Serum Replacement (Life Technologies) + 10% Plasmanate (Biocare) + GlutaMAX + NEAA (Life Technologies) and supplemented with 10 M SB431542 and 1 M Dorsomorphin dihydrochloride (Tocris) on feeder-free dishes. Cells were maintained in SB431542 and Dorsomorphin until day 18 of differentiation. On days 4, 5, and 6 of differentiation, hEB media was mixed with N2 Base media (D-MEM/F12 + GlutaMAX, 1% N2 Supplement + 4.5 mM D-Glucose, 0.05 mM Ascorbic Acid (Sigma)) at a ratio of 70:30, 50:50, and 50:50, respectively. On days 7 and 8 of differentiation, cells were maintained in 50:50 combination of hEB media and maturation media (D-MEM/F12 + GlutaMAX, 2% N2 Supplement, 4% B27 Serum-Free Supplement (Invitrogen), 9.0 mM D-Glucose, 0.1 mM Ascorbic Acid (Sigma)) supplemented with 2 ng/mL each of ciliary neurotrophic factor (CNTF), brain-derived neurotrophic factor (BDNF), and glial cell-derived neurotrophic factor (GDNF)

(Peprotech). From day 7 to day 22 of differentiation, cells were treated with 200 nM Smoothened Agonist (SAG; EMD Biosciences) and 1.5 M Retinoic Acid (RA; Sigma). On day 18, cells were dissociated using Accutase, and transferred to dishes coated with Poly-D-Lysine (Sigma) + Laminin (Life Technologies) and maintained in maturation media supplemented with RA and SAG. On day 22, cells were maintained in maturation media containing 2 M DAPT (Tocris). On day 26, cells were maintained in maturation media only. Throughout the differentiation protocol media was changed daily. The identity and purity of motor neurons were analyzed by immunofluorescence for markers of stem cells, motor neurons, astrocytes, and glial cells.

2.5.4 Generation of motor neurons from fibroblast-derived iPSCs

Adult human primary fibroblasts were obtained by Franca Cambia, Edward Kasarskis, and Haining Zhu (University of Kentucky). Informed consent was obtained from all subjects before sample collection. The use of patient fibroblasts for research was approved by the University of Kentucky Institutional Review Board (IRB 05-0265). Briefly, adult human primary fibroblasts were cultured at 37C and 5% CO2 in DMEM supplemented with 10% FBS, NEAA, and L-glutamine. To generate iPS cells, control and ALS patient fibroblasts were transduced with CytoTune iPS Sendai Reprogramming Kit, as described in manufacturers protocol (Invitrogen). Colonies were manually passaged onto Matrigel-coated plates and grown in mTeSR1 growth media. After several passages, colonies were expanded using Accutase (Innovative Cell Technologies) and grown as a monolayer prior to differentiation. Motor neuron differentiation was performed as described above with the following modifications: CHIR99021 (Tocris) was added at 4 m until day 7 and the cells were either fixed for immunostaining or harvested for RNA in TRIzol (Life Technologies) 35 days post neural induction. Three ALS patient lines GY6.2, GY7.3, and GY7.6 are referred to as FUS R521G Line 1, Line 2 and Line 3 respectively, and two wild-type sibling control lines KIN1ALS17.3 and KIN1ALS17.4 are referred to as WT sibling control Line 1 and Line 2, respectively.

2.5.5 Lentiviral infections and transfections

Lentiviral shRNA constructs (Open Biosystems) complementary to human TAF-15 (TRCN0000020140, TRCN0000020141 or TRCN0000020143,), human FUS/TLS (TRCN0000010450, TRCN0000039824, or TRCN0000039825), and human TDP-43 (TRCN0000016038) in the pLKO.1 vector system were used to produce lentivirus as previously described[YCL$^+$09]. Virus produced from a pLKO.1 construct containing a control sequence was used as the control. At 60-70% confluency, NPCs were infected with virus (MOI=3) for 24 hours, followed by a complete media change and further incubation for 72 hours until cells were either collected and frozen in TRIzol (Invitrogen) or pelleted and frozen in liquid nitrogen for RNA and protein analysis, respectively. For lentiviral infection of motor neurons, media containing virus (MOI=5) was added to cells on day 28 of the motor neuron differentiation protocol. After 24 hours, a complete media change was performed and cells incubated for an additional 48 hours. A second round of infection, similar to the first, began on day 31. On day 34 of MN differentiation, corresponding to a six-day exposure period to shRNA expression, cells were either collected and frozen in TRIzol (Invitrogen) or pelleted and frozen in liquid nitrogen for RNA and protein analysis, respectively. For transfection of HEK293T cells, cells were plated in DMEM high glucose media (Life Technologies) supplemented with 10% FBS. Cells were transfected with plasmid expressing human FUS-Myc cloned into pcDNA5 or TAF15-V5 cloned into pEF5-DEST using FuGENE 6 (Promega) according to the manufacturers protocol for 24 hours and then harvested for protein analysis.

2.5.6 CLIP-seq library preparation and sequencing

Brains from 8-week-old female C57Bl/6 mice were rapidly dissociated by forcing the tissue through a cell strainer with a pore size of 100 m (BD Falcon) before ultraviolet crosslinking. CLIP-seq libraries were constructed as previously described47 using 10 g of a polyclonal antibody

against TAF15 (300A 308, Bethyl Laboratories). Libraries were subjected to sequencing on a HiSeq2000 platform for 50 cycles. For each CLIP-seq library, the brain of one mouse was used.

2.5.7 Computational analysis of CLIP-seq experiments

CLIP-seq alignment and peak calling were performed as previously described21. Briefly, reads with the sequencing adapter or homopolymeric runs were trimmed and then mapped to the repeat-masked mouse genome (mm9) using Bowtie (version 0.12.2) with parameters q p 4 e 100 a m 10 beststrata. Reads that were flagged as PCR duplicates were removed. Significant clusters of reads were identified using a Poisson distribution with two different frequencies to determine a p-value. First, a transcriptome-wide frequency was calculated by dividing the total length of all pre-mRNAs by the total number of CLIP reads mapping to the whole pre-mRNA transcriptome. Second, a gene-specific frequency was calculated by dividing the size of the gene-specific pre-mRNA by the total number of CLIP reads mapping to that gene-specific pre-mRNA. A significant cluster was annotated if it had sufficient reads to exceed a Bonferroni-corrected p¡10e-4 using both frequencies against the Poisson distribution

2.5.8 De novo motif analysis

Motif analysis was performed as previously described[LGM$^+$13]. Briefly, HOMER[HBS$^+$10] was used to call de novo motifs using the command findMotifs.pl ¡foreground¿ fasta ¡outloc¿ -nofacts p 4 rna S 20 len 5,6,7,8,9 noconvert nogo fasta ¡background¿. Where foreground was a fasta sequences taken from all called clusters, or all called clusters in a specific transcriptome region and background was randomly located clusters within the same genic regions as predicted TAF15 clusters.

2.5.9 Peak Annotations

Transcriptome regions and gene classes were defined using annotations found in GENCODE version 17[HFG+12]. Depending on the analysis, clusters were either associated by the GENCODE annotated 5UTR, 3UTR, exon, or intronic regions. If a cluster overlapped multiple regions or a single part of a transcript was annotated as multiple regions, clusters were iteratively assigned first as exon, then 3UTR, 5UTR, and finally as proximal or distal introns (as defined as 500 bp or greater from an exon-intron boundary). Overlapping peaks were calculated using bedtools[QH10, DPQ11].

2.5.10 Enrichment of peaks relative to region size

To compute the fold enrichment of peaks in a given region, the fraction of peaks in that region was calculated as described above. The fractional region size was derived by dividing the total number of base pairs in that region relative to the total number of base pairs in all regions. Fold enrichment was computed using the equation log2 (FCLIP / Fregion).

2.5.11 Distance of peaks from motifs

Distance from peaks was computed by using the annotatePeaks function in HOMER[HBS+10] with the arguments annotatePeaks.pl ¡peaks¿ mm9 -m ¡motif¿, -hist 10-size 1000 noann. Identification of peaks and motifs were determined as described above.

2.5.12 RNA Bind-n-Seq (RBNS)

RBNS was performed as previously described26. Briefly, truncated reading frames of FUS (amino acids 204-415) and TAF15 (amino acids 235-418), which contain all RNA binding domains, were cloned downstream of a tandem GST-SBP tag into a modified pGex6p-1 vector (GE). Truncated proteins were recombinantly expressed and purified via the GST tag, and used

for RBNS, which was performed at 5 concentrations (0 nM, 5 nM, 20 nM, 80 nM, and 320 nM) with a pool of randomized 20mer RNAs, flanked by short primers. Preparation of the randomized RNA pool and all reaction conditions were identical to previous descriptions[LRJ$^+$14]. Further computational analysis details can be found in Supplementary Methods.

2.5.13 RNA-seq library preparation, sequencing, and analysis

Total RNA was extracted from mouse tissues and human cells using TRIzol (Invitrogen) according to the manufacturers instructions. 0.5-3 g of total RNA was DNase treated and subjected to poly(A) selection or Ribo-Zero treatment followed by library preparation using TruSeq Stranded mRNA and Total RNA Sample Preparation Kit (Illumina). Barcoded libraries were pooled at equal concentrations and sequenced on the HiSeq 2000 or HiSeq2500 platform for 50 cycles. RNA-seq reads were trimmed of polyA tails, adapters, and low quality ends using Cutadapt[Mar11] with parameters –match-read-wildcards –times 2 -e 0 -O 5 –quality-cutoff' 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTCGGGTGCCAAGG -b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase (version 18.05) using Bowtie (version 1.0.0) with parameters -S -q -p 16 -e 100 -l 20[LTPS09]. Reads that did not map to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR (version 2.3.0e)[DDS$^+$13] with parameters –outSAMunmapped Within outFilterMultimap-Nmax 10 outFilterMultimapScoreRange 1. Counts were calculated with featureCounts[LSS14] and RPKMs were computed. Differential expression was calculated using DESeq2[LHA14], individually pairing each knockdown experiments with their respective controls.

2.5.14 Test of overlapping significance between gene sets

Genes from each differential expression experiment were considered significant if —log2 fold change— ¡ log2(1.5) and the adjusted p¡0.05. Significant genes between two sets were overlapped and the total set of genes was defined as genes that were expressed (RPKM¿1) in the corresponding control experiment. A hypergeometic test was performed to determine if the overlap of two gene sets was statistically significant. Regression analysis was performed using the scipy linear regression function on genes that were significantly differentially expressed in both samples.

2.5.15 RT-PCR of splicing events

To validate alternative splicing events, RT-PCR (2427 amplification cycles) was carried out using poly-Aselected and reverse transcribed (Superscript III, Invitrogen) cDNA from mice (n=3) treated with either a control ASO or ASO targeting the indicated RBP. Isoform products were visualized using the Agilent 2200 TapeStation System (Agilent Technologies) or on an agarose gel and quantified using ImageJ to calculate ratios between inclusion and exclusion products. Statistical significance in differences between control and ASO samples was calculated by Students t test. Primer sequences are listed in Supplementary Data 9.

2.5.16 Quantitative RT-PCR

qRT-PCR was performed using Power SYBR Green Master Mix (Life Technologies) using poly-Aselected and reverse transcribed (Superscript III, Invitrogen) cDNA on an iQ5 real-time PCR detection system (Bio-Rad). For each biological replicate, qRT-PCR was carried out in technical triplicates. GAPDH and Actin were used as reference genes for human and mouse targets, respectively. Analysis was performed using the iQ5 optical system software (Bio-Rad; version 2.1). Expression values were normalized to the reference gene and expression values were

expressed as a fold-change relative to control samples. Inter-group differences were assessed by two-tailed Student's t test. Primer sequences were designed using Primer3 software[UCK+12] or obtained from PrimerBank[WSWS12]. Primer sequences are listed in Supplementary Data 9.

2.5.17 RNA immunoprecipitation qPCR (RIP-qPCR)

NPCs were resuspended in lysis buffer (50 mM Tris pH 7.4,100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% Sodium Deoxycholate) supplemented with 1x Protease Inhibitor cocktail (Roche) and 80U of RNAse Inhibitor (Roche). Clarified lysates were pre-cleared with Protein G agarose beads (Life Technologies). Aliquots of the supernatant (equivalent to 5% of supernatant) were saved as input protein and RNA. The remainder of the supernatant was incubated with 10 g of antibody at 4C for 4 hours. The protein-RNA-antibody complex was precipitated by incubation with Protein G magnetic beads overnight at 4C. Beads were washed twice with lysis buffer and three times with wash buffer (5 mM Tris pH 7.5, 150 mM NaCl, 0.1% Triton X 100). Ten percent of the bead slurry was reserved for Western blot analysis. The remaining bead slurry was resuspended in TRIzol (Life Technologies) and RNA was extracted as per the manufacturers instructions. Input and immunoprecipitated RNA was converted into cDNA and gene expression was measured by qPCR. RIP-qPCR studies were performed in biological duplicates. Primer sequences are listed in Supplementary Data 9.

2.5.18 Antibodies for Western blot analysis

The primary antibodies used are as follows: FUS/TLS (ProteinTech 1:1,000), FUS/TLS (Santa Cruz Biotechnology, clone 4H11, sc-47711, 1:100), TAF15 (Bethyl Laboratories 300A-308, 1:1,000), TDP-43 (Proteintech, 10782, 1:2,000), and GAPDH (Abcam, AB8245, 1:10,000). Images have been cropped for presentation. Full size images are presented in Supplementary Fig. 8.

2.5.19 Immunofluorescence

Cells were fixed in 4% paraformaldehyde for 20 min, washed 3 times in PBS, and simultaneously blocked and permeabilized with 5% donkey serum and 0.1% Triton-X100 in PBS for 1 hour at room temperature. Cells were then rinsed once in PBS and incubated with primary antibody overnight at 4C. After 5 washes with PBS, secondary antibodies consisting of goat anti-rabbit Alexa Fluor 488 and goat anti-mouse Alexa Fluor 555 (Life Technologies) were added at a dilution of 1:1,000 for 2 hours at room temperature. Following incubation, the cells were rinsed 3 times with PBS, and nuclei were labeled with 1 g/ml DAPI for 10 min. The following primary antibodies were used: HB9 (1:100, DSHB), Islet1 (1:500, Santa Cruz Biotechnology), Oct4 (1:500, Cell Signaling), Olig2 (1:500, Millipore), Sox2 (1:500, Cell Signaling), Tra1-60 (1:1000, Millipore), Tra1-81 (1:1000, Millipore), Tuj1 (1:500, Millipore).

2.5.20 RBNS Computational Analysis

RBNS analysis was performed as previously described[CVP$^+$16]. Briefly, motif enrichment (R) values were calculated for 6mers as the motif frequency in the RBP-selected pool over the frequency in the input RNA library. R values were considered significant if they had a Z-score 2 (mean and standard deviation calculated over all 6mers). Values in Fig. 2 and Supplementary Fig. 2 are for the protein concentration library with the highest overall enrichment (80 nM for both proteins). RBNS datasets have been deposited at the ENCODE DCC under accession IDs ENCSR936LOF for FUS and ENCSR827QYL for TAF15.

Motif logos were generated following an iterative procedure on the most enriched 6mer library precipitated from the GST-SBP tagged protein: the most enriched 6mer was given a weight equal to its enrichment over the input library (=R-1), and all occurrences of that 6mer were masked in both the precipitated and input libraries. All enrichments were recalculated on the masked read sets to obtain the most enriched remaining 6mer and its corresponding weight,

with this process continuing until the R Z-score was less than 2. All 6mers determined from this procedure were aligned to minimize mismatches to the most enriched 6mer, and a new motif was generated if the number of mismatches was greater than 2. The frequencies of each nucleotide in the position weight matrix, as well as the overall percentage of each motif, were determined from the weights of the individual aligned 6mers that went into that motif.

For comparison with CLIP-seq data, RBNS enrichments were determined from the concentration with the largest enrichment. For enrichment in CLIP-seq 6mers, FASTQ sequences were extracted from all clusters, and a matched number of random clusters from the same genomic region (5UTR, exon, 3UTR, proximal introns, and distal introns). EMBOSS compseq was performed on the real and background set, and a delta between real and background k-mers was calculated with the equation: $\delta kmer = f_{CLIP}/N_{CLIP} - f_{background}/N_{background})/\sqrt{((1/N_{CLIP} + 1/N_{background} * g * (1-g)))}$, for $g = (f_{CLIP} + f_{background})/(N_{CLIP} + N_{background})$ where N is the number of times the motif occurs in the set and f is observed frequency of the motif. To plot enrichment, all 6mers with the 4mer of interest were highlighted and a KDE plot was created for all 6mers. The KolmogorovSmirnov two-tailed test determined statistical significance in differences between distributions.

2.5.21 RNA stability analysis

NPCs were infected with virus as described above. 96 hours post infection, cells were treated with Actinomycin D (10 g/mL) for the indicated times. Cells were washed with cold PBS and harvested for RNA extraction using TRIzol (Life Technologies) or protein for Western blot analysis. 1 g of total RNA was subjected to DNase treatment and poly(A) enrichment, and used to prepare RNA-seq libraries as described above. To calculate RNA half-lives, RPKMs from each experiment were calculated and decay rates were generated by fitting RPKMs for each gene to a log-linear regression using the equation $lnN(t) = ln_{N_0} + -\lambda t$, where t is time and N(t) is the RPKM at time t. Half-lives were derived from the decay rate using the equation t1/2 = ln(2)/ .

Genes were included in the analysis if their decay rate was positive (i.e., RPKMs decreased over time) and the linear regression line had a R2 fit greater than 0.6.

2.5.22 Correlation of gene expression to CLIP binding and motifs

Mouse brain CLIP-seq data for FUS and TDP-43 were previously described21,22. The binding location of each peak was assigned using the peak annotation method as described above. For each RBP, mouse brain CLIP-seq data and mouse striatum knockdown RNA-seq data was used to classify genes into the following categories: target and regulated, non-target and regulated, target and not regulated, and non-target and not regulated. A Fishers exact test was performed determine if binding and regulation were significantly correlated. Motif analysis was performed similarly, by determining if a TAF15 GGUAA or FUS GUGG motif was present in the 3UTRs or introns of genes.

2.5.23 Splicing-sensitive microarray analysis

Total RNA from three individual control and TAF15 ASO-treated mice were prepared for hybridization to splicing-sensitive microarrays (Affymetrix). Separation scores (Sep scores) were generated as previously described[HVA^{+}12]. For clustering of splicing events, a splicing event was included in clustering if, for any of the three experiments, TAF15, FUS, or TDP-43 knockdown was significantly (—Sep score—¿0.5, q-value¡0.05), differentially expressed. Hierarchal clustering was performed using Seaborn/SciPy on the Sep scores for each splicing event. Overlap analysis of splicing-sensitive microarray results and TAF15 mouse CLIP-seq data was performed as previously described61.

2.5.24 Gene ontology analysis

Significantly enriched gene ontology (GO) terms were identified using a hypergeometric test that compared the number of genes that were either regulated (RNA-seq data) or bound (CLIP-seq data) in each GO term to genes expressed (background) in each GO term. The background gene set was defined as genes that were expressed (RPKM¿1) in the corresponding control experiment.

2.5.25 Data availability statement

The accession number for the sequencing data deposited in GEO for this paper is GSE77707.

2.6 AUTHOR CONTRIBUTIONS

GWY conceived the study. SCH, JPD, LS, and MA performed splicing microarray analyses. PF, NJL, and CBB performed the RNA Bind-n-Seq experiments. TYL performed the CLIP experiments. AV, FJM, JC, and KK generated iPSC lines and performed neural differentiation. BS assisted with the RNA stability experiments. KRH and GP performed the bioinformatics analyses. FR and SC performed the antisense-oligonucleotide experiments. HZ, JZ, FC, and EK provided the FUS R521G fibroblasts. RB and JR contributed sporadic ALS and control patient RNA-seq data and analyses. KK, GP, and GWY analyzed data and wrote the manuscript.

2.7 COMPETING FINANCAL INTERESTS

F.R. is a paid employee of Ionis Pharmaceuticals.

2.8 ACKNOWLEDGMENTS

The authors would like to thank members of the Yeo lab, especially Stefan Aigner for critical reading of the manuscript. S.C.H. and G.P. were funded by National Science Foundation Graduate Research Fellowships. G.P. was also partially supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number T32GM008666. This work was supported by grants from the National Institutes of Health (HG004659, NS075449, and HG007005 to G.W.Y; NS077284 to H.Z), the California Institute of Regenerative Medicine (RB3-05009 and RB4-06045 to G.W.Y.) and ALS Association (VC8K27 to G.W.Y; 6SE340 to H.Z.). This work was partially supported by NIH grant HG007005 to C.B.B. We would also like to thank Ionis Pharmaceuticals for sharing reagents and unpublished results. G.W.Y. is an Alfred P. Sloan Research Fellow.

Chapter 2, in full, is a reprint of the material as it appears in Nature Communications 2016. Katannya Kapeli, Gabriel A. Pratt, Anthony Q. Vu, Kasey R. Hutt, Fernando J. Martinez, Balaji Sundararaman, Ranjan Batra, Peter Freese, Nicole J. Lambert, Stephanie C. Huelga, Seung J. Chun, Tiffany Y. Liang, Jeremy Chang, John P. Donohue, Lily Shiue, Jiayu Zhang, Haining Zhu, Franca Cambi, Edward Kasarskis, Shawn Hoon, Manuel Ares Jr., Christopher B. Burge, John Ravits, Frank Rigo, Gene W. Yeo Nature Publishing Group, 2016. The dissertation/thesis author was the primary investigator and author of this paper.

Figure 2.

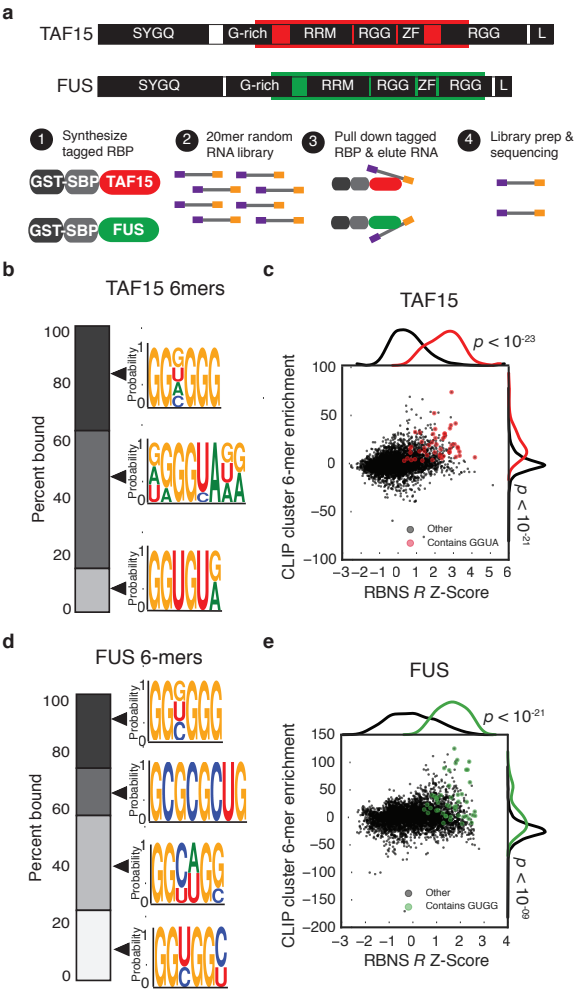

Figure 2.3: **Figure 2. RNA Bind-n-Seq confirms enrichment for GGUA motifs in RNAs that bind TAF15 in vitro** (a) Experimental overview of RNA Bind-n-Seq (RBNS). Truncated versions of TAF15 and FUS (highlighted in red and green, respectively) were tagged and incubated at different concentrations with a diverse pool of RNA oligonucleotides flanked by adapters. The tagged proteins were retrieved with streptavidin-coated beads and bound RNAs were sequenced. Input RNA was sequenced in parallel to quantify the proportions of bound RNA molecules. (b) RNA binding preferences for truncated TAF15 shown as motif logos made from aligning RBNS 6mers weighted by their enrichments. Motif proportions were determined by summing the enrichments of each motifs aligned 6mers. (c) Scatter plot correlating the percent enrichment above background of 6mers in TAF15 mouse brain CLIP-seq vs. TAF15 RBNS R Z-scores. Red dots represent all significant 6mers containing the GUAA motif. Histograms show normalized distributions of 6mers containing (red) or not containing (black) the GUAA motif in CLIP-seq (right) or RBNS (top). p values shown are computed by a Kolmogorov-Smirnov statistic. (d) RNA binding preferences for truncated FUS as in (b). (e) Scatter plot and histogram analyses are as described in (c) using FUS mouse brain CLIP-seq vs. FUS RBNS R Z-scores. Green dots represent all significant 6mers containing the GUGG motif.

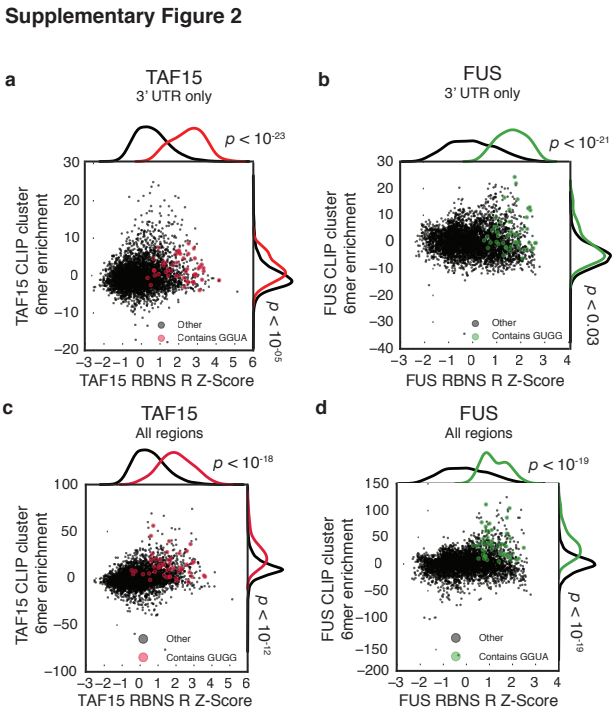

Figure 2.4: **Supplementary Figure 2. RNA Bind-n-Seq confirms enrichment for GGUA motifs in RNAs that bind TAF15 in vitro** (a) Scatter plot correlating the percent enrichment above background of 6mers in TAF15 mouse brain CLIP-seq for peaks in the 3UTR vs. TAF15 RBNS R Z-scores. Red dots represent all significant 6mers containing the GGUA motif. Histograms show normalized distributions of 6mers containing (red) or not containing (black) the GGUA motif in CLIP-seq (right) or RBNS (top). p values shown are computed by a Kolmogorov-Smirnov statistic. (b) Scatter plot as in (a) for 6mers in FUS mouse brain CLIP-seq peaks residing only in the 3UTR vs. FUS RBNS R Z-scores. Green dots represent all significant 6mers containing the GUGG motif. (c) Scatter plot correlating the percent enrichment above background of 6mers in TAF15 mouse brain CLIP-seq vs. TAF15 RBNS R Z-scores. Red dots represent all significant 6mers containing the FUS GUGG motif. Histograms show normalized distributions of 6mers containing (red) or not containing (black) the GUGG motif in CLIP-seq (right) or RBNS (top). p values are shown. (d) Scatter plot as in (c) for 6mers in FUS mouse brain CLIP-seq vs. FUS RBNS R values. Green dots represent all significant 6mers containing the TAF15 GGUA motif.

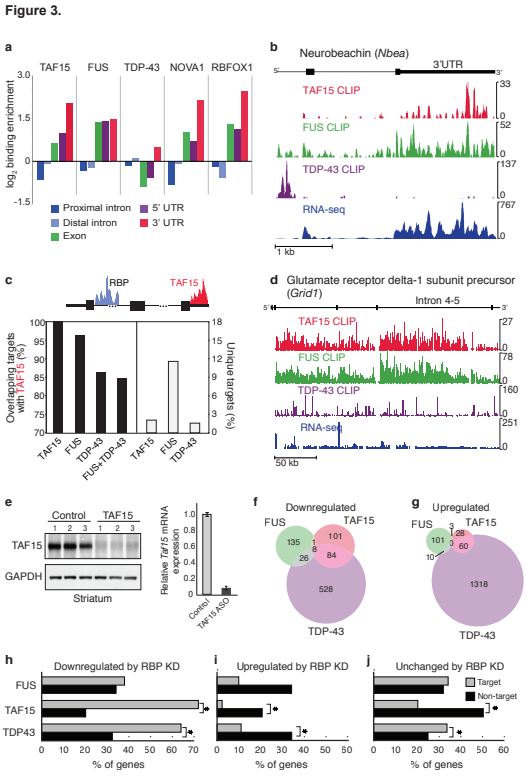

Figure 2.5: **Figure 3. TAF15 and FUS exhibit similar RNA interaction profiles in the mouse brain** (a) Fold change in binding enrichment of TAF15, FUS, TDP-43, NOVA1 and RBFOX1 after normalization to the average length of proximal introns (dark blue), distal introns (light blue), exons (green), 5UTRs (purple) or 3UTRs (red). (b) An example of 3UTR binding by TAF15 (red) and FUS (green) but not TDP-43 (purple) to Neurobeachin (Nbea) mRNA (chr3:55,428,730-55,433,169) in the mouse brain. RNA-seq results showing expression of Nbea is shown in blue. (c) Bar graph showing the percent of gene targets that are common (black bars) or unique (white bars) to TAF15 and FUS, TDP-43, or both FUS and TDP-43. (d) An example of intronic saw-tooth binding by TAF15 (red) and FUS (green) but not TDP-43 (purple) to the Glutamate receptor delta-1 subunit precursor (Grid1) mRNA (chr14:35,634,350-36,071,292) in the mouse brain. RNA-seq results showing expression of Grid1 is shown in blue. (e) Confirmation of reduced TAF15 expression in the mouse striatum by Western blot analysis (left) and qPCR (right). Knockdown was achieved by intrastrial injection of ASOs complementary to TAF15 or a non-murine/human gene (Control). Error bars represent standard deviation. (f-g) Venn diagrams showing overlap of genes downregulated (f) and upregulated (g) upon loss of TAF15, FUS, or TDP-43 in the mouse striatum. (h-j) Percent of genes that are downregulated (h), upregulated (i), or unchanged (j) upon ASO-mediated knockdown of the indicated RBP that has at least one binding site (Target, gray) or no binding sites (Non-target, black) by that RBP. Asterisks denote significant difference between target and non-target genes by Fishers exact test at p¡0.05.

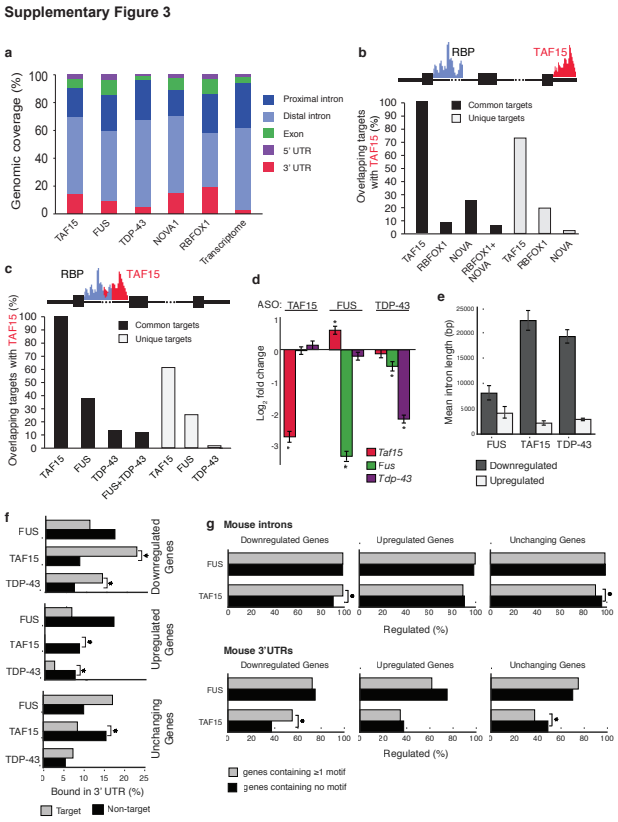

Figure 2.6: **Supplementary Figure 3. TAF15 and FUS exhibit similar RNA interaction in the mouse brain** (a) Percent binding enrichment of TAF15, FUS, TDP-43, NOVA1 and RBFOX1 in proximal intron (dark blue), distal intron (light blue), exon (green), 5UTR (purple) or 3UTR (red) as compared to the observed percentage of nucleotides in the annotated transcriptome. (b) Bar graphs showing the percent of gene targets that are common (black bars) or unique (gray bars) to TAF15 and NOVA1, RBFOX1, or both NOVA1 and RBFOX1. (c) Bar graph showing the percent of CLIP clusters that are common, i.e. within 100 bp of each other (black bars) or unique (gray bars) to TAF15 and FUS, TDP43, or both FUS and TDP-43. (d) Bar graph of log2 fold change in gene expression (from RNA-seq, as measured by DESeq2) and the standard error for TAF15, FUS, and TDP-43 upon TAF15, FUS, or TDP-43 knockdown. (e) Mean intron length for all introns in genes that are either upregulated or downregulated by FUS, TAF15 or TDP-43. Error bars represent standard deviation. (f) Percent of genes that are downregulated, upregulated, or unchanged upon ASO-mediated knockdown of the indicated RBP that are either bound (Target, gray) or not bound (Non-target, black) by that RBP in the 3UTR. Asterisks denote significant difference between target and non-target genes by Fishers exact test. Error bars represent standard deviation.

Figure 4.

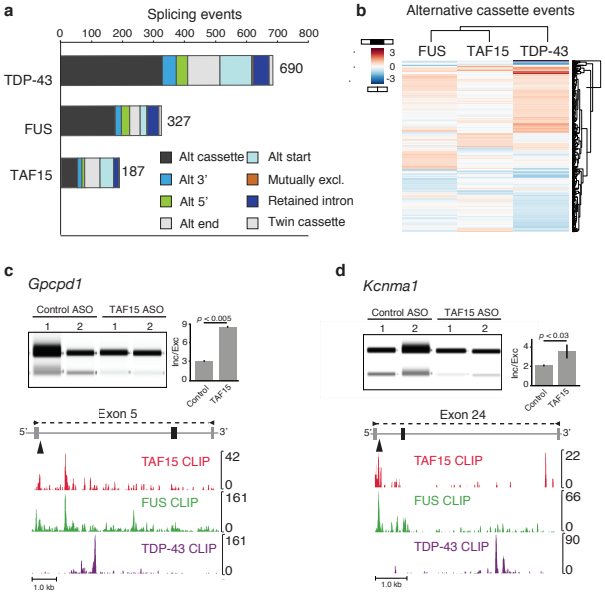

Figure 2.7: **Figure 4. TAF15 influences alternative splicing for a small subset of transcripts** (a) Bar graph showing the number of alternative splicing events altered upon ASO-mediated depletion of TDP-43, FUS, or TAF15 in the mouse striatum, as detected by splicing-sensitive microarray analyses. (b) Heatmap of alternative cassette events in (a) altered by FUS, TAF15, or TDP-43 depletion. Hierarchical clustering analysis was performed using separation (Sep) scores. Higher Sep scores (red) indicate inclusion events and lower Sep scores (blue) indicate exclusion events. (c) RT-PCR for exon 5 of glycerophosphocholine phosphodiesterase 1 (Gpcpd1) (chr2:132,382,646-132,390,412) to assess alternative splicing in TAF15 knockdown samples compared to controls. Quantification of biological replicates is shown. Error bars represent standard deviation. Binding of TAF15 (red), FUS (green), and TDP-43 (purple) in the mouse brain is shown below. (d) RT-PCR of exon 24 in the potassium channel, calcium activated large conductance subfamily M alpha, member 1 (Kcnma1) (chr14:24,149,961-24,156,401) to assess alternative splicing in TAF15 knockdown samples compared to controls. Quantification of biological replicates is shown. Error bars represent standard deviation. Binding of TAF15 (red), FUS (green), and TDP-43 (purple) in the mouse brain is shown below.

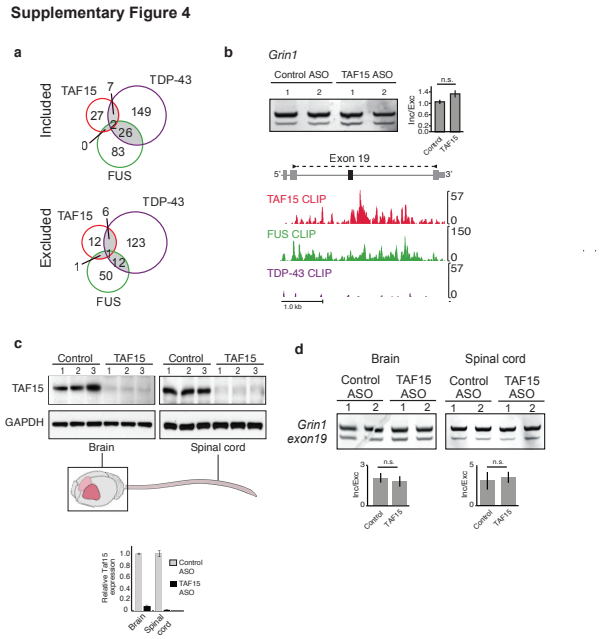

Figure 2.8: **Supplementary Figure 4. Effect of TAF15 depletion on alternative splicing** (a) Venn diagrams showing common and unique inclusion (top) or exclusion (bottom) cassette events upon loss of FUS (green), TAF15 (red), or TDP-43 (purple). (b) RT-PCR of exon 19 of Glutamate receptor ionotropic, NMDA 1 (Grin1; chr2:25,294,438-25,294,549) to assess alternative splicing in TAF15 knockdown samples compared to controls. Quantification of biological replicates is shown. Error bars represent standard deviation. Binding of TAF15 (red), FUS (green), and TDP-43 (purple) in the mouse brain is shown below. (c) Confirmation of reduced TAF15 expression in the mouse brain or spinal cord by Western blot analysis (top) and qPCR (bottom). Knockdown was achieved by intracerebroventricular injection of ASOs complementary to TAF15 or a non-murine/human gene (Control). (d) Splicing validation by RT-PCR for exon 19 of Grin1 upon ASO-mediated TAF15 depletion in the mouse brain or spinal cord compared to a Control ASO. Quantification of biological replicates is shown. Error bars represent standard deviation. n.s. indicates no significant difference.

Figure 5.

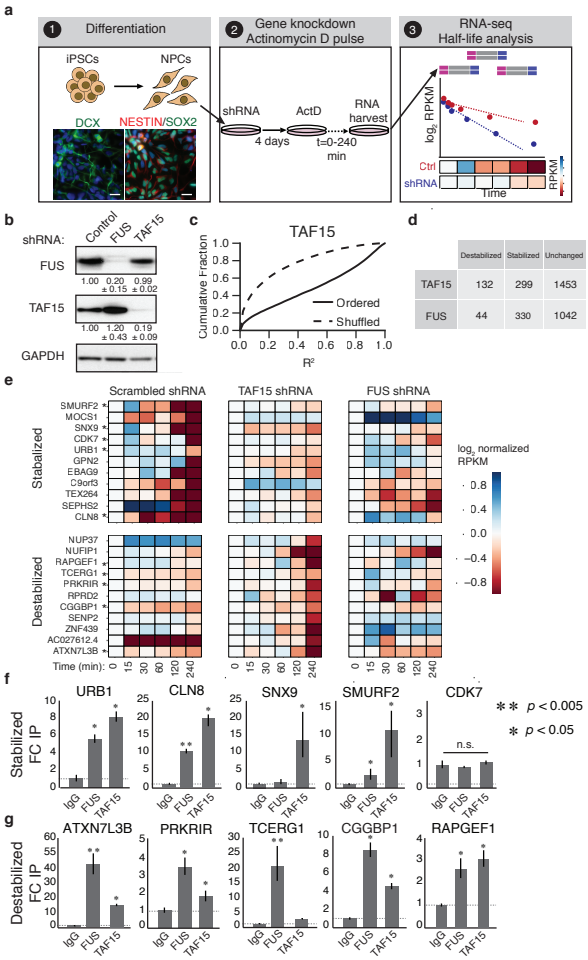

Figure 2.9: **Figure 5. Loss of TAF15 or FUS affects mRNA stability in human neural precursor cells** (a) Schematic of workflow. (1) Human iPSCs were differentiated into neural progenitor cells (NPCs) and stained for neuronal lineage markers DCX (left, green), NESTIN (right, red), SOX2 (right, green) to confirm differentiation. Dapi (blue) was used to locate nuclei. Scale bar: 25 m. (2) NPCs were infected with virus expressing shRNAs against TAF15 or FUS and then treated with Actinomycin D for indicated duration. (3) Poly(A)-selected RNA was converted into libraries for sequencing and sequencing reads were used to calculate mRNA half-lives. (b) Validation of shRNA-mediated knockdown of FUS or TAF15 in NPCs by Western blot analysis. Representative Western blot is shown from one replicate with quantifications from biological triplicate knockdown experiments. (c) For each gene, the coefficient of determination (R2) reflecting the fit of the RPKM values to a log linear regression was computed. The cumulative distribution functions of the R2 values for all genes in the TAF15 depletion experiment are depicted for real and shuffled values. (d) Table displaying the number of mRNAs whose half-lives were destabilized, stabilized, or unchanged by depletion of TAF15 or FUS. Half-life changes, measured as log2 (knockdown/control), that were greater than 1 were considered. (e) Heatmap of normalized RPKMs for stabilized and destabilized mRNAs upon shRNA-mediated knockdown of TAF15 or FUS. RPKMs are normalized for each gene to its RPKM at time 0. An asterisk indicates that the gene was examined for binding in (f). (f) RNA immunoprecipitation was performed using antibodies against IgG (Control), TAF15 and FUS in NPCs. The relative fold change compared to the IgG control for genes that are stabilized by TAF15 loss, was determined by qPCR. Values are means standard deviation for biological duplicates. Asterisk denotes a significant difference compared to IgG by Students t test where **p¡0.005 and *p¡0.05. (g) RNA immunoprecipitation analysis as in (g) for mRNAs that were destabilized upon TAF15 loss.

Supplementary Figure 5

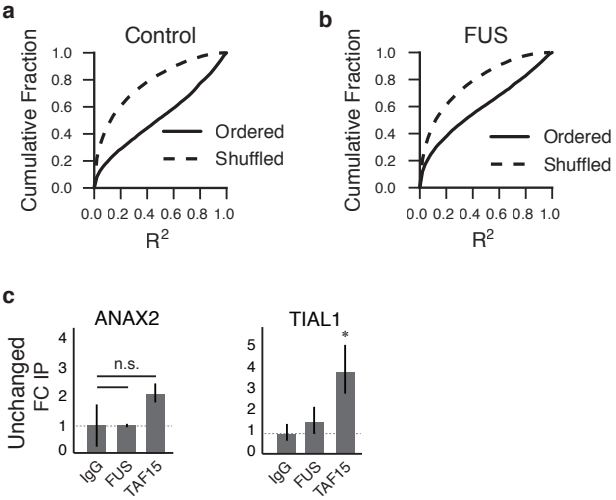

Figure 2.10: **Supplementary Figure 5. Transcriptome-wide analysis of mRNA decay upon loss of TAF15 or FUS** (a, b) For each gene, the coefficient of determination (R2) reflecting the fit of the RPKM values to a log linear regression was computed. The cumulative distribution functions of the R2 values for all genes in the FUS depletion and scrambled shRNA treated experiments are depicted for real and shuffled values. (c) RNA immunoprecipitation was performed using antibodies against IgG (Control), TAF15 and FUS in NPCs. The relative fold change compared to the IgG control for ANAX2 and TIAL1, RNAs that are not altered by TAF15 loss, was determined by qPCR. Values are means SD for biological duplicates. Asterisk denotes a significant difference compared to IgG by Students t test where p¡0.05 and n.s. indicates no significant difference.

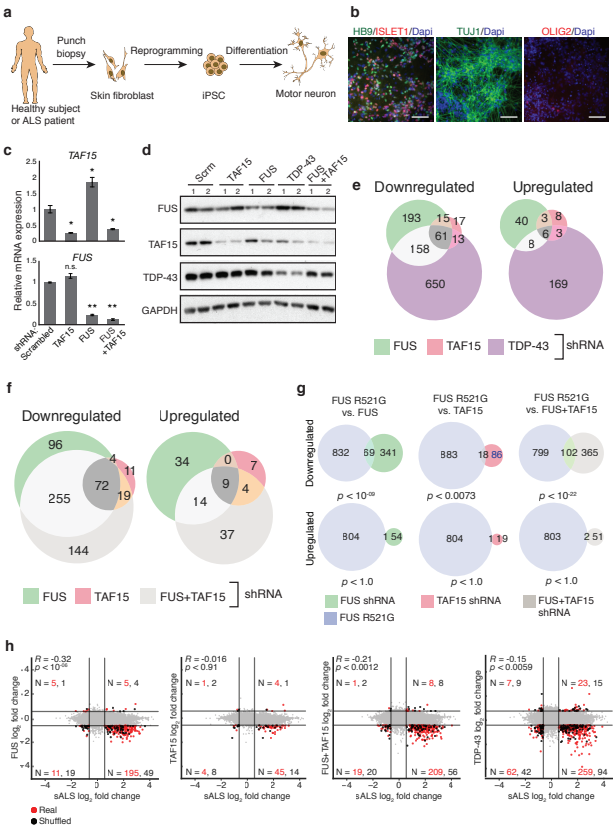

Figure 2.11: **Figure 6. Comparison of motor neuron RNA signatures upon TAF15, FUS, or TDP-43 loss to two models of ALS** (a) Schematic of workflow to reprogram iPSCs and differentiate into motor neurons (MNs). (b) Immunofluorescence of human iPSC-derived MNs for motor neuron marker HB9 (green), post-mitotic neuronal marker TUJ1 (green), neural stem cell marker ISLET1 (red), and oligodendrocyte marker OLIG2 (red). Dapi stain marks cell nuclei (blue). Scale bar: 25 m. (c) qRT-PCR and (d) Western blot validation of shRNA-mediated depletion of TAF15, FUS, and TDP-43 in MNs. Error bars represent standard deviation from biological duplicate experiments. (e,f) Venn diagrams showing overlap of up- and down-regulated genes in MNs upon depletion of TAF15, TDP-43, FUS or simultaneously, FUS and TAF15 (FUS+TAF15). (g) Venn diagrams showing overlap of up- and down-regulated genes between MNs with the FUS R521G mutation and knockdown of TAF15, FUS or FUS+TAF15. Statistical significance was determined by a hypergeometric test using genes expressed in MNs as background. (h) Scatter plots comparing gene expression changes (log2 RPKM) in MNs from sALS patient samples compared to loss of TAF15, FUS, TDP43 or FUS+TAF15. Each quadrant of a scatter plot shows genes (red dots) and gene counts (N, in red) that are significantly changing in sALS and RBP depletion experiments. Genes from a randomly ordered comparison are also shown (black dots) along with gene counts (N, in black). R2 and p values from linear regression analyses of genes significantly changing in both sALS and RBP knockdown experiments are shown.

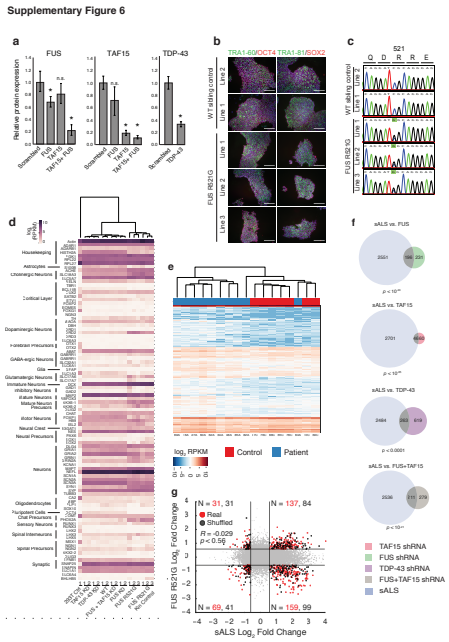

Figure 2.12: **Supplementary Figure 6. Characterization of motor neuron model systems of sALS** (a). Densitometric quantification of Western blots in Fig. 6d. Asterisk denotes a significant difference compared to control (scramble) shRNA by Students t test at p¡0.05. Error bars represent standard deviation. (b) Immunofluorescence staining of ALS-patient and wild-type (WT) sibling control iPSCs for pluripotency markers TRA1-60, OCT4, TRA1-81, and SOX1. Dapi stain marks cell nuclei (blue). Scale bar = 100 m. (c) Sanger sequencing results of iPSC lines in (b) showing the wild-type (c1561) or R521G mutation (c1561g) in the FUS gene. (d) Heatmap comparing the expression in log2 RPKMs of 93-gene RNA signature of glial, astrocyte, oligodendrocyte and neuronal subtype markers in HEK293T cells, iPSC-derived human motor neurons (MNs) where TAF15, TDP-43, FUS, or both TAF15 and FUS have been knocked down (KD) in WT cells, and iPSC-derived MNs from ALS-patient fibroblasts harboring the FUS R521G mutation (FUS R521G) or the sibling control (FUS R521G Kin Control) and a scrambled shRNA treated FUS R521G Kin control (Line 3). Low expression is grey and high is purple. (e) Clustergram of log2 RPKMs from sALS patient samples (blue) and control samples (red) used in the differential gene analysis. (f) Venn diagrams showing overlap of genes upregulated in sALS patient samples and genes downregulated upon loss of TAF15, FUS, TDP-43, or FUS+TAF15. P values were derived by hypergeometric test. (g) Scatter plot comparing gene expression changes in log2RPKMs in MNs from sALS patient samples compared to expression in FUS R521G mutants. Each quadrant of a scatter plot shows genes (red dots) and gene counts (N, in red) that are significantly changing in sALS and RBP depletion experiments. Genes from a randomly ordered comparison are also shown (black dots) along with gene counts (N, in black). R2 and p values from linear regression analyses of genes significantly changing in both sALS and RBP knockdown experiments are shown.

Supplementary Figure 7

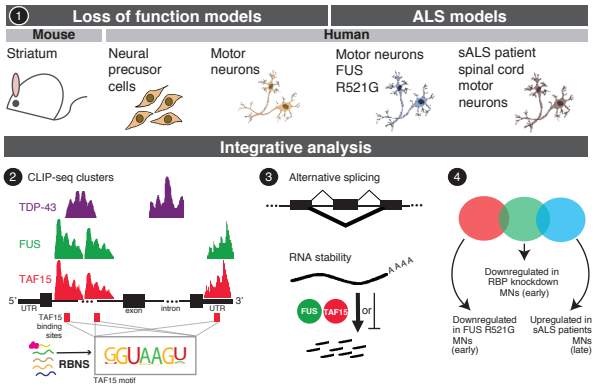

Figure 2.13: (1) Multiple model systems were employed to examine the neuronal functions of the ALS-associated RBPs TAF15, FUS, and TDP-43 through loss-of-function studies and compare these findings to models of ALS. Genome-wide studies such as (2) CLIP-seq revealed similar binding profiles for FUS and TAF15, but not TDP-43, and identified a novel TAF15 binding motif that was validated by RBNS. (3) TAF15 loss causes minimal changes in alternative splicing compared to FUS and TDP-43, but generally promotes exon skipping (bold line). TAF15 and FUS affect the stability of distinct mRNAs. (4) There is a small yet significant overlap of RNA signatures from motor neurons (MNs) depleted of TAF15, FUS, or TDP-43 with MNs differentiated from FUS R521G ALS patient fibroblasts (representative of an early stage of sALS) or obtained from sALS patient spinal cords (representative of a late stage of sALS).

Chapter 3

Guidelines and Best Practices for enhanced CLIP experiments and analysis

3.1 ABSTRACT

Enhanced cross-linking immunoprecipitation (eCLIP) featuring a size-matched input control has been recently applied to profile the binding sites of more than one hundred RNA binding proteins (RBPs). However computational pipelines and quality control metrics needed to process CLIP data at scale have yet to be well defined. Here, we describe our ENCODE eCLIP processing pipeline (https://github.com/YeoLab/eclip), enabling users to go from raw reads to processed peaks that are enriched above paired input, reproducible across biological replicates, and can be directly compared against the public ENCODE eCLIP resource. In particular, we discuss processing steps designed to address common artifacts, including properly quantifying unique RNA fragments bound by both unique genomic- and repetitive element-mapped reads. Using manual quality annotation of 350 ENCODE eCLIP experiments, we develop metrics for quality assessment of eCLIP experiments prior to and after sequencing, including library yield, number of unique fragments in the library, total binding relative information, and biological

reproducibility. In particular, we quantify the commonly believed linkage between depth of sequencing and peak discovery, and derive methods for estimating required sequencing depth based on pre-sequencing metrics. Finally we provide recommendations for the common question of integrating RBP binding information with RNA-seq to generate splicing maps representing the positional effect of binding on alternative splicing. These pipelines and QC metrics enable large-scale processing and analysis of eCLIP data, and will help to standardize rigorous analysis of RBP binding data.

3.2 INTRODUCTION

RNA binding proteins (RBPs) are major factors in the post-transcriptional regulation of gene expression. Recent estimates indicate that approximately 1,500 RBPs exist in the human genome[RZB$^+$08]. Collectively, these RBPs affect base modification, splicing, translation, nuclear export, subcellular localization, and stability of messenger RNAs, as well as processing of non-coding RNAs including microRNAs, ribosomal RNAs, and repetitive elements such as retrotransposons [Bar04, HPCO$^+$15, HPS$^+$15, ZKT$^+$13]. Manipulation of protein-RNA interactions is important in many aspects of human development, underscored by the identification of causal mutations in RBPs for diseases such as amyotrophic lateral sclerosis, fragile X syndrome, and cancers including medulloblastoma and chronic lymphocytic leukemia [DVK$^+$09, VPS$^+$91, PWA$^+$12, WLW$^+$11].

A critical first step in elucidating the function of RBPs is the identification of its RNA targets in vivo. Approaches such as UV crosslinking and immunoprecipitation followed by sequencing (CLIP-seq or HITS-CLIP) have enabled the transcriptome-wide discovery of RBP-RNA interactions at high resolution [WRV80, Gre79, UJR$^+$03]. These protein-RNA maps can reveal insights into the molecular roles of RBPs [YCL$^+$09, UJR$^+$03]. However, traditional CLIP-seq suffers from low ligation and crosslinking efficiencies, leading to poor experimental success rates

and low reproducibility [VPS+16]. Variants of CLIP have addressed this in many ways, including increasing crosslinking efficiency with 4-thiouridine (PAR-CLIP) [HLB+10] and the incorporation of unique molecular identifiers (UMIs) to identify PCR duplicates (iCLIP) [KZR+10]). Recently, enhanced CLIP (eCLIP) [VPS+16] and infrared-CLIP (ir-CLIP)[ZFS+16] describe major improvements in the efficiency of generating sequencing libraries that led to dramatic increases in experimental success. This improved efficiency has enabled the incorporation of a paired size-matched input control experiment in eCLIP, which increases specificity in identifying high-confidence binding sites [VPS+16]. This approach has opened the door to large-scale profiling of RBP targets in vivo, enabling the ENCODE consortiums efforts to profile 126 RBPs by eCLIP (Van Nostrand et al., in preparation).

With this considerable increase in the scale of eCLIP experimentation, there is a concomitant demand for improved methods to process eCLIP datasets and assess the quality of eCLIP datasets in an automated and scalable manner. Due to the variation in methodologies and low number of RBPs profiled in typical studies, current CLIP-seq quality control (QC) has traditionally been performed on a case-by-case basis, looking at individual binding locations [KZR+10, UJR+03], comparing results to known motifs [RKC+13, CKZ+11] or referring to previously published binding sites or known functions for the same RBP [WVMY+14, ZKT+13]. Indeed, the first pass of quality scoring for ENCODE eCLIP datasets was performed by manual inspection [VPS+16]. However, these methods cannot often be performed for previously unstudied RBPs and do not scale, necessitating more structured metrics for the analysis of large-scale eCLIP datasets.

A similar requirement for robust, scalable quality metrics led the Chromatin ImmunoPrecipitation followed by sequencing (ChIP-seq) field to create a set of guidelines for sequencing depth, experimental design, and data quality. Including the use of Irreproducible Discovery Rate (IDR) analysis to determine reproducibility between biological replicates and provide highly confident and reproducible binding sites [LBHB11, KTP08, JLH+14]). These approaches were

invaluable to standardizing the large-scale data generation performed as part of the ENCODE project and have become useful tools for the genomics community at large [DKA+12, LMK+12]. Strikingly, a re-analysis of public ChIP-seq data using these standards found that 20% of published datasets were of low quality, confirming the value of such metrics in increasing robustness and reproducibility of trusted public datasets [MKPW14].

Here we describe the development and implementation of data quality metrics and standards for eCLIP experiments performed as part of the ENCODE consortium efforts (Fig. 1A). Each of these datasets was manually assessed during data production, creating a reference set of high- and low-quality eCLIP experiments for proper evaluation of derived metrics. First, we developed experimental quality metrics to rigorously assay eCLIP experimental success prior to sequencing. Next, we developed a uniform processing pipeline to analyze 350 eCLIP datasets and evaluate the downstream effects of various processing choices [VPS+16]. Using these datasets, we derived effective computational quality control metrics to aid in analysis decisions. Finally, similarly to previous CLIP-seq pipelines [WXX14, ZD11, UBSB+12, AGVB+11, LGM+13, CGM+11], we provide a portable set of eCLIP processing tools to reproduce data generated by the ENCODE project, process data newly generated data, and create splicing maps of position-dependent regulatory activities. These tools and metrics will serve as a reference for future experiments within the ENCODE project and for the increasing number of other labs performing eCLIP.

3.3 RESULTS

3.3.1 Experimental Quality Control Considerations

The generation of high-quality, reliable RBP-RNA interaction datasets requires careful validation of two key experimental components: successful immunoprecipitation of the desired RBP, and successful library generation and sequencing. For these two components, we developed

recommended metrics that enable rigorous validation of experimental success.

First, eCLIP experiments require the successful immunoprecipitation of a RBP of interest. This prerequisite requires the identification of a RBP-specific immunoprecipitation-grade antibody, which we previously addressed by screening over 700 antibodies to identify 438 IP-grade antibodies against 365 RBPs in K562 cells [SZB$^+$16]. However, when we performed eCLIP we found that 18% (54 out of 302) IP-grade antibodies failed to successfully immunoprecipitate the desired RBP in K562 cells (Supplemental Fig. S1A-B). This is perhaps unsurprising as eCLIP has numerous on-bead enzymatic steps and additional wash steps. Nevertheless this indicates that it is critical to perform an IP-western experiment as part of the standard eCLIP experimental protocol, as successful IP during eCLIP is not a given based on successful IP with other protocols. As such, we incorporated IP-western as a routine component of the eCLIP procedure, and IP-western images are provided for each ENCODE eCLIP experiment as part of the antibody metadata available at https://www.encodeproject.org.

A high-quality library is defined as one that is complex (i.e., it contains many unique RNA fragments relative to PCR duplicated fragments or other artifacts) and is critical for a successful eCLIP experiment. Thus, a quantitative metric for library complexity that can be applied prior to sequencing enables rapid culling of poor quality experiments and could help guide a desired sequencing depth by estimating an upper bound on the number of recovered RNA fragments. We previously introduced the extrapolated CT (eCT) metric that estimates the number of PCR cycles needed to obtain sufficient material for sequencing. This metric had appealing characteristics, as it was RBP-specific, showed high correlation with PCR duplication rate, and could be directly compared against eCLIP experiments performed with IgG isotype controls or antibodies in null cell lines [VPS$^+$16, VGBW$^+$17].

However, we found that additional refinements were required to use eCT to derive recommendations for the optimal depth of sequencing. Although the initial eCT calculation assumed an idealized 2-fold amplification rate per PCR cycle, we observed that this rate is

frequently lower in practice. To properly estimate PCR efficiency during eCLIP, we noted that at our standard sequencing depths some experiments had saturated the discovery of unique fragments, which enabled us to accurately estimate the total number of pre-PCR unique fragments for these datasets. Using 6 datasets with a PCR duplication rate of greater than 90%, we observed that the best fit between the number of observed unique fragments, and the estimated number of unique fragments, occurred at a PCR efficiency of 1.84 (Supplemental Fig. S1C,D). We defined an accurate-eCT (a-eCT) as the eCT calculated with 1.84-fold amplification per cycle instead of 2-fold.

To validate the a-eCT metric, we considered datasets that were beginning to saturate (PCR duplication rate greater than 60%). We observed that a-eCT showed strong predictive power for the number of unique RNA fragments observed (R^2 = .48, p $<$ 3.7e-38) (Fig. 1B), an improvement on the prior eCT metric (MSE 0.16 vs 0.82), confirming that a-eCT provides a robust estimate of library complexity (Supplemental Fig. S1E).

Next, we compared a-eCT against our manual annotation of experiment quality (Fig. 1A), and observed that experiments that pass manual quality assessment have a significantly lower a-eCT than experiments that failed manual quality assessment with mean a-eCTs of 13.2 vs 14.4 respectively (Fig. 1C, students t-test; p $<$ 10-7). Low a-eCT (corresponding to a highly complex library) did not always indicate high-quality eCLIP datasets, with failures due to poor reproducibility, lack of significant binding signal, and other failure modes (further discussed below). However, a high a-eCT value was a strong predictor of failure, typically due to a lack of the required number of unique fragments to produce reproducible binding regions. To establish a maximum a-eCT threshold beyond which data are unreliable, we observed that the mean a-eCT for IgG control eCLIP experiments (which only pull down background RNA) was 19.6. With that threshold applied, 18 out of 21 datasets with an a-eCT $>$ 19.6 also independently failed manual QC. In all datasets examined no successful experiment had an a-eCT ¿ 20.7, while there were still 9 experiments that did not pass manual quality control that had a higher a-eCT (Fig. 1C).

In addition to providing a maximum cutoff for experiment quality, the strong predictive power among saturated datasets suggests that a-eCT can be used to estimate the total number of unique immunoprecipitated RNA fragments prior to sequencing (Fig. 1B). Considering all RBPs profiled, we observed a wide range of unique fragments from nearly 2 billion estimated for HNRNPU, to less than 320,000 for DROSHA (Fig. 1D). Thus, a-eCT can also be used to help guide desired sequencing depth, as high sequencing depth for samples with high a-eCT is essentially wasted sequencing of PCR duplicates (see further discussion below).

3.3.2 eCLIP Processing Pipeline

Processing of raw eCLIP sequencing data is complex, as adapter sequences, double-adapter ligation products, retrotransposable elements and other multi-copy sequences, PCR duplicates, and underlying differences in RNA abundances all contribute to false negatives and false positives at both the read mapping and peak identification stages. To address these issues, we developed a rigorous standard eCLIP processing and analysis pipeline (Fig. 2A) [VPS$^+$16].

First, adapter sequences are aggressively trimmed to decrease potential false-negative mapping and to remove reads $<$18nt in length, which typically retains 91% of sequenced reads (Fig. 1B). We observed that if adapter trimming is not performed correctly, the remaining adapter sequences often lead to mapping failures. To illustrate, 53% of eCLIP fragments for the RBP RBFOX2 in HepG2 cells uniquely map to the genome if proper adapter trimming is performed, but only 35% map if adapter trimming is not performed (Supplemental Fig. S2A).

Next, we considered repetitive elements and other multi-copy RNAs. These can represent true RBP associations: for RPS11 an average of 80% of fragments mapped to repetitive elements including 68% to the 18S rRNA, which is biologically consistent as RPS11 is a component of the 40S ribosomal subunit [TDH$^+$09]. However, we observed that inclusion of these fragments in standard peak calling created significant artifacts, as these elements can have thousands of degenerate copies throughout the genome that lead to spurious peaks. For example, TIA1 and

PTBP1 both bind the 3 UTR of the MYADM mRNA, with enriched fragment density observed at a SINE element that is highly homologous to thousands of loci throughout the genome (Fig. 2C). If repetitive element fragments are not removed this SINE element is detected as a spurious peak between two authentic binding sites in both experiments (Fig. 2C). Thus, we developed two approaches: a standard peak calling and analysis approach that removes repeat-mapping fragments and only incorporates fragments that uniquely map to the genome, and another pipeline that quantitatively measures mapping to multi-copy element families (see below) (Fig. 2A). On average, 42% of fragments were removed due to repetitive element mapping, and 33% of sequenced fragments were uniquely mapped (with a standard deviation of 0.16 across 362 ENCODE eCLIP datasets passing manual quality control) (Fig. 2B). Although high, this variability is largely due to RBPs that associate with specific multi-copy RNAs (although we note that in cases of extremely low RNA yield, we have observed that bacterial RNA present in nitrocellulose membranes can contribute to the lack of mapping to the human genome) [VNGB$^+$17].

After mapping, PCR duplicate fragments are removed (using UMIs incorporated in library preparation prior to PCR amplification) to yield unique genomic fragments. An average of 66% of unique genomic fragments remained after removal of PCR duplicates, although we observed an extreme range from 99.5% of fragments remaining for hnRNPK in K562 cells to 4.6% of fragments remaining for SF3B1 in K562 cells (Fig. 2B). We note that proper PCR duplicate removal is essential, as the presence of PCR duplicate reads can lead to regions of artificially high fragment coverage (often referred to as skyscrapers) that can be falsely identified as significant peaks, as in the case of hnRNPC (Fig. 2D). Overall, on average of 22% of sequenced fragments remained after removing reads that were too short, non-uniquely mapping or PCR duplicates (Fig. 2B).

As multi-copy signals are not captured with our standard pipeline we developed an independent approach to quantify fragments that map to either multi-copy RNAs or unique

genomic elements (Van Nostrand et al. in preparation). Datasets subjected to mapping to either repetitive or unique segments of the genome retained an average of 73% of sequenced fragments for all RBPs, and were highly consistent, ranging from 71% of sequenced fragments retained for SF3B1 in K562 to 85% of sequenced fragments retained for CENPI in HepG2 (Fig. 2B). The fraction of PCR duplicate removed fragments was highly variable, with an average of 46% of sequenced fragments retained after duplicate removal (see further discussion below) (Fig. 2B). In manual quality assessment of these datasets, we identify 25 experiments that did not pass based on standard analysis but were deemed high quality due to significant association to specific repetitive elements (Van Nostrand et al. in preparation).

Finally, significant peaks are identified using a two-step approach in which we first identify clusters enriched relative to local background, followed by input normalization. Input normalization is critical for the removal of common false-positive signals at abundant transcripts and enriches for motifs as well as RBP-responsive targets [VPS$^+$16]. Cluster identification is performed using CLIPper, which uses spline-fitting to identify regions of enriched read density relative to both unspliced and spliced transcript-level background [LGM$^+$13]. Reverse transcriptase enzymes will often terminate at the crosslinked nucleotide (due to amino acid adducts that remain after proteinase K treatment), which can be utilized to map crosslink sites with single-nucleotide resolution [KZR$^+$10]. As such, CLIPper was run using only the second, paired-end read in ENCODE eCLIP datasets, which is the read that begins at the putative site of reverse transcription termination. This yielded an average of 142,360 clusters per dataset, with a range from 647,011 clusters for hnRNPC in HepG2 to 1,910 clusters for SERBP1 in K562 (Fig. 2E). The read density within these regions is then compared in IP versus paired size-matched input to identify significantly enriched peaks [VPS$^+$16]. We observed that a significant fraction (averaging 13%) of clusters are depleted in IP relative to size-matched input, matching previous observations [VPS$^+$16]. Applying a stringent threshold of requiring clusters to satisfy both fold-enrichment 8 and p 10-3 significance in IP versus size-matched input yielded an average of

7,146 high-confidence peaks, with a range from 48,173 peaks for PRPF8 in HepG2 to 43 peaks for AUH in HepG2 (Fig. 2E). On average, 6% of clusters met this stringent threshold for enrichment (Fig. 2F). While we find that filtered clusters frequently fail both thresholds (an average of 68% of clusters fail to meet both the fold-enrichment and significance threshold) (Supplemental Fig. S2B).

To identify reproducible and significantly enriched peaks across biological replicates, we used a modified Irreproducible Discovery Rate (IDR) method (Supplemental Fig. S2C). IDR requires that peaks are ranked by an appropriate metric, but we found undesirable results ranking peaks by either significance (due to the dependence on underlying expression) or fold-enrichment (due to the large variance of fold-enrichment when few reads are observed in input). Thus, we adapted relative entropy to better estimate the strength of binding in IP relative to input by defining the information content of a peak as $p_i * \log_2 \frac{p_i}{q_i})$, where p_i and q_i are the fraction of total reads in IP and input respectively that map to $peak_i$. To confirm that this metric captures true binding signal, we considered the RBFOX2 eCLIP datasets. We observed 14,595 reproducible clusters when we ranked by fold enrichment, whereas 32,431 clusters were reproducible when we ranked by information content (Fig. S2D). Given the increased number of reproducible clusters detected, we used information content to perform standard IDR analysis to identify reproducibly bound regions [LBHB11]. We then identified the set of non-overlapping peaks from both replicates that maximized information content to define a final set of reproducibly enriched peaks that corresponded to CLIPper-identified regions (see Methods). This method revealed that an average of 53.1% of peaks were identified as significantly enriched in individual replicates, confirming high reproducibility for most experiments (Fig. 2G).

3.3.3 Depth of sequencing does not significantly affect peak quality

How deeply to sequence a CLIP-seq dataset is a major consideration (particularly at large scale), as samples must be sequenced sufficiently to robustly detect true binding signals

while minimizing experimental cost. To develop recommendations for required sequencing depth, we considered two questions: first, how does sequencing depth affect identification of true binding sites, and second, how many reads are required to detect binding sites in any gene when accounting for variability in gene expression.

First, we asked whether peaks discovered at deeper sequencing depths were still likely to be biologically relevant. To do this we looked at RBFOX2, which is known to bind to the GCAUG motif. Overall, we observed significant enrichment for RBFOX2 binding to its motif, with 36% of RBFOX2 peaks overlapping the motif vs a mean of 6% of peaks overlapping the motif in all other datasets (Supplemental Fig. S3A). We then down-sampled the unique genomic fragments, re-called peaks, and asked how many peaks discovered at each down-sampling step overlapped the RBFOX2 motif. We observed that peaks discovered using only 10% of unique genomic fragments showed the highest motif overlap (38% on average), whereas peaks that were only discovered when going from 90% to 100% of unique genomic fragments were less likely to contain GCAUG (27% on average) (Fig. 3A). Although this suggests that signal to noise is highest among the most abundantly covered peaks, we note that later discovered peaks were still 2.8- to 7.0-fold enriched above non-RBFOX2 datasets, indicating they still contain significant true binding signal (Fig. 3A). Supporting this, we observed that conservation of later-discovered peaks was similar to those discovered earlier with a mean phastcons conservation score of 0.136 versus 0.132 (Supplemental Fig. S3B). Considering an independent dataset, PRPF8, we observed similar results when testing its known association with the 5 splice site: although peaks discovered at low sequencing depth were less enriched for true signal, we continued to see significant true positive signal throughout the range of down-sampling, indicating that it is true that deeper sequencing allows for the continued discovery of high quality peaks (Supplemental Fig. S3C, D).

Second, we considered the identification of binding sites as a function of sequencing depth. To explore if there was a correlation between sequencing depth and the discovery of peaks in lowly expressed genes, we calculated the correlation between gene expression and the number

of reads in each peak for RBFOX2. We observed that lowly expressed genes had fewer reads per peak (as expected), whereas highly expressed genes displayed a large variation in the number of reads per peak, with only a weak correlation overall for both RBFOX2 (R^2 = .03) (Fig. 3B). All other RBPs showed a similar week correlation (mean R^2 = 0.24) (Supplemental Fig. S3E).

Next, we asked whether peaks at lowly expressed genes could be detected at standard sequencing depths. Surprisingly, we found that lowly expressed genes (defined as those with TPM < 1) need on average only 870,000 unique genomic fragments to allow for detection of a peak in the gene, and this estimate was similar when varying the fraction of peaks required to be discovered or TPM thresholds (Fig. 3C, Supplemental Fig. S3F-H). As ENCODE eCLIP datasets have a mean sequencing depth of 4,509,000 unique genomic fragments, these results suggest that an inability to detect peaks on lowly expressed genes is not a major concern in eCLIP data sequenced to standard depths.

3.3.4 Saturation analysis suggests optimal sequencing depth

Our analysis above indicates that continued sequencing until fragment saturation can recover true binding sites even at extremely high read depths. However, sequencing until fragment saturation is not typically economically reasonable. Thus, we set out to quantify diminishing returns upon deeper sequencing to identify the sequencing depth that optimally balances the tradeoff between gaining additional binding information and increased sequencing depth for an average eCLIP experiment (Fig. 4A).

First we developed a metric to quantify the diminishing returns of deeper sequencing in eCLIP datasets. Considering the discovery of significant peaks, we queried how many peaks were newly discovered when comparing peaks observed when 90% or 100% of fragments in a dataset were used to identify peaks. We observed that 71% of experiments passing manual QC saturated the discovery of significant peaks (defined as the discovery of fewer than 5% new peaks in the above metric), suggesting that simple peak detection was saturating for most but not all

high-quality datasets (Fig. 4B).

We next considered whether binding information by total information content was saturating even when peak discovery was not. Summing the total information content across all peaks, we observed that information recovered saturated for 98% of manually accepted datasets (using the same 5% or less discovery metric between 90% and 100% of fragments used to call peaks in a dataset) (Fig. 4B). Thus, these results suggest that although additional peaks can be identified, the vast majority of binding information is already captured. We next asked at what sequencing depth total information content tends to saturate. We found that 90% of all eCLIP datasets that passed manual quality assessment had saturated information discovery by 7.7M unique fragments (corresponding to 4.1M unique genomic fragments) (Fig. 4C, Supplemental Fig. S4A).

Next, we developed a model to estimate the number of sequenced reads that would typically be necessary to obtain the 7.7M suggested number of unique fragments. We observed that an average of 8.9% and 4.7% of reads are lost during adapter trimming and mapping, respectively (Fig. 2B). We also noted that when datasets are not near saturation (fraction usable $>$90%) the PCR duplicate removal rate was on average 5.0% (Supplemental Fig. S4B). Thus, in situations where there is a highly complex library, the number of unique fragments from sequenced reads can be estimated as $U = (1-P(a))(1-P(m))(1-P(d)) * S$, where U is the number of unique fragments, S is the number of sequenced reads, and $P(a)$, $P(m)$, and $P(d)$ are the probabilities of losing a read due to adapter trimming, mapping and PCR duplicate removal, respectively. This model works well when datasets have a low PCR duplication rate (fraction usable $>$90%) (R^2 = 0.91) (Supplemental Fig. S4C). However, because PCR duplicates often account for a large fraction of fragments lost, this method performs poorly when taking saturated datasets into account (R^2 = .55) (Supplemental Fig. S4C).

To address this limitation, we modeled PCR duplicate removal. Considering a pre-amplified library with E unique fragments PCR amplified to obtain a final library of 100 femtomoles, we model the random sampling of individual reads as Poisson distributed. Specifically,

the probability that a fragment is observed at least once can be calculated as $P(\leq 1) = (1 - e^{-\lambda})$, where λ can be modeled as $\frac{M}{E}$ where M represents the number mapped fragments and E represents the number of unique pre-amplified fragments (which can be estimated from a-eCT). Thus, the number of unique fragments U obtained from M mapped reads can be estimated as $U = E(1 - e^{-(\frac{M}{E})})$ (Fig. 4D). This model predicted the actual number of unique fragments in each dataset with generally high accuracy (Fig. 4E, R^2 = .0.75). To further test this model, we compared the estimated number of unique fragments against the observed number of unique fragments identified in downsampling experiments as described above, and observed high correspondence for most datasets (median R^2 = 0.91) (Supplemental Fig. S4D). Thus, the Poisson model provides a highly accurate estimate for unique fragments obtained that is independent of sequencing and relies solely on the experimentally-obtained a-eCT.

Finally, merging the two above approaches allows us to back-calculate the number of sequencing reads S necessary to observe 7.7M unique fragments $S = -E * \frac{ln(1-(U/E))}{P(a)P(m)}$. Using estimates for $P(a)$, $P(m)$ above yields final model $S = -E * \frac{\ln 1 - \frac{7{,}700{,}000}{E}}{0.868}$ where $E = \frac{100_{fm} * 6.022 * 10^8_{\frac{molecules}{fm}}}{1.84^{aeCT}}$ (Fig. 4F). Considering the full range of a-eCT values, we observe that 10M fragments are sufficient for the majority of values, with slightly higher requirements for a-eCT values between 12 and 15 (Fig. 4G). Above a-eCT of 15 (where there are fewer than 7.7M estimated unique fragments in the pre-amplified library) we recommend targeting sequencing to observe 90% of unique fragments (Fig. 4G, Supplemental Table S1). We find that the same depth of sequencing has proven sufficient for size-matched input experiments as well.

3.3.5 Automated QC Metrics verify data quality

We next developed a set of quality control metrics for ENCODE eCLIP experiments to assess the quality of individual datasets as well as the reproducibility across biological replicates (Fig. 5A). We ultimately arrived at two metrics for individual datasets: a minimal unique fragment cutoff, and a total information in peaks cutoff. To evaluate these metrics, we used manual quality

assessment of datasets to define a reference set of high- and low-quality eCLIP datasets.

The number of unique fragments per dataset varies widely, depending on library complexity and sequencing depth (as described above). We observed that a required number of 1.5M unique fragments maximized the predictive power for datasets passing manual quality assessment (f-score = .86) (Supplemental Fig. S5A). Only 7 of 516 manually passed datasets do not meet this threshold: two (TBRG4, PABPC4) are not yet saturated and thus could be rescued by re-sequencing, whereas the 5 other datasets (one replicate of SLBP, two replicates of SF3B1, and SUPV3L1) are already highly saturated, but were considered high quality due to presence of signal at a small number of specific RNAs matching previous studies of these RBPs (histones, the U2 snRNP, and mitochondrial RNA respectively) ([TTPMW06, HB94, BDH$^+$13]. Conversely, 30 of 184 manually failed datasets do not meet the criteria (Fig. 5B). Thus, although the classification power of this model is low (AUC = 0.61), datasets not meeting this threshold were more than 16-fold more likely to fail manual quality assessment (Supplemental Fig. S5B).

Next, we considered a metric based on whether the dataset contained significant binding signal. As described above, we observed that the relative information of a peak better captures the binding information of peaks across genes with widely varying expression levels. Thus, to validate that a dataset contains significant binding information, we calculated the sum of relative information across all peaks in the dataset. We observed that this total information content score maximized the f-score of manually annotated high- and low-quality datasets at a total information content of .044 (f-score = .89) (Fig. 5C, Supplemental Fig. S5C). The information content model was highly accurate (AUC = 0.75), accurately classifying 77% of ENCODE datasets with 0.48 specificity and 0.92 sensitivity (Supplemental Fig. S5D).

Next, we developed criteria to assay biological reproducibility, using two metrics based upon the Irreproducible Discovery Rate (IDR) approach that has previously been used to assay reproducibility of ChIP-seq peaks: reproducibility between real and pseudo-replicates (Rescue Ratio) (Supplemental Fig. S5E) and confirmation that the number of reproducible peaks be-

tween both replicates is similar (Self-Consistency Ratio) (Supplemental Fig. S5F) (Landt et al. 2012). We found that cutoffs previously used for ChIP-seq data could be similarly applied to eCLIP [LMK$^+$12], and observed that 81.4% of experiments have a passing rescue ratio of <2 (Supplemental Fig. S5G) and 70% of experiments have a passing self-constancy ratio of <2 (Supplemental Fig. S5H). 221 experiments pass both thresholds, while 89 are borderline (passing one of the two thresholds), and 40 fail both thresholds (Supplemental Fig. S5I). Notably, these IDR metrics have high specificity, as 180 out of 221 (79%) of experiments that meet unique fragment and total information content cutoffs and were manually judged to be high quality passed this IDR critera. In contrast, IDR detects potential false positives by correctly failing 9 out of 28 (32%) datasets that met read depth and information content metrics, but failed manual inspection (Fig. 5D).

Finally, we combined these metrics into one overall automated quality call requiring that each experiment passes minimum read and entropy cutoffs as well as either being classified as passing or borderline based on IDR metrics (Fig. 5E). Overall our model accurately classified 83% of eCLIP datasets with a sensitivity of 0.84 and a specificity of 0.79 (Fig. 5F), better than any individual classification scheme.

3.3.6 eCLIP Pipeline Implementation

We previously described the implementation of our eCLIP processing pipeline [VPS$^+$16]. However, implementation of the pipeline is non-trivial, and difficult to reproduce in other computational environments. To address the problem we have made available a reference version of the pipeline runs locally on a cluster through a pipeline definition in Common Workflow Language (CWL). Our eCLIP processing pipeline handles fastq demultiplexing, PCR duplicate removal and peak calling including input normalization (Fig. S6A). Additionally, we produce an automated quality control report, describing the metrics developed throughout the paper. A major advantage to this structure is to both simplify and standardize the definition of all associated metadata, which

is now explicitly structured in a single input manifest. Key output files match those deposited at the ENCODE Data Coordination Center, and include mapped reads (in compressed bam format), and identified clusters and significant peaks (in bed format). Due to being created with CWL this pipeline is highly portable, able to run on any mac or Linux installation with minimal effort.

3.3.7 Integration of eCLIP with RNA-seq to generate regulatory maps

Once direct binding sites are identified with eCLIP or related methods, it is common to integrate this information with independent datasets (such as transcriptome profiling upon RBP knockdown or over-expression) to gain insight into which interactions drive differential regulation of the bound transcript. This has proven particularly insightful for regulation of splicing, where RBP binding can cause either inclusion or exclusion of alternative exons depending on whether binding occurs in the upstream or downstream intron [YCL$^+$09]. This location-dependent splicing regulation is often visualized as a splicing map, and the generation of splicing maps for different RBPs has become an important visualization tool to provide mechanistic hints into the global regulation of alternative splicing by RNA binding proteins. However, we observed that multiple decision points regarding integration of eCLIP data into the splicing map, for example using either read density or peaks (each of which can be normalized in various ways) can yield significantly different maps [WU11, HVA$^+$12]. The scale of data generated here enabled us to query the effects of these choices across multiple datasets.

The first decision regards what events to consider in a splicing map. For visualization purposes, only a meta event is shown - for example, cassette exon events (the most commonly visualized splicing maps) are represented as a three exon region to show the upstream splice site, the 5 and 3 splice sites of the cassette exon, and the downstream splice site, although analogous approaches can be used to visualize alternative 5 and 3 splice sites or retained intron events (Fig. 6A, top). A set of exons is then defined to build the map, typically based off of a set of RBP-responsive events that show significant splicing alteration upon RBP knockdown or over-

expression. In generating these sets of exons, we noted that it was common for splicing analysis pipelines to identify overlapping alternative exons. However, if two or more skipped events overlap at any position, these positions become susceptible to integrating the same CLIP signal multiple times (Supplemental Fig. S7A). To avoid this, we conservatively group overlapping events and select only those with the highest inclusion junction count. We found that standard differential splicing analysis tools yielded an average of 20% overlapping events, indicating that this can represent a significant artifact if not considered (Supplemental Fig. S7B).

Next, we explored options for incorporating eCLIP data. First, we considered splicing maps built off of density of significantly enriched peaks [HVA$^+$12]. These maps are typically simplest to calculate, as peak density over the meta region is simply summed across all events. We found that this approach yielded clear maps for RBPs such as RBFOX2 with a large number of bound differential events (Supplemental Fig. S7C). However, the main drawback of this approach is limited power, as the mean number of peaks covering a skipped exon region in a ENCODE datasets is only 31 (Supplemental Fig. S7D). Indeed, for SRSF9, a map made using read density shows enrichment at events excluded upon SRSF9 knockdown; however, because there are only 4 peaks overlapping skipped exon regions, peak based analysis shows no significant sites (Fig. 6A).

We hypothesized that incorporating read density instead of simply peak location could improve splicing map power. Considering eCLIP data, we first noted that traditional splicing maps built on iCLIP and CLIP data do not contain normalization against a paired input [KZR$^+$10]. We observed that this approach could be applied to eCLIP data for RBPs such as HNRNPK that already have high signal-to-noise, where the map showed a significant enrichment for binding signal at exons included upon knockdown that recapitulates the general repressive role for HNRNP proteins (Fig. 6B, Supplemental Fig. S7E). However, we noted that if one generated a splicing map from input alone, there exists significant variation in binding density (with particular enrichment of signal at exons in general) (Fig. 6C, Supplemental Fig. S7F), suggesting that input normalization could improve signal-to-noise in identifying regions enriched above background.

To explore input normalization within CLIP density maps, we applied two normalization strategies: background subtraction and information content-based normalization (Fig. 6D). Background subtraction first normalizes the binding density across each event, and then subtracts the input density from a corresponding IP experiment. In this case, the binding profile at event is weighted equally, resulting in a map that reflects the global shape of binding at the cost of muting regions with signal from a small number of events (Fig. 6E, Supplemental Fig. S7G). As an alternative approach, we calculated the relative information (in IP versus input) at each position for each queried event. The distribution across all events is then used to create the splicing map (using the mean and standard error calculated across all events for each position). As relative information is dependent on abundance, in this approach more abundant binding events will contribute more to the overall splicing map, events with high input, and low IP abundances contribute less strongly to reduction in overall signal, which can be a desired property in some cases (Fig. 6F, Supplemental Fig. S7H-J).

In both cases, we observed that individual highly abundant positions at single events could dominate the composite signal. Manual inspection suggested that these often arise from miRNAs, pseudogenes, and other multi-copy or highly abundant transcripts present within these intronic regions. To address this, we performed outlier removal on the top and bottom 2.5% signal at each position across the splicing map. For example, we observed a site of significant enrichment approximately 250bp downstream of knockdown-excluded skipped exons for HNRNPC. However, after removing a single outlier element, the HNRNPC splicing map shows primarily 3 splice site and exonic signal at exons included upon knockdown, consistent with the splicing-repressive role of HNRNP proteins (Fig. 6G, Supplemental Fig. S7K-L)

Finally we explore the utility of utilizing CLIP crosslink-diagnostic events that allow for the detection of the exact theoretically crosslinking site [RHK$^+$14, KZR$^+$10]. For the previous analyses, we calculated read density by including bases covered by the entire read. However, reverse transcription often terminates at the site of protein-RNA crosslinking, which causes the

5 end of reads to often correspond to the site of RBP-RNA interaction (with some variability due to the positioning of available crosslinkable amino acids and bases within the binding site) [KZR+10]. When we queried how using only the 5 end of reads could affect splicing map signal and noise, we observed that the effect was heavily dependent on different RBPs. For some RBPs such as U2AF2, we observed that the use of only the 5 end of reads provided significant clarity to the splicing map by resolving binding to specific regions (for example, in U2AFC the intronic 3 splice site region is enriched as opposed part of the alternative exon) (Fig. 6H, Supplemental Fig. S7M). However, for other RBPs such as RBFOX2 we observed that using 5 read ends yielded a similar map, but with dramatically increased noise relative to using whole reads (Supplemental Fig. S7N). Thus, these results suggest that this method can improve resolution for some RBPs (particularly those with highly specific splice site-proximal binding), but that factors with broader crosslinking and binding patterns may suffer unacceptable loss of signal.

3.4 CONCLUSION

The continuing refinement of methods to identify RNA binding protein binding sites enabled large-scale profiling of targets for 350 RBPs by the ENCODE consortium. This compendium of manually quality assessed datasets provided a unique opportunity to develop rigorous quality metrics that can be applied in an automated fashion to assist in determining reliability of eCLIP experiments, both for large-scale efforts as well as individual eCLIP experiments performed by other groups.

3.5 METHODS

3.5.1 Sequencing and data generation for eCLIP

eCLIP experiments were performed as previously described [VPS+16]. Datasets passing manual quality assessment were deposited at the ENCODE Data Coordination Center (http://www.encodeproject.org) (Supplemental Table S2). Datasets that did not pass but were important for this study are deposited at the Gene Expression Omnibus (accession pending).

3.5.2 eCLIP data processing and peak calling

An exact description of the eCLIP processing pipeline can be found on the ENCODE DCC website (https://www.encodeproject.org/documents/dde0b66909094f8b946d3cb9f35a6c52/download/attachment/e Briefly inline barcodes are used to demultiplex datasets, these and unique molecular identifiers (UMIs) are then removed and recorded using custom scripts. Reads are trimmed using cutadapt (version 1.9), and then mapped against a database of repetitive elements (Repbase version 18.05) using STAR (version 2.4). Reads not successfully mapped to repetitive elements are stringently mapped to the human genome (hg19) using STAR. Finally PCR duplicates were removed using a custom PCR duplicate removal script, and clusters were called using CLIPper (version 1.0). Cluster significance was calculated against a paired input dataset, and significant peaks were defined as clusters with a fold-enrichment 8 over input and p-value 0.001.

Repetitive element quantification was performed using a customized pipeline to identify unique mapping to retrotransposable and other multi-copy element families (Van Nostrand et al., in preparation).

Unique fragments were defined as the number of PCR duplicate removed fragments mapping uniquely to the human genome or to a repetitive element family. Unique genomic fragments were defined as the number of PCR duplicate removed fragments uniquely mapping to the human genome. PCR duplication rate was estimated as the fraction of unique fragments to

mapped fragments.

For analysis of untrimmed datasets, reads were demultiplexed exactly as described in the eCLIP standard operating procedure, they were then mapped directly to the human genome without any additional processing steps. Percent mapped was reported via STAR mapping output.

For analyses with repetitive elements kept, adapter trimming was performed as normal. Instead of mapping against (and discarding reads mapping to) the database of repetitive elements, reads were immediately mapped to the human genome, and peak calling were performed as previously described. Input normalization was performed using a paired input dataset that also contained repetitive element reads.

For analyses with PCR duplicates kept, all processing, up to PCR duplicate removal was performed as previously described to obtain mapped genomic fragments. Those fragments were then visualized.

3.5.3 Identification of biologically reproducing peaks by IDR

Peaks reproducing across biological replicates were identified by a modification of the Irreproducible Discovery Rate method [Li11] (Supplemental Fig. S2B). Information content for each peak was calculated as $I = p_i * \log_2(\frac{p_i}{q_i})$, with p_i and q_i represent the fraction of total reads in IP and input respectively that map within peak i. Peaks were then ranked by information content, and processed with the IDR software (version 2.0.2) to identify reproducible regions at an IDR threshold of $<= 0.01$. Next, CLIPper-identified clusters within these reproducible regions were queried against both replicates to calculate fold-enrichment and significance in IP versus input. Clusters were ranked by the geometric mean of fold-enrichment between replicates, and a final set of reproducible peaks was identified as the set of non-overlapping CLIPper-identified peaks meeting standard enrichment cutoffs (fold-enrichment $>=8$ and p-value $<= 0.001$) that were highest ranked by fold-enrichment within each reproducible region.

3.5.4 Estimation of unique fragments with a-eCT

PCR amplification can be modeled as $F = I * e^{CT}$, for the number of final molecules yielded F, the number of initial molecules I, the number of performed PCR cycles CT, and PCR efficiency per amplification cycle e. Library yield for eCLIP was initially described as an estimated CT (eCT) required to obtain 100 femtomoles of library approximating e = 2: $eCT = CT - \log_2(\frac{C_{fm}}{100_{fm}})$ for measured library yield C (in femtomoles)(Van Nostrand et al. 2016). Concentrations, PCR Cycle numbers and eCTs for all experiments used in this study can be found in Supplemental Table S2. Concentrations, PCR cycle numbers and eCTs for control IgG experiments were obtained from (Van Nostrand et al. 2017a).

To obtain an empirical estimate for e, experiments with a PCR duplication rate of >90% were selected. Then a non-linear least squares curve fitting approach was applied to identify the e that minimized the sum of squared error between estimated and observed fragments yielding $\hat{e}$ = 1.84. Accurate-eCT (a-eCT) was the defined as: $aeCT = CT - \log_{1.84}(\frac{C_{fm}}{100_{fm}})$

Calculation of mean squared error was performed by comparing the a-eCT-estimated versus observed number unique fragments in each sequencing experiment.

3.5.5 Peak identification dependence on sequencing depth

The relationship between peak detection and sequencing depth was analyzed by creating pseudo-datasets by subsampling defined fractions of unique genomic fragments. A downsampled series was created by subsampling 10% of reads, and then iteratively adding 10% additional reads until all reads were used. Standard CLIPper cluster identification and input normalized peak identification was performed on each downsampled dataset using the same (fully sequenced) input control. Next, each significantly enriched peak identified in the full dataset was associated with the smallest downsampling fraction it was discovered in. The minimum sequencing depth needed to detect peaks in genes with a given TPM was defined as the number of unique genomic

fragments in the downsampled dataset the peak was first discovered in. K562 and HepG2 total RNA TPM values were obtained from ENCODE project accession numbers ENCFF286GLL and ENCFF533XPJ respectively

To consider the correlation between gene expression and number of reads in peaks for eCLIP, the number of reads in each peak was quantified and normalized by peak length. For each gene only the peak with the largest number of reads was selected as representative, and other peaks not considered, to avoid over-weighting large or highly-bound genes. Linear regression comparing the $log_{10}(TPM)$ of the gene to the $log_{10}(normalizedreadcount)$ of the most highly expressed peak was then calculated.

3.5.6 Motif or region presence near eCLIP peaks

To estimate specificity of peaks identified in the downsample series, motif presence (for RBFOX2) and peak location (for PRPF8) were considered. For RBFOX2, a peak was considered as containing a RBFOX2 motif if it overlapped a GCAUG 5-mer on the same strand as the transcript. For PRPF8, a peak was considered properly located if it overlapped the 3 ends of an exon in GENCODE (v19) (Harrow et al. 2012). For each fraction downsampled, the peaks initially discovered in that fraction were overlapped with the feature of interest to calculate specificity.

To consider sequence conservation within eCLIP peaks, the mean (mammalian) phastcons score was calculated for each peak. Then peaks were grouped by fraction discovered, and the mean conservation of each fraction was calculated. To estimate background conservation, peak locations were shuffled while preserving the number of peaks found in 5 UTRs, CDS, 3 UTRs, exons, proximal introns and distal introns, and the mean conservation was calculated. For all RBFOX2 and PRPF8 analysis, downsampling and peak analysis was performed 10 times.

3.5.7 Peak and information content saturation analysis

For saturation analysis, peaks were taken from the above downsampling series. For each downsampling experiment, the fraction of all peaks or fraction of total information content discovered up until the current fraction was calculated. Where total information content is defined as the sum of the information content across all significant peaks.

Saturation was calculated using the equation $\frac{(N_i - N_{(i-i)}}{Total}$, where N_i is the number of peaks or total information content recovered at downsampling fraction i. Total is the total number of peaks or total information content in the full dataset.

3.5.8 Estimates of required reads

Experiments were defined as saturating when a 10% increase in reads used lead to a $<5\%$ gain in total information. Because peak calling downsampling was performed on unique genomic fragments, the number of unique fragments needed to saturate each dataset was estimated by calculating the number of fragments downsampled if unique fragments had instead been downsampled.

Experiments were rank ordered by read depth when saturated, and the number of unique fragments in the dataset in the 90th percentile was used to estimate the number of unique fragments needed to allow 90% of datasets to saturate.

Estimation of read loss due to adapter trimming and mapping Read loss due to adapter trimming and mapping was modeled by calculating the mean fraction of reads lost during each step. Specifically for mapping the fraction lost was calculated as the fraction of mapped fragments from trimmed fragments.

3.5.9 PCR duplication downsampling

Downsampling was performed as described above, however, pre-PCR duplicate mapped fragments, rather than unique genomic fragments were used. The downsampled files were then removed of PCR duplicates as in the main eCLIP processing pipeline and total reads before and after downsampling were counted.

3.5.10 Poisson Modeling of PCR Saturation

For each dataset the relationship between mapped fragments and unique fragments was modeled as a Poisson distribution. The probably that a single fragment was observed at least once is modeled with the equation $P(\leq 1) = 1 - e^{-\lambda}$. Where λ is the mean number of times a molecule is observed. λ is defined as $\lambda = \frac{M}{E}$. M is the number of mapped fragments and E is the estimated number of unique fragments, as calculated by the a-eCT equation. To estimate the total number of unique fragments (U) from mapped fragments we parameterize the previous equation obtaining $E(1 - e^{(-\frac{M}{E})}) = U$.

3.5.11 Estimates of required reads to sequence

Estimation of number of sequenced reads needed to observe the required number of unique fragments was performed by integrating the three previously described read loss models. The final model to estimate usable reads from input reads is defined as $S = -E * \frac{ln(1-(U/E))}{P(a)P(m)}$. Where S is the number of reads to sequence, E is the number of estimated unique fragments in the total experiment, U desired number of unique fragments, and $P(a)$ and $P(m)$ are the probably of retaining reads after adapter trimming or mapping.

The naive sequencing depth model was calculated as described in the main text. Specifically $U = (1 - P(a))(1 - P(m))(1 - P(d)) * S$, where U is the number of unique fragments, S is the sequenced reads, and $P(a)$, $P(m)$, and $P(d)$ are the probabilities of losing a read due to

adapter trimming, mapping or PCR duplicate removal, respectively.

3.5.12 Usable read and total information content cutoff calculations

An optimal unique fragment cutoff for all experiments was calculated by optimizing the f-score. Because many deeply sequenced datasets failed manual QC, for model fitting, only experiments $<$ 1,000,000 reads uniquely genomic fragments were used. The minimum unique fragment threshold was calculated by maximizing the f-score when setting a pass/fail threshold on the number of unique fragments.

A threshold for total information content was developed first by calculating each experiments total information content as previously described. For training all datasets passing the minimum unique fragment cutoff were used. To identify an optimal total information content cutoff the cutoff was calculated for all values between 0 and the largest total information content in all experiments. The total information content cutoff that maximized the f-score was determined to be the optimal threshold.

Usable read and total information content values for all datasets used to generated these cutoffs can be found in Supplemental Table S3.

3.5.13 Rescue and Self-consistency Ratio

Rescue and Self-consistency Ratios were calculated as in [LMK$^+$12]. To calculate the rescue ratio, unique genomic fragments from both biological replicates were combined, shuffled, and split into two pseudo-biological replicates. Peak calling and input normalization was then performed. Peaks were ranked by information content and IDR was performed to determine reproducible peaks in the pseudo replicates (Np). Additionally IDR was performed on the real replicates and the number of reproducible peaks was counted (Nt). Rescue Ratio was calculated as $\frac{max(N_p,N_t)}{min(N_p,N_t)}$.

To calculate self-consistency ratio, uniquely mapped reads from each replicate were independently split into two sub replicates. Peak calling and input normalization were then performed on each set of reads. Within each replicate IDR was performed on information content ranked peaks to determine a reproducible set of peaks. The number of reproducible peaks was then counted for replicate 1 (N1) and replicate 2 (N2). Self-consistancy ratio was calculated as $\frac{max(N_1,N_2)}{min(N_1,N_2)}$

Self-consistency ratio and rescue ratio values for all datasets used in this study can be found in Supplemental Table S3.

3.5.14 Identification of Alternatively Spliced Events

rMATS (version 3.2.1.beta) JunctionCountsOnly files were used to identify alternatively spliced (AS) events. Significant AS events were defined as having a p-value > 0.05, FDR > 0.1 and > 0.05. Elimination of redundant splicing events was performed by identifying groups of overlapping AS events and selecting the event with the highest inclusion junction count (IJC) among the overlapped events using the bedtools (v2.26) command merge (-o collapse -c 4) and pybedtools (v0.7.9). As a background control 2555 (HepG2) and 3148 (K562) unchanging cassette exons, with $0.1 < < 0.9$ in at least half of control RNA-seq datasets were selected.

3.5.15 Generation of splicing maps

RBP splicing maps were generated using one of four different methods. For every method all information for significant splicing was extracted in windows near the exon/intron boundary; maximally 50nt into each exon and maximally 300nt into each intron. For shorter exons ($<$100 nt) and introns ($<$600nt), information was only counted until the boundary of the neighboring feature.

For peak binding splicing maps the number of peaks overlapping an AS event were simply

summed. IP and input density splicing maps were calculated by computing the mean of the eCLIP normalized (reads per million) read densities over each AS event.

For the background subtraction approach for each splicing event, read densities were first normalized across all regions separately for IP and input, in order to equally weigh each event. Per-position input probability densities were then subtracted from IP probability densities to attain position-level enrichment or depletion.

For information content approach each splicing events read densities were normalized by dividing the read depth by the total number of reads in a sample separately for IP and input. Per-position entropy probabilities were calculated using the equation $p(ip) * log_2(\frac{p(ip)}{p(input)})$ where $p(ip)$ and $p(input)$ are the per-position read probabilities at a given base.

To calculate mean normalized enrichment scores for background subtraction and information content methods the mean score for each base was first calculated. Then each individual event score at each base was normalized by the mean at each base.

3.5.16 Outlier removal

Outliers were moved on a base by base basis. For each base the highest (2.5%) and lowest (2.5%) values were removed, and results were visualized

3.5.17 Alternative splicing map approaches

For peak coverage splicing maps the per-base peak coverage of each significant AS event (included or excluded upon knockdown) was summed and reported.

5 end splicing maps were generated as background subtracted, outlier removed splicing maps, with the exception that full read densities were replaced with just the read density at the 5 end of the R2 read for each read near an AS event.

3.5.18 DATA ACCESS

Data passing ENCODE quality standards are available on the ENCODE DCC (encodedcc.org). Data not passing quality standards have been submitted to GEO (accession pending). All datasets used in this study and their locations are listed in Supplemental Table S2.

3.6 DISCLOSURE DECLARATION

E.L.V.N. and G.W.Y. are co-founders and consultants for Eclipse BioInnovations Inc. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies.

3.7 ACKNOWLEDGMENTS

The authors would like to thank Cricket Sloan, Jean Davidson, Eurie Hong, and Mike Cherry for assistance with data deposition and distribution to the public. This work was funded by the National Human Genome Research Institute ENCODE Project, contract U54HG007005, to GWY. ELVN is a Merck Fellow of the Damon Runyon Cancer Research Foundation [DRG-2172-13] and is supported by a K99 grant from the NIH [HG009530]. GAP is supported by the NSF graduate research fellowship. GWY was partially supported by grants from the NIH (HG007005, NS075449).

Chapter 3, in full, has been submitted for publication of the material as it may appear in Nucleic Acids Research, 2018. Gabriel A. Pratt, Eric L. Van Nostrand, Brian A. Yee, Alain Domissy, Steven M. Blue, Chelsea Gelboin-Burkhart, Thai B. Nguyen, Ines Rabano, Ruth Wang, Balaji Sundararaman, Keri Garcia, Rebecca Stanton, Gene W. Yeo. Nucleic Acids Research, 2018. The dissertation/thesis author was the primary investigator and author of this paper.

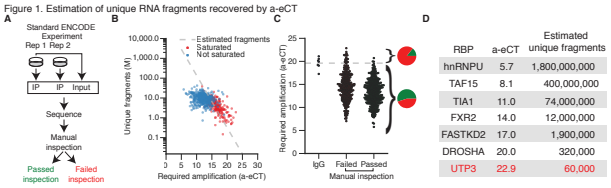

Figure 3.1: **Figure 1. Estimation of unique RNA fragments recovered by a-eCT** (A) Schematic of standard ENCODE experiment including manual quality control. (B) Scatter plot indicates accurate-eCT (a-eCT) (see Methods) versus unique fragments observed (including non-PCR duplicate reads mapped either to unique genomic loci or repetitive elements, in millions of reads mapped) for all ENCODE eCLIP experiments. Non-saturated (<60% PCR duplicates) datasets are indicated in blue, and saturated (>60% PCR duplicates) datasets are indicated in red. Dashed line indicates the number of unique molecules expected based on a-eCT. (C) Points indicate the a-eCT value of all ENCODE eCLIP experiments, separated into IgG controls (blue), datasets that failed manual quality assessment (red) and datasets passing manual assessment (green). Dotted line indicates average a-eCT of IgG control experiments (19.6). (D) Representative RBPs are listed along with their a-eCT and corresponding estimate of the number of unique RNA molecules isolated in eCLIP. UTP3 (in red) did not pass quality control metrics.

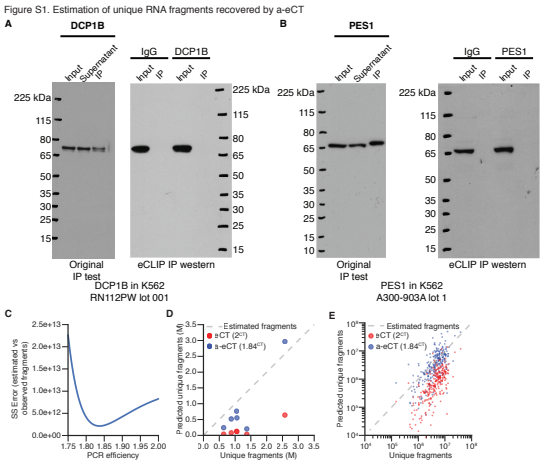

Figure 3.2: **Supplementary Figure 1. Estimation of unique RNA fragments recovered by a-eCT** (A-B) Example immunoprecipitation (IP) western blots for proteins (A) DCP1B and (B) PES1 with successful IP using simplified conditions that exclude enzymatic library preparation steps, but failed to IP during full eCLIP. (C) Plot indicates sum of squared error for varying PCR efficiency when comparing true observed number of unique molecules to estimated number of unique molecules for six highly saturated (>90% PCR duplicated) experiments. (D-E) Scatter plot of estimated unique molecules at two estimates of PCR efficiency, 2 (red) and 1.84 (blue) versus unique fragments obtained after sequencing. Shown are (D) six highly saturated (>90% PCR duplicated) experiments, or (E) 255 moderately saturated experiments (>60% PCR duplicated).

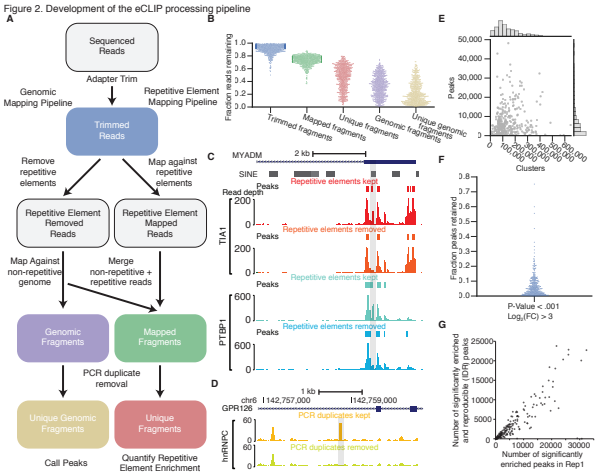

Figure 3.3: **Figure 2. Development of the eCLIP processing pipeline**(A) Schematic of eCLIP processing for both unique genomic mapping and repetitive element mapping. (B) Points indicate the fraction of reads remaining after trimming (blue), mapping to either repetitive elements or unique segments of the genome (green), PCR duplicate removal on both repetitive elements and unique elements (red), mapping to only unique genomic elements (purple), and PCR duplicate removal on only unique genomic elements (yellow) for all ENCODE eCLIP experiments. (C) Effect of repetitive element masking on the gene MYADM. Tracks indicate raw read number for TIA1 (HepG2) and PTBP1 (HepG2) eCLIP datasets. Significant (fold-enrichment 8 and p-value 0.001) peaks are shown. SINE elements are derived from the RepeatMasker UCSC Genome browser track. The highlighted area indicates a peak overlapping a SINE element that is no longer detected after repetitive element removal. (D) Tracks indicate HNRNPC eCLIP read density in HepG2 in GPR126 for (top) without and (bottom) with PCR duplicate removal. Highlighted region represents skyscraper that is lost during PCR duplicate removal. (E) Scatter plot of clusters versus significant peaks remaining after filtering (fold-enrichment 8 and p-value 0.001 in IP versus input). Histograms indicate cluster or peak distribution respectively. (F) Points indicate fraction of clusters meeting the above thresholds for p-value and fold-enrichment. (G) Plot indicates the number of significantly enriched peaks in replicate 1 of each eCLIP experiment versus the number of enriched and reproducible peaks in the total experiment after IDR analysis.

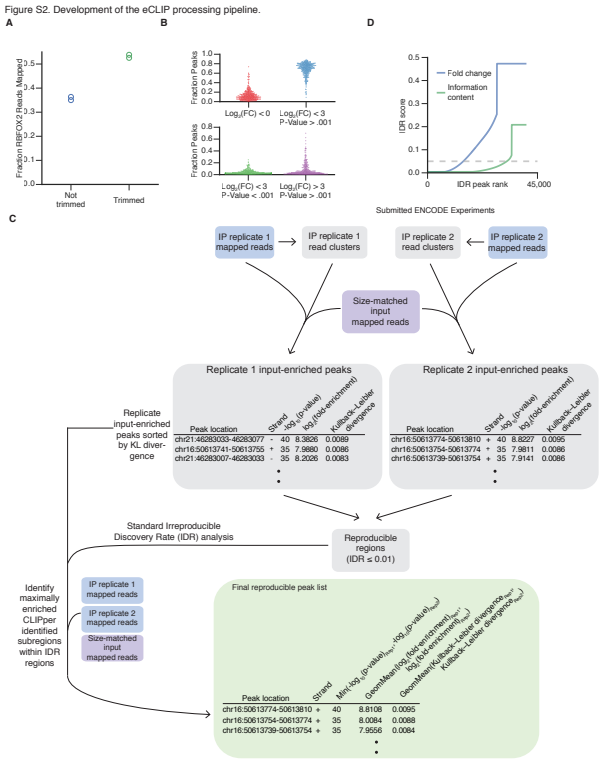

Figure 3.4: **Supplementary Figure 2. Development of the eCLIP processing pipeline** (A) Circles indicate fraction of reads effectively mapped in HepG2 RBFOX2 experiments if adapter trimming is not performed (blue), or is correctly performed (green) (B) Points indicate fraction of clusters removed due to specific filtering criteria, generally not enriched above input (red), enriched, but fail to meet p-value and fold change cutoff (blue), fails to meet only fold change cutoff (green) and fails to meet only p-value cutoff (purple) (C) Schematic of adaption of Irreproducible Discovery Rate (IDR) analysis to identification of reproducible eCLIP peaks. First, input-normalized clusters are identified separately for two biological replicates. Next, these peaks are ranked by relative information content, defined as $I_i = p_i * \log_2(\frac{p_i}{q_i})$, for proportion of IP reads within peak i represented by pi and fraction of input reads within the peak as qi. Next, standard IDR analysis is performed on the ranked peak lists to identify reproducible regions at IDR cutoff of 0.01. Next, we considered all CLIPper-identified subregions within these IDR regions, and calculated the fold-enrichment in IP versus input for each subregion in each replicate. Subregions were ranked by the geometric mean of fold-enrichment between the two replicates, and the set of non-overlapping subregions that were significantly enriched ($p<0.001$ in both replicates) with geometric mean of fold-enrichment 8 in both replicates were obtained as the set of reproducible peaks. (D) Plot indicates each peak ranked by IDR score, when IDR score is calculated by ranking peaks based on (blue) fold-enrichment above input or (green) information content.

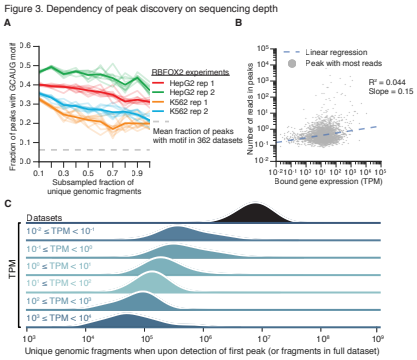

Figure 3.5: **Figure 3. Dependency of peak discovery on sequencing depth**(A) Plot of fraction of peaks containing a GCAUG motif for peaks identified in a series of subsamples of eCLIP unique genomic fragments for RBFOX2 in HepG2 and K562. Shown are RBFOX2 HepG2 replicate 1 (red) and replicate 2 (green), and RBFOX2 K562 replicate 1 (orange) and replicate 2 (blue). The dashed grey line indicates the mean fraction of GCAUG-containing peaks observed across all released eCLIP datasets. (B) One point for each gene indicates the TPM (Transcripts Per Million reads) of the gene (x-axis) and the number of reads (normalized by peak size) in the peak with the highest number of reads for RBFOX2 in HepG2. Dashed line indicates simple linear regression. (C) Joy plot indicates (top) the distribution of unique genomic fragment values for released ENCODE eCLIP experiments, versus (bottom) the distribution of total eCLIP unique genomic fragments in the downsampled subsample where the first peak was identified in each gene, separated into

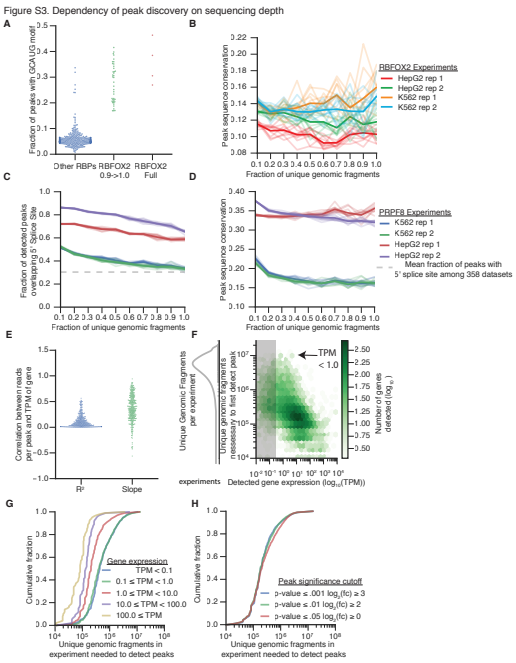

Figure 3.6: **Supplementary Figure 3. Dependency of peak discovery on sequencing depth** (A) Points indicate fraction of significant peaks detected in between the 90% to 100% subsampled fraction that contain a GCAUG motif for all RBPs except RBFOX2 (blue) versus RBFOX2 (green) or fraction of all RBFOX2 peaks that contain GCAUG motif (red) (B) Plot indicates mean mammalian phastons conservation for all peaks newly discovered in each downsampled subsample for RBFOX2 eCLIP in HepG2 replicate 1 (red) and replicate 2 (green) and K562 replicate 1 (orange) and replicate 2 (blue). (C-D) Downsampling analysis for PRPF8 eCLIP in HepG2 replicate 1 (blue) and replicate 2 (green), and K562 replicate 1 (red) and replicate 2 (purple). (C) Plot indicates the fraction of peaks newly discovered in each downsampled subsample that overlap the 5 splice site. The dashed grey line indicates mean fraction overlap with 5 splice sites for all 179 non-PRPF8 released ENCODE datasets. (D) Lines indicate the average conservation for newly discovered peaks at the indicated downsampling fraction. (E) Points indicate the R2 (blue) and slope (green) for the linear regression between gene TPM and maximum peak read density for all released ENCOE eCLIP experiments. Each point represents an individual dataset as shown in Fig. 3B. (F) Density plot indicates the log10 number of reads in experiment (y-axis) when a gene with a specific log10 TPM is first detected to contain a peak (x-axis) for all released ENCODE eCLIP experiments. Green shading indicates the number of genes detected for each bin of read depth and TPM combination. The shaded region indicates number of reads needed to detect peaks when gene TPM < 1. (left) Curve indicates the number of unique genomic fragments per released ENCODE eCLIP experiment. (G) Cumulative distribution function plot indicates the number of reads needed to first detect peaks for the set of genes in indicated bins separated by gene TPM: TPM $< .01$ (blue), $.01$ TPM < 1.0 (green), 1.0 TPM < 10.0 (red), 10.0 TPM < 100.0 (purple), 100.0 TPM (gold). (H) Plots indicate the cumulative fraction of genes with peaks discovered at given experimental sequencing depth, for the indicated cutoffs for peak enrichment in IP versus input (p-value .001 and fold-enrichment 8 (blue), p-value .01 and fold-enrichment 4 (green), p-value .05 and fold-enrichment 0 (red).

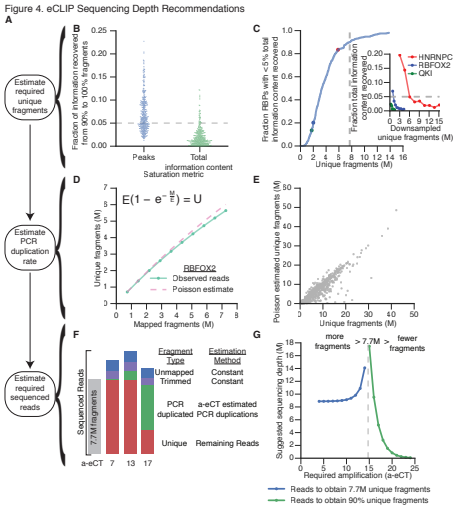

Figure 3.7: **Figure 4. eCLIP Sequencing Depth Recommendations** (A) Schematic of method used to estimate number of suggested reads for sequencing based on a-eCT calculation. (B) Points indicate saturation rate for peak or total information content between the 90% retained and 100% (full dataset) subsampled fraction for all submitted ENCODE eCLIP experiments. Grey dashed line is 5% saturation cutoff. (C) (right) Lines show percent of additional information recovered when adding 10% additional reads for hnRNPC (red), RBFOX2 (blue), and QKI (green) in HepG2. Dotted line indicates the saturation point at which less that 5% additional information is gained. (left) Cumulative fraction plot indicates the distribution of unique fragments when each eCLIP dataset reaches saturation. Colored points indicate depth of sequencing when hnRNPC, RBFOX2 and QKI saturate. (D) Plot indicates the relationship between mapped reads pre-PCR duplicate removal (x-axis) or unique fragments remaining post-PCR duplicate removal (y-axis) for actual downsampling experiments (green) or modeled using a Poisson estimate (red) for RBFOX2 in HepG2. (E) Scatter plot indicates observed versus estimated usable reads using Poisson model for all eCLIP datasets. (F) Schematic indicates how datasets with varying a-eCT values can be modeled to estimate required sequencing depth to obtain 7.7 million unique molecules. (G) Plot of a-eCT versus estimated number of input reads needed to obtain 7.7M unique molecules. For datasets with less than 7.7M estimated unique fragments (a-eCT $>$ 14.7), plot indicates the estimated number of reads to observe 90% of total unique molecules (green).

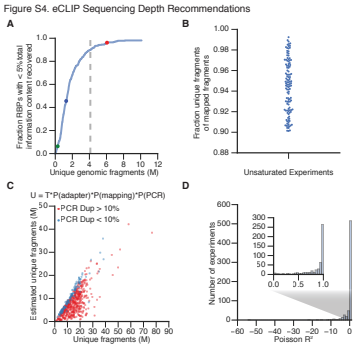

Figure 3.8: **Supplementary Figure 4. eCLIP Sequencing Depth Recommendations** (A) Plot indicates the distribution of unique genomic fragments when each eCLIP dataset reaches saturation. Colored points indicate depth of sequencing when hnRNPC, RBFOX2 and QKI saturate. (B) Plot of fraction of unique fragments remaining after PCR duplicate removal is performed on mapped fragments in datasets with PCR duplicate rate of $<$10% (C) Plot indicates (x-axis) the actual number of unique fragments versus (y-axis) the estimated number of unique reads using a nave read estimation model with static fragment loss estimates for adapter trimming, mapping, and PCR duplicate removal. Shown are non-saturated (PCR duplication rate $<$ 10%, blue) and all other datasets (red). (D) Histogram indicates R2 fits for all experiments using Poisson PCR duplicate read loss model (inset focuses R2 between 0 and 1).

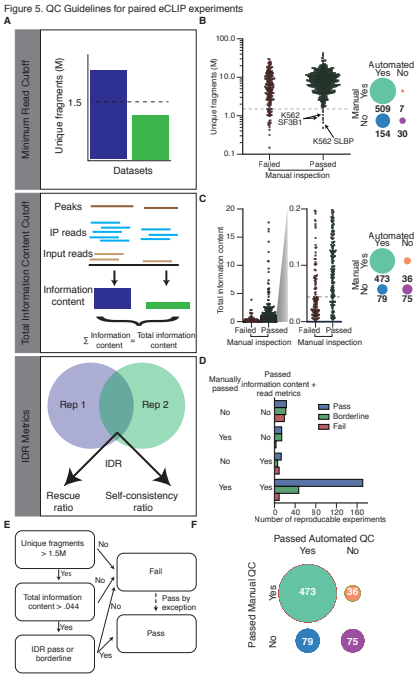

Figure 3.9: **Figure 5. QC Guidelines for paired eCLIP experiments** (A) Schematic of eCLIP data quality standards. (B) Swarm plot indicates number of unique fragments observed in each eCLIP dataset separated by failing (red) or passing (green) manual quality inspection. Dashed line indicates 1.5 million read quality threshold that maximizes predictive power on manual classification (Supplemental Fig. 5A), and inset indicates confusion matrix for this threshold versus manual inspection. Three datasets judged to be high quality despite low unique fragment number are indicated. (C) Similar swarm plot indicates the total information content across all peaks in each eCLIP dataset that passes the unique fragment threshold in (B), separated by failing (red) or passing (green) manual quality inspection. Dashed line indicates the information content threshold that maximizes predictive power on manual classification (Supplemental Fig. 5C), and inset indicates confusion matrix for this threshold versus manual inspection. (D) Bar chart indicates the count of all ENCODE experiments that pass or fail manual or automated QC approaches, broken into three groups based on their IDR thresholding metric status: passed (blue), borderline (green), and failed (red). (E) Schematic detailing final recommended quality assessment decision flowchart. (F) Confusion matrix of final classification scheme versus manual quality assessment.

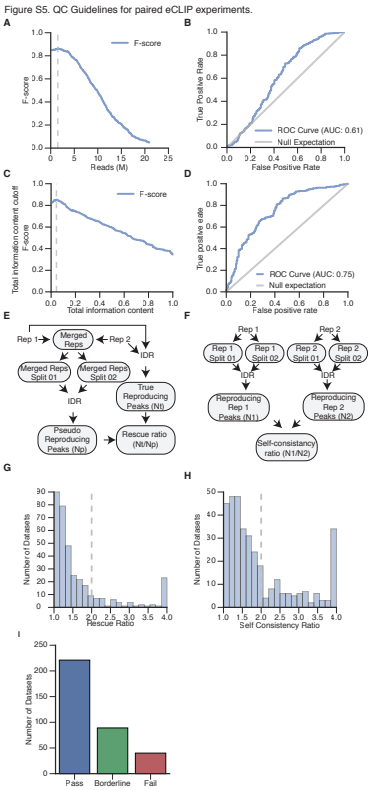

Figure 3.10: **Supplementary Figure 5. QC Guidelines for paired eCLIP experiments** (A) Plot indicates f-score for classification of datasets relative to manual quality assessment based on unique fragments present. Maximal classification of datasets was obtained at a cutoff of 1.5 million unique fragments. (B) ROC curve for classifying datasets based upon varying minimum unique fragment thresholds. (C) Plot indicates f-score for classification of datasets based on the total information content in all significantly enriched peaks. Only datasets passing the unique fragment cutoff in (A) were considered. (D) ROC curve for classifying datasets based upon varying total information in peak cutoff. (E) Schematic for calculation of Irreproducible Discovery Rate rescue ratio. (F) Schematic for calculation of IDR self-consistency ratio. (G) Bar plot indicates rescue ratio for all ENCODE eCLIP experiments. (H) Bar plot indicates self-consistency ratio for all ENCODE eCLIP experiments. Dashed line indicates a cutoff of 2 previously used for ChIP-seq analysis. (I) Bars indicate the number of ENCODE eCLIP experiments that either pass both rescue ratio and self-consistency ratio (pass, blue), passed just one of the two tests (borderline, green) or failed both tests (fail, red).

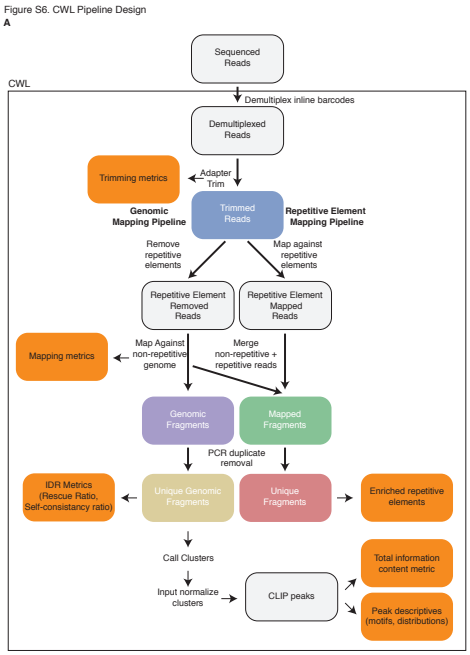

Figure 3.11: **Supplementary Figure 5. CWL Pipeline Design** Schematic of eCLIP CWL processing pipeline

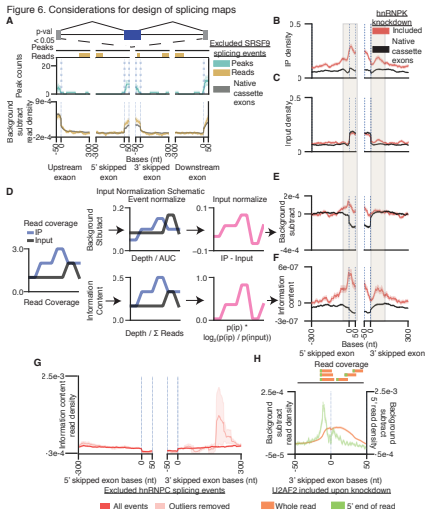

Figure 3.12: **Figure 6. Considerations for design of splicing maps** (A) (top) Schematic of cassette exon splicing. (bottom) Comparison of splicing maps generated for SRSF9 (HepG2) knockdown-excluded cassette exons based on peak (blue), background subtracted normalized read (gold) or non-changing background read (black) density. Bars above plot indicate areas of significant enrichment (by Mann-Whitney U test) above background when either peaks or reads are used. (B) Splicing map for HNRNPK (HepG2) binding at knockdown-included (red) or native (black) events generated from read density in IP. The region shown includes 300 nucleotides of intronic sequence upstream or downstream of the skipped exon, and 50bp of exonic sequence on both the 5 and 3 sides. Boxes indicate regions discussed in the text. (C) As in (B), but showing read density in input. (D) Schematic shows calculation of background subtraction and entropy normalization methods. (top) For background subtraction, read coverage is first normalized within each event by dividing each position by the total read coverage. Input signal is then subtracted from IP signal to create a map for each event. Final splicing maps represent the mean across all events. (bottom) Read density within events is first normalized for overall sequencing depth. Then relative information is calculated at each position to yield an event-level map. All events are then averaged to yield the final splicing map. (E) Similar to (B), plot shows the background subtraction splicing map for hnRNPK. (F) Similar to (B), plot shows the information content splicing map for hnRNPK. (G) Effect of removing outliers (defined as the 2.5% of the top and bottom most enriched elements at each position). Plot indicates mean information content normalized density before (dark red) and after (light red) removing outliers. (H) (top) Schematic of whole read (orange) versus 5 end (green) read coverage. (bottom) plot indicates U2F2 eCLIP mean background subtracted normalized read density at cassette exon 3 splice sites for U2AF2 knockdown-included events in HepG2.

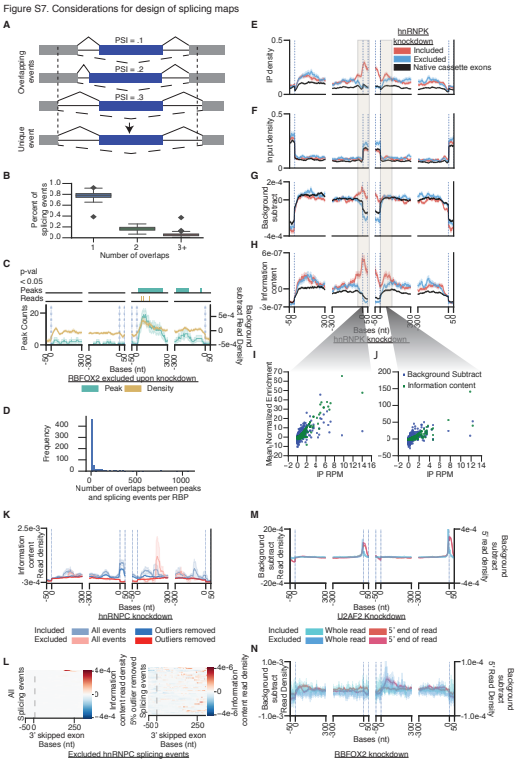

Figure 3.13: **Supplementary Figure 6. Considerations for design of splicing maps** (A) Schematic of overlapping splicing event removal. For all events that overlap with their upstream and downstream exon, the event with the largest percent spliced in (PSI) is chosen as the canonical event. (B) Box plot shows the fraction of events per knockdown experiment removed, based on (x-axis) the number of events within an overlapping window as shown in (A). (C) Comparison of splicing maps generated for RBFOX2 knockdown-excluded cassette exons based on peak density (gold) and background subtracted normalized read density (blue) in HepG2. (D) Histogram indicates the distribution of the number of peaks overlapping significant splicing events for each RBP with both ENCODE eCLIP and knockdown RNA-seq datasets. (E-J) Cassette exon splicing maps for RBP hnRNPK in HepG2, showing 50 exonic and 300 intronic nucleotides for the upstream exon 5 splice site, cassette exon 3 and 5 splice sites, and downstream exon 3 splice site. Binding around knockdown-excluded (blue), included (red), and native (black) events are shown. (E) Lines indicate read density in IP. (F) Read density in input. (G) Background subtraction splicing map. (H) Information content splicing map. (I-J) Splicing maps indicate mean centered normalized background subtraction and information content scores as compared to IP RPM counts for each significant splicing event at two specific locations: (I) 8bp upstream of the 3 splice site of the cassette exon, and (J) 42bp downstream of the 5 splice site of the cassette exon. (K) Full version of Fig. 6H, showing average information content normalized read density before outlier removal at hnRNPC knockdown-excluded (light red) and included (light blue) events and after performing outlier removal for excluded (dark red) and included (dark blue) events in HepG2. (L) Heatmaps show information content normalized read density in the 3 section of cassette exon events for hnRNPC (HepG2) before (left) and after (right) outlier removal. (M-N) Comparison of whole-read based inclusion (light blue) and exclusion (dark blue) splicing maps versus 5 end-only based inclusion (light red) and exclusion (dark red) background subtraction normalized maps for (M) U2AF2 (HepG2) and (N) RBFOX2 (HepG2).

Chapter 4

Insights gained from individual CLIP-seq experiments

4.1 ABSTRACT

Analysis of a single RBP, or comparison of binding patterns for mutant and wildtype RBPs can lead to novel biological insights. In this chapter I present two short vignettes, both part of larger stories published elsewhere that illustrate the information gained when computational analysis on individual RNA binding proteins is performed. In this chapter I briefly frame the biological question we sought to answer when analyzing two RBPs, UPF1 and MSI2, describe the computational analysis and summarize how these results contributed to the overall study.

4.2 INTRODUCTION

ts

4.2.1 UPF1

The dynamic interaction of RNA binding proteins (RBPs) with RNA is critical to every aspect of RNA metabolism[Moo05]. An important question yet to be fully addressed is how RNA regulators faithfully distinguish their target RNAs from the large complement of non-targets in the cell. Most models for RBP-target specificity invoke RBP affinity for target-specific RNA sequences, structures or bound proteins[ÄN12, GBYD08]. However, for RNA quality control pathways, which detect and destroy faulty or non-functional RNAs, target-specific mechanisms for RBP recruitment are harder to envision, as these aberrant RNAs have the potential to differ widely in sequence composition and associated proteins[VW11, PL13].

The first mRNA quality control pathway discovered was nonsense-mediated decay (NMD)[LPJC91, LL79, MKRR81, PA93]. This translation-dependent pathway is conserved among eukaryotes and degrades transcripts on which translation is halted by termination codons recognized as premature. In this way, NMD prevents accumulation of truncated polypeptides arising from aberrant mRNAs bearing premature termination codons (PTCs) and also serves as a potent gene suppression mechanism for select naturally occurring mRNAs, impacting the expression of up to 10% of protein-coding genes in diverse eukaryotes[SRZ$^+$13].

The key RNA binding regulator in NMD is Upf1, a helicase belonging to the SF1 family of DNA and RNA helicases[FWGJ10]. Degradation of target mRNAs involves assembly of Upf1 with other NMD protein factors including Upf2 and Upf3 and, in most eukaryotes studied to date, the kinase Smg1 and one or more Smg5-7 proteins[KJ12, SRZ$^+$13]. In humans, Smg1 phosphorylates Upf1 in a manner stimulated by Upf2, Upf3 and the exon junction complex[KYI$^+$06], which promotes association of phospho-binding proteins Smg5, Smg7 and mRNA decapping and deadenylation machinery[CBS$^+$14, CHC$^+$13, LJI13, OKYK$^+$12]. Smg6 is itself an endonuclease, which has both phospho-dependent and -independent interactions with Upf1 [CBS$^+$14, ELAMJ09, NJK$^+$14, OKYK$^+$12]. In addition, the ATPase activity of Upf1 itself has been implicated in a late step of target mRNA degradation to remodel the mRNP for enhanced

nuclease access[FSLA10].

Despite the wealth of information regarding the processes involved in NMD target degradation, a fundamental, and yet poorly understood, aspect of the NMD pathway is what enables the central NMD factor Upf1 to distinguish target mRNAs from non-targets in the first place. NMD targets that have been well-studied share in common the fact that translation termination occurs at an unusual position in the mRNA distal to the poly(A) tail due to an extended 3UTR or with an exon-exon junction located downstream of termination[SRZ$^+$13]. One model for NMD target recognition proposes that stalled or aberrant termination complexes recruit Upf1[KJ12, SRZ$^+$13]. Indeed, ribosome toe-printing assays in S. cerevisiae extracts and rabbit reticulocyte lysates revealed that ribosome dissociation at an NMD-inducing PTC compared to a normal termination codon (NTC) is stalled aberrantly[AGK$^+$04, PIB$^+$12] and interactions between Upf1 and the ribosome release factors, eRF1 and eRF3, have been found in both yeast and human cells[CREP$^+$98, IGK$^+$08, KYI$^+$06, SRLA08]. Though the determinants leading to aberrant termination at a PTC have yet to be fully elucidated, it has been suggested that the absence of proximal mRNP factors that promote normal termination, such as poly(A) binding protein, underlies the difference between NMD targets and non-targets[CBC$^+$02, IGK$^+$08, KJ12, UiHI$^+$02].

Recent reports cloud the simple view that aberrantly stalled translation termination complexes account fully for Upf1 recruitment to mRNAs. For example, Upf1 was found to associate with both target and non-target mRNAs even in the absence of translation, and the manner and degree to which translation affects Upf1-mRNA accumulation appears to vary among different mRNAs [HG10, KM13, KLH$^+$14, ZGZM13]. Additionally, genome-wide crosslinking-immunoprecipitation studies have revealed that translation inhibitors induce a shift in Upf1 distribution across mRNAs, from a 3 untranslated region bias to increased association with protein-coding regions [GSM$^+$14, HRB13, ZGZM13], suggesting that translation influences where Upf1 associates along the length of an mRNA, in addition to how strongly it associates with NMD target and non-target mRNAs overall.

Thus, the mechanisms involved in Upf1 discrimination of NMD target from non-target mRNAs have remained unclear. Here we present evidence that Upf1 ATPase activity is required for NMD target selection. In the absence of Upf1 ATPase activity, Upf1-mRNA selectivity is disrupted and NMD complexes accumulate indiscriminately on target and non-target mRNAs.

4.2.2 MSI2

Umbilical cord blood (CB)-derived hematopoietic stem cells (HSCs) are essential in many life saving regenerative therapies, but their low number in CB units has significantly restricted their clinical use despite the advantages they provide during transplantation[MKE13]. Select small molecules that enhance hematopoietic stem and progenitor cell (HSPC) expansion in culture have been identified[BWR$^+$10, FCG$^+$14], however, in many cases their mechanisms of action or the nature of the pathways they impinge on are poorly understood. A greater understanding of the molecular pathways that underpin the unique human HSC self-renewal program will facilitate the development of targeted strategies that expand these critical cell types for regenerative therapies. Whereas transcription factor networks have been shown to influence the self-renewal and lineage decisions of human HSCs[NSL$^+$11, LDZ$^+$13], the post-transcriptional mechanisms guiding HSC fate have not been closely Users investigated. By performing a global analysis of MSI2-RNA interactions, we determined that MSI2 directly attenuates aryl hydrocarbon receptor (AHR) signaling through post-transcriptional downregulation of canonical AHR pathway components in CB HSPCs. Our study provides new mechanistic insight into RBP-controlled RNA networks that underlie the self-renewal process and give evidence that manipulating such networks ex vivo can provide a novel means to enhance the regenerative potential of human HSCs.

4.3 RESULTS

4.3.1 Upf1-mRNA selectivity is lost on a transcriptome-wide level in Upf1 ATP-binding and ATP-hydrolysis mutants

To examine the contribution of Upf1 ATPase activity to mRNA selectivity among endogenous mRNAs, we next turned to a global approach. We employed RIP-seq (RNA immunoprecipitation followed by strand-specific high-throughput sequencing) with Flag-tagged Upf1 WT, DEAA and KA expressed at endogenous levels (Figure 1 and S1A) to query the enrichment in IPs over inputs for endogenous RNAs in comparison to a parental cell line expressing no exogenous Upf1 used as a negative control. RIP-seq libraries were sequenced to a mean depth of 23 million reads and approximately twenty thousand genes had $>$0.1 RPKM per library. Significantly, Upf1 WT, DEAA and KA RIPs were all enriched for transcripts annotated as protein-coding with a smaller fraction derived from pseudogenes, indicating that the ATPase mutations do not disrupt Upf1 transcript specificity for mRNAs as a class.

Using the background recovery of RNAs in the negative control IPs to establish a 5% false-discovery rate (Figure S1B), a distinct population of 2,040 Upf1-associated RNAs was identified as enriched by at least 2-fold over input levels (Figure 1A, Upf1-enriched genes indicated in red; Figure 1D). Based on observations by others [HG10, KLH$^+$14, ZGZM13], these RNAs likely include a mix of NMD sensitive mRNAs and mRNAs that are less sensitive to NMD but limited in downstream steps of the NMD pathway. In striking contrast to WT Upf1, RIPs for Upf1 DEAA and KA did not show enrichment for any RNAs when subjected to the same FDR cutoff (Figure S1B), and, accordingly, the population of WT Upf1-enriched RNAs was not enriched in Upf1 DEAA and KA RIPs over a WT Upf1-non-enriched RNA population defined by 0.97- to 1.03-fold enrichment in WT Upf1 RIPs over inputs (Figures 1A-C, compare red and blue; Figure 1D). These global findings generalize our observations for individual mRNA reporters to the human transcriptome, supporting the conclusion that RNA selectivity is lost in Upf1 mutants deficient in

ATP binding or hydrolysis, despite their preserved specificity in associating with mRNAs as an RNA class.

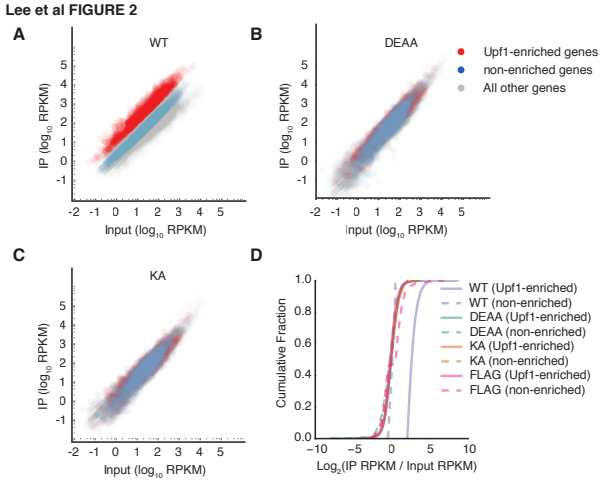

Figure 4.1: **Figure 1. Selectivity in mRNA association is lost on a transcriptome-wide level in Upf1 ATP binding- and ATP hydrolysis-deficient mutants** (A-C) Scatter plots of reads per kilobase transcript per million mapped reads (RPKM) from RNA-seq of input samples versus IPs for Flag-Upf1 WT (A), DEAA (B) and KA (C). Genes with IP/input ratios for WT Upf1 of greater than 2.05 (cut-off based on 5% false discovery rate (FDR) established by comparison to negative control cells expressing Flag epitope only, Figure S2) are shown in red (Upf1-enriched), while genes with log2 (IP/input) between -0.5 and +0.5 are shown in blue (non-enriched). All remaining genes are shown in grey. (D) Cumulative fraction of Upf1-enriched and non-enriched genes with IP enrichment represented as log2 (IP RPKM/input lysate RPKM) for Flag-Upf1 WT, KA, and DEAA, along with Flag only. Difference between WT Upf1 (Upf1-enriched) curve compared to all other curves was statistically significant (p-value <0.05 for all comparisons, KS-test).

4.3.2 ATP binding- and ATPase-deficient Upf1 accumulate on mRNA 3UTRs and are enriched near termination codons and 3 ends

Our observations suggest that Upf1 ATPase activity is required for preventing Upf1 from accumulating and promoting NMD complex formation on translated non-target mRNAs. Recent studies employing Upf1 UV cross-linking IPs followed by high throughput sequencing (CLIP-seq) have reported that while Upf1 cross-links can be found all along the length of mRNAs, the overall distribution of binding sites has a distinct 3UTR bias [GSM$^+$14, HRB13, ZGZM13]. To gain

insight into how binding to mRNA of Upf1 disrupted in the ATPase cycle might differ from WT Upf1, we performed UV CLIP-seq on Flag-Upf1 WT, DEAA and KA expressed at near endogenous levels (Figures S2A, B, C).

Notably, the binding site distributions of Upf1 DEAA and KA were nearly identical to each other and exhibited preference for mRNAs as a transcript class, with a 3UTR bias similar to WT Upf1 (Figures 2A, and S2D, E). However, Upf1 mutant binding was strikingly shifted in comparison to WT Upf1 towards greater binding in 3UTRs on average (Figure 2A; compare read densities in 3UTRs for DEAA, KA with WT) and across all genes, as seen by the increased fraction of 3UTR derived reads per mRNA normalized to length in Upf1 DEAA and KA CLIPs compared to WT Upf1 CLIP (Figure 2B, left graph, see right-shifted curves for DEAA, KA compared to WT; p-value < 0.05, KS test). Examination of read distributions for the WT Upf1-enriched and non-enriched RNA subpopulations identified by RIP-seq (Figure 2) revealed that a greater fraction of the Upf1 enriched mRNAs had a stronger 3UTR bias in WT Upf1 distribution than the non-enriched mRNAs, while the two sets of mRNAs were indistinguishable in their enhanced 3UTR distribution bias for both Upf1 mutants (Figure 2B, right graph, compare dashed and solid lines). These findings suggest that Upf1 ATPase activity is needed to limit Upf1 association preferentially with 3UTRs, and particularly so for the mRNA subpopulation that is less highly bound by WT Upf1 at steady state.

A closer examination of CLIP-seq read density at nucleotide resolution around translation termination codons and at mRNA 3 ends revealed two peaks that are stronger for Upf1 DEAA and KA than WT Upf1. The first centers around 45 nucleotides downstream of the termination codon (Figure 2C, left), while the second peak was observed upstream of transcript 3 ends (Figure 2C, right). This difference in binding site distribution between Upf1 mutants and WT Upf1 in regions proximal to the termination codon and the poly(A) tail was observed of both the WT Upf1-enriched and non-enriched mRNAs (Figure S2F). These observations suggest that the Upf1 ATPase cycle plays a particularly critical role in limiting accumulation of Upf1 at these specific

sites in the 3UTR.

4.3.3 MSI2 Global Analysis

To identify key RNA targets that underlie MSI2 function, we analysed global MSI2 protein-RNA interactions using cross-linking immunoprecipitation followed by sequencing (CLIPseq)[YCL$^+$09] (Figure S3a, b). Replicates were highly correlated via gene RPKMs (reads per kilobase of transcript per million mapped reads) and 5,552 protein-coding genes were bound in both replicates (Figure S3c and Figure 3a, b). Within the top 40% of reproducible clusters, MSI2 bound predominantly to the 3 untranslated regions (3UTRs) of mature mRNAs (Figure 3c). Importantly, 9% of annotated protein-coding gene mRNAs were reproducible MSI2 targets, compared to 0.2% of long non-coding RNAs (Figure S3d), suggesting that MSI2 controls the stability or translation of coding mRNAs. Motif analysis identified a consensus pentamer (U/G)UAGU resembling the known mouse Msi1-binding sequence[ONT$^+$12, KLL$^+$14] within binding sites in all genic regions; additionally, MSI2-binding sites were generally significantly more conserved than background and tended to occur after the stop codon (Figure 3d and Figure S3eh). The presence of MSI2 binding sites within Msi1 targets[KLL$^+$14] across species indicates that Musashi proteins may bind the same genes through 3UTR-embedded motifs (Figure S3i). Finally, target gene ontology analysis revealed 186 biological processes categories, among the most significant of which were electron transport, oestrogen receptor signalling regulation and metabolism of small molecules, all processes known to be transcriptionally influenced by AHR signalling[Tij05].

Among the top 2% of enriched CLIPseq targets were the 3UTRs of the genes for two AHR pathway components: heat shock protein 90 (HSP90) and CYP1B1. Each exhibited multiple MSI2-binding motifs correlating with overlapping clusters of CLIPseq reads (Figure 3e and Figure S4a). To investigate the ability of MSI2 to post-transcriptionally regulate these genes during HSPC expansion, we looked for instances of uncoupled transcript and protein expression. HSP90 displayed uncoupling of transcript (1.6-fold up) and protein (1.6-fold down)

expression early in culture, but after 7 days showed further upregulated transcript expression (2.5-fold) and variable protein levels (Figure 3f and Figure S4b). As AHRHSP90 binding is essential for ligand-dependent transcriptional activity[MFK03], downregulation of HSP90 protein at the outset of HSPC culture would be expected to reduce latent AHR complex formation and attenuate AHR signalling. Indeed, CYP1B1 transcript and protein expression displayed twofold reductions early in culture, consistent with decreased AHR pathway activity; however, at day 7, CYP1B1 transcripts were upregulated 1.7-fold and uncoupled from protein expression, which was downregulated twofold (Figure 3g and Figure S4c). To test whether MSI2 directly mediates post-transcriptional repression of these targets, the 3UTRs of CYP1B1 and HSP90 were coupled to luciferase. MSI2 overexpression induced significant reductions in luciferase signal from both reporters, and this effect was mitigated when the core CLIPseq-identified UAG motifs were mutated (Figure S4d, e). As MSI2 overexpression-mediated post-transcriptional downregulation of the AHR pathway converged on CYP1B1 protein repression throughout culture, we explored the effects on HSPCs of inhibiting CYP1B1 independently with (E)-2,3‘,4,5‘-tetramethoxystilbene (TMS). During culture, TMS increased the frequency and total numbers of CD34+ cells by 1.5-fold and 2-fold, respectively (Figure 3h, i), phenocopying the effects of MSI2 overexpression. Finally, overexpression of both CYP1B1 lacking its 3UTR and MSI2 decreased secondary CFU-GEMM replating efficiency (Figure S4f, g); this suggests that CYP1B1, while typically used to report AHR signalling, itself promotes HSPC differentiation.

4.4 METHODS

4.4.1 RIP sample preparation for Western, Northern or RNA-seq analysis

To monitor protein depletions, if used, and recovery of Flag- or Myc-epitope tagged proteins and/or coprecipitating proteins, a fraction of the input lysate and IP was set aside for Western blotting in 1x SDS loading buffer (60 mM Tris-HCl, pH 6.8, 2% SDS, 4% -mercaptoethanol, 10%

glycerol, 0.1% bromophenol blue and xylene cyanol). To monitor RNA, Trizol (Ambion) was added to input, unbound, and IP samples. RNA was isolated according to the manufacturer, with 25 g linear polyacrylamide used as carrier for either library preparation for RNA-seq analysis (see below) or Northern blotting.

4.4.2 RIP-seq and CLIP-seq analysis

Sequence alignment of CLIP-seq and RIP-seq data to the human genome. Sequencing reads from CLIP-seq and RIP-seq libraries were first trimmed of polyA tails, adapters, and low quality ends using cutadapt with parameters –match-read-wildcards –times 2 -e 0 -O 5 –quality-cutoff' 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTCGGGTGCCAAGG -b AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA -b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase18.05. Bowtie version 1.0.0 with parameters -S -q -p 16 -e 100 -l 20 was used to align reads against an index generated from Repbase sequences [LTPS09]. Reads not mapped to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR [DDS+13] version 2.3.0e with parameters –outSAMunmapped Within outFilterMultimapNmax 1 outFilterMultimapScoreRange 1.

4.4.3 CLIP-seq Cluster Identification and analyses.

Reads that were PCR replicates were removed from each CLIP-seq library using a custom script. Briefly one read was kept at each nucleotide position when more than one reads 5' end was mapped. Clusters were then assigned using the CLIPper software with parameters –bonferroni –superlocal –threshold- software [LGM+13]. Clusters that overlap by at least one base pair are considered overlapping clusters, as determined using bedtools [QH10] and pybedtools[DPQ11].

4.4.4 Read Distribution Region Counting and comparisons.

Genic features were defined using gencode v17 annotations [HDF$^+$06]. For each gene, all annotated transcripts for that gene were combined and gene level features were generated. 5' UTR, CDS, and 3' UTR regions from gencode genes were identified and merged at the gene level. The number of reads mapping to each annotated meta-region for each gene was counted. Each gene region (3' and 5' UTRs and CDS) was binned into 100 bins, and the mean of reads in each bin was calculated. Then for all genes, the mean of each bin was calculated and plotted. Finally, for each experiment the distribution was normalized to compute a probability density across each bin. To compare different regions, the total number of reads within each region was totaled. For each region the RPK (reads per 1,000 bases) was calculated. To find the percent of reads falling into each region, the percent of RPK normalized reads that were contained within a given region per gene was calculated.

4.4.5 Read Distribution Feature Counting.

For read distributions around specific genic features (termination codons and 3 transcript ends), features from gencode v17 were selected, the number of reads around each feature was counted, and the mean for each base was calculated across all features. Finally, for each experiment the distribution was normalized to show a probability density across each bin. A background model was generated assuming a uniform distribution of reads distributed to each base, and reads were normally distributed. Each feature at each base is then an independent observation and a z-test was applied to each base (Bonferroni corrected) to see if the reads at that base differed significantly from the background distribution.

4.4.6 RIP-seq Analysis.

RPKMs for each gene annotated in gencode v17 were calculated from RIP-seq data using custom scripts. We determined the fold-change (log2) threshold by which at most 5% of genes in the FLAG RIP-seq sample were enriched. This threshold reflected a false discovery rate (FDR) of 5%, and was applied to both WT Upf1 and mutant Upf1 RIP-seq samples to identify Upf1 target genes. Non-targets were defined as having a log2 fold change in WT Upf1 RIPs of between -0.05 and 0.05 RPKM

4.4.7 UV CLIPseq library preparation

CLIPseq was performed as previously described[YCL$^+$09]. Briefly, 25 million NB4 cells (a transformed human cell line of haematopoietic origin) were washed in PBS and UV-cross-linked at 400mJcm2 on ice. Cells were pelleted, lysed in wash buffer (PBS, 0.1% SDS, 0.5% Na-deoxycholate, 0.5% NP-40) and DNase-treated, and supernatants from lysates were collected for immunoprecipitation. MSI2 was immunoprecipitated overnight using 5g of anti-MSI2 antibody (EP1305Y, Abcam) and Protein A Dynabeads (Invitrogen). Beads containing immunoprecipated RNA were washed twice with wash buffer, high-salt wash buffer (5 PBS, 0.1% SDS, 0.5% Na-Deoxycholate, 0.5% NP-40), and PNK buffer (50mM Tris-Cl pH 7.4, 10mM MgCl2, 0.5% NP-40). Samples were then treated with 0.2U MNase for 5min at 37 with shaking to trim immunopreciptated RNA. MNase inactivation was then carried out with PNK + EGTA buffer (50mM Tris-Cl pH 7.4, 20mM EGTA, 0.5% NP-40). The sample was dephosphorylated using alkaline phosphatase (CIP, NEB) at 37 for 10min followed by washing with PNK+EGTA, PNK buffer, and then 0.1mgml1 BSA in nuclease-free water. 3RNA linker ligation was performed at 16 overnight with the following adaptor: 5P-UGGAAUUCUCGGGUGCCAAGG-puromycin. Samples were then washed with PNK buffer, radiolabelled using P32-y-ATP (Perkin Elmer), run on a 4-12% Bis-Tris gel and then transferred to a nitrocellulose membrane. The nitrocellulose

membrane was developed via autoradiography and RNAprotein complexes 1520kDa above the molecular weight of MSI2 were extracted with proteinase K followed by RNA extraction with acid phenol-chloroform. A 5RNA linker (5HO-GUUCAGAGUUCUACAGUCCGACGAUC-OH) was ligated to the extracted RNA using T4 RNA ligase (Fermentas) for 2h and the RNA was again purified using acid phenol-chloroform. Adaptor ligated RNA was re-suspended in nuclease-free water and reverse transcribed using Superscript III reverse transcriptase (Invitrogen). Twenty cycles of PCR were performed using NEB Phusion Polymerase using a 3PCR primer that contained a unique Illumina barcode sequence. PCR products were run on an 8% TBE gel. Products ranging between 150 and 200bp were extracted using the QIAquick gel extraction kit (Qiagen) and re-suspended in nuclease-free water. Two separate libraries were prepared and sent for single-end 50-bp Illumina sequencing at the Institute for Genomic Medicine at the University of California, San Diego. 47,098,127 reads from the first library passed quality filtering, of which 73.83% mapped uniquely to the human genome. 57,970,220 reads from the second library passed quality filtering, of which 69.53% mapped uniquely to the human genome. CLIP-data reproducibility was verified through high correlation between gene RPKMs and statistically significant overlaps in the clusters and genes within replicates. CLIPseq data have been deposited in NCBIs GEO and are accessible through GEO Series accession number GSE69583.

4.4.8 CLIPseq mapping and cluster identification

Before sequence alignment of CLIPseq reads to the human genome was performed, sequencing reads from libraries were trimmed of polyA tails, adapters, and low quality ends using Cutadapt with parameters -match-read-wildcards -times 2 -e 0 -O 5 -quality-cutoff 6 -m 18 -b TCGTATGCCGTCTTCTGCTTG -b ATCTCGTATGCCGTCTTCTGCTTG -b CGACAGGTTCAGAGTTCTACAGTCCGACGATC -b TGGAATTCTCGGGTGCCAAGG -b AAAAAAAAAAAAAAAAAAAA -b TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT. Reads were then mapped against a database of repetitive elements derived from RepBase (version 18.05). Bowtie

(version 1.0.0) with parameters -S -q -p 16 -e 100 -l 20 was used to align reads against an index generated from Repbase sequences[LTPS09]. Reads not mapped to Repbase sequences were aligned to the hg19 human genome (UCSC assembly) using STAR (version 2.3.0e)[DDS+13] with parameters -outSAMunmapped Within -outFilterMultimapNmax 1 -outFilterMultimapScoreRange 1. To identify clusters in the genome of significantly enriched CLIP-seq reads, reads that were PCR replicates were removed from each CLIPseq library using a custom script of the same method as in [Dar12]; otherwise, reads were kept at each nucleotide position when more than one reads 5-end was mapped. Clusters were then assigned using the CLIPper software with parameters -bonferroni -superlocal-threshold[LGM+13]. The ranked list of significant targets was calculated assuming a Poisson distribution, where the observed value is the number of reads in the cluster, and the background is the number of reads across the entire transcript and or across a window of 1000bp$\pm$the predicted cluster.

4.4.9 Gene annotations for CLIPseq

Transcriptomic regions and gene classes were defined using annotations found in gencode v17. Depending on the analysis, clusters were associated by the Gencode-annotated 5UTR, 3UTR, CDS or intronic regions. If a cluster overlapped multiple regions, or a single part of a transcript was annotated as multiple regions, clusters were iteratively assigned first as CDS, then 3UTR, 5UTR and finally as proximal ($<$500 bases from an exon) or distal ($>$500 bases from an exon) introns. Overlapping peaks were calculated using bedtools and pybedtools[QH10, DPQ11].

4.4.10 Gene ontology analysis for CLIPseq

Significantly enriched gene ontology (GO) terms were identified using a hypergeometric test that compared the number of genes that were MSI2 targets in each GO term to genes expressed in each GO term as the proper background. Expressed genes were identified using the control

samples in SRA study SRP012062. Mapping was performed identically to CLIPseq mapping, without peak calling and changing the STAR parameter outFilterMultimapNmax to 10. Counts were calculated with featureCounts37 and RPKMs were then computed. Only genes with a mean RPKM$>$1 between the two samples were used in the background expressed set.

4.4.11 De novo motif and conservation analysis for CLIPseq

Randomly located clusters within the same genic regions as predicted MSI2 clusters were used to calculate a background distribution for motif and conservation analyses. Motif analysis was performed using the HOMER algorithm as in [LGM$^+$13]. For evolutionary sequence conservation analysis, the mean (mammalian) phastCons score for each cluster was used.

4.5 ACKNOWLEDGEMENTS

Chapter 4, in part, is a reprint of the material as it appears in Molecular Cell 2015. Suzanne R. Lee, Gabriel A. Pratt, Fernando J. Martinez, Gene W. Yeo, Jens Lykke-Andersen. Elsevier, 2015. The dissertation/thesis author was an investigator and author of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Nature 2016. Stefan Rentas, Nicholas T. Holzapfel, Muluken S. Belew, Gabriel A. Pratt, Veronique Voisin, Brian T. Wilhelm, Gary D. Bader, Gene W. Yeo, Kristin J. Hope. Nature, 2016. The dissertation/thesis author was an investigator and author of this paper.

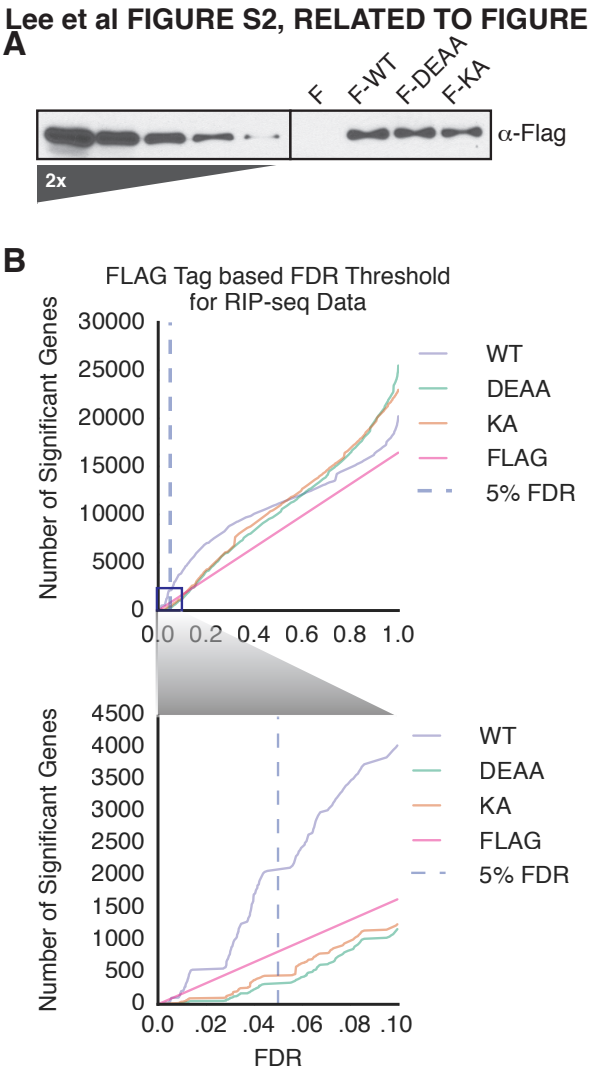

Figure 4.2: **Supplementary Figure 1. Supplemental data related to RIP-RNAseq** (A) Western blot of Flag-Upf1 recovered in RNA-IPs used for the RNA seq in Figure 2 alongside a two-fold titration of Flag-Upf1 WT input lysate. (B) Plot showing derivation of empirical false discovery rate (FDR) based on IP/input ratios for FLAG only, with X-axis as FDR and Y-axis as the number of genes remaining at that FDR threshold for all experiments. Inset depicts region around 5% FDR cut-off (0.05) used for designating genes identified in WT Upf1 RNA-IP as Upf1-enriched.

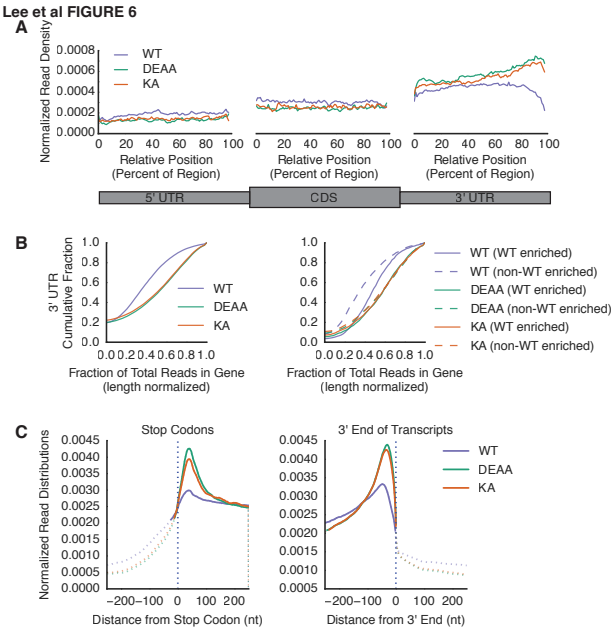

Figure 4.3: **Figure 2. Upf1 WT and ATPase mutants cross-link preferentially in 3UTRs, with elevated crosslinking for ATPase mutants downstream of termination codons and near 3 ends** (A) Mean read density across the metagene, normalized to the total number of reads per gene. (B) Cumulative fraction of genes with 3UTR read abundance represented as a fraction of total reads in the gene, normalized to nucleotide length, shown for all mRNAs on the left, and for WT enriched and non-WT enriched mRNAs, as defined in Figure 2, on the right. Differences between WT and mutant curves were statistically significant (p-value < 0.001; KS statistic 0.30 for both DEAA and KA compared to WT for non-WT Enriched and 0.20 and 0.19 for DEAA and KA, respectively, compared to WT for WT Enriched) (C) Mean read densities, shown as percentages of total reads mapped in the region depicted per mRNA, around the first nucleotide of the translational stop codon, shown on the left, and the 3 end of annotated transcripts, shown on the right. Solid lines represent regions where differences were found to be significant with a P-value <0.05 (Bonferroni corrected).

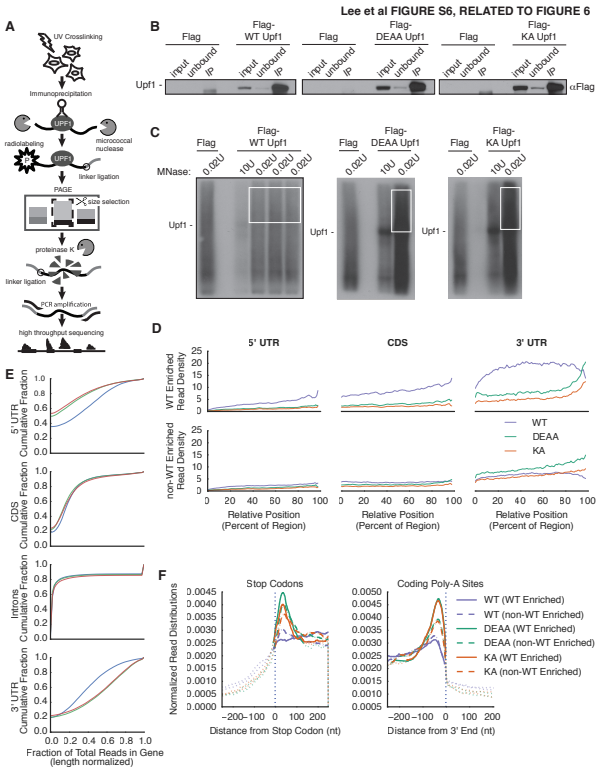

Figure 4.4: **Supplementary Figure 2. Supplemental data related to CLIP-seq** (A) Overview of workflow for CLIP-seq performed with Flag-Upf1 WT, DEAA and KA using α-Flag antibodies for immunoprecipitation. (B) Anti-Flag Western blots of Flag-protein purification for CLIP-seq. (C) Film exposure of polyacrylamide resolved, 32P-end-labeled RNA that remained associated with Flag-Upf1 WT, DEAA and KA compared to Flag epitope-only, after 0.02U or 10U MNase digestion. The prominent band made apparent with 10U MNase treatment only in Flag-Upf samples migrates at a size consistent with Flag-Upf1. White box delineates region excised for CLIP-seq library construction. (D) Mean read density across the metagene, shown as a percentage of total reads in CLIPs for WT enriched and non-WT enriched genes, as defined in Figure 2, not normalized to the total number of reads per gene. (E) Cumulative fraction of genes with regional read abundance represented as a fraction of total reads in the gene, normalized to nucleotide length. UTR, untranslated region; CDS, coding sequence (F) Read density around stop codons and mRNA 3 ends for WT-enriched and non-WT enriched genes. Solid and long-dash lines represent regions where differences were found to be significant with a P-value <0.05 (Bonferroni corrected).

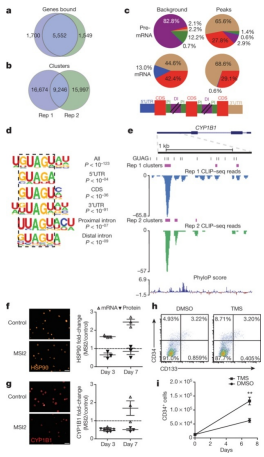

Figure 4.5: **Figure 3. MSI2 overexpression post-transcriptionally downregulates AHR pathway components** a, Overlap between MSI2 target genes from separate CLIPseq experiments. b, Statistically significant overlap (P<0.0001, hypergeometric test) of clusters between the replicates. c, Percentage of CLIPseq clusters in different genic regions. d, Consensus motifs within MSI2 clusters in different genic regions. P values presented for the top 40% of clusters. e, CLIPseq reads (blue, replicate 1; green, replicate 2) and clusters (purple) mapped to the 3UTR of CYP1B1. Matches to the GUAG motif are shown in black. f, g, Immunofluorescence for HSP90 and CYP1B1 3 days after transduction and summary of fold-changes in HSP90 and CYP1B1 protein and transcript levels with MSI2 overexpression at 3 and 7 days after transduction (scale bar, 20m; dotted line indicates no change; n=3 experiments). h, HSPC marker expression by CD34+ cells treated with TMS for 10 days. i, Absolute CD34+ cell number with TMS (n=4 experiments). Data are presented as mean±s.e.m. Unpaired t-test, **P<0.01.

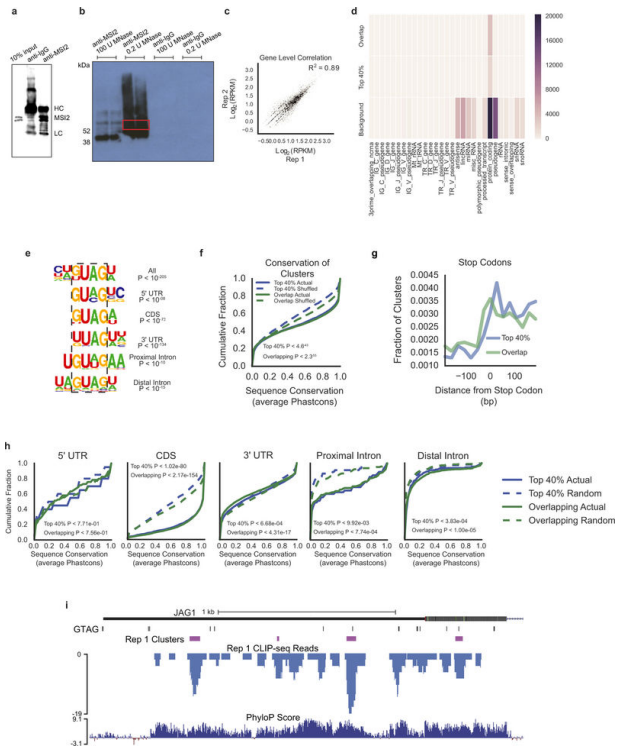

Figure 4.6: **Supplementary Figure 3. MSI2 preferentially binds mature mRNA within the 3'UTR** a, Validation of the capacity of the anti-MSI2 antibody to immunoprecipitate MSI2 compared to IgG control pulldowns. b, Autoradiogram showing anti-MSI2 immunoprecipitated, MNase digested and radiolabelled RNA isolated for CLIP library construction and sequencing (red box). Low levels of MNase show a smearing pattern extending upwards from the modal weight of MSI2. c, Scatter plot of total number of uniquely mapped CLIP-seq reads for each gene, comparing both replicates. d, Heatmap indicating the number of different classes of Gencode annotated genes that contain at least one predicted MSI2 binding site. e, Consensus motifs within MSI2 clusters in the different genic regions. P-values for the most statistically significant enriched motif is presented for all overlapping clusters between replicates. f, Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for all overlapping clusters and the top 40% of overlapping clusters. P- values were obtained by a Kolgomorov-Smirnov two-tailed test comparing the distributions from actual and shuffled locations. g, Number of clusters within 200 bases of the annotated stop codon in known mRNA transcripts for all overlapping clusters between replicates and the top 40% of overlapping clusters. h, Cumulative distribution function of mean conservation score (Phastcons) of MSI2 clusters, compared to a shuffled background control, computed for overlapping clusters between the replicates and the top 40% of overlapping clusters found in different genic regions. Similarity in the 3'UTR conservation for the top 40% with the shuffled background control is likely due to MSI2 sites being small and not needing structural contexts for conservation. P-values were obtained by a Kolgomorov-Smirnov two-tailed test comparing the distributions from actual and shuffled locations. i, Genome browser views displaying CLIP-seq mapped reads from replicate 1 (blue), predicted clusters (purple), exact matches for the GUAG sequence (black) and mammal conservation scores (PhyloP) in the 3' UTRs for a previously predicted Msi1 target.

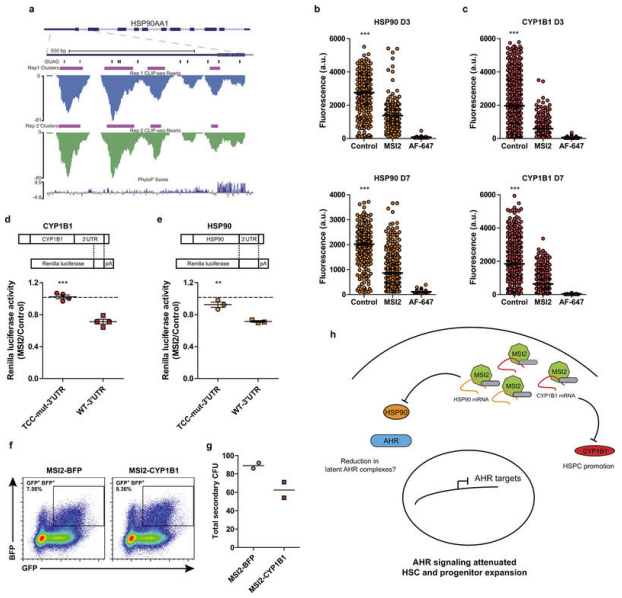

Figure 4.7: **Supplementary Figure 4. MSI2 OE represses CYP1B1 and HSP90 3'UTR Renilla Luciferase reporter activity** a, CLIP-seq reads (replicate 1 in blue and replicate 2 in green) and clusters (purple) mapped to the 3'UTR of HSP90. Matches to the GUAG motif are shown in black. Mammal PhyloP score listed in last track. b and c, Representative data of mean per cell fluoresence for HSP90 and CYP1B1 protein in transduced CD34+ CB. Protein level in cells during in vitro culture was analyzed 3 days (D3) and 7 days (D7) after transduction and sorting for GFP. Corresponding secondary alone antibody staining is shown for each experiment. Each circle represents a cell, and greater than 200 cells were analyzed per condition. d and e, Levels of renilla luciferase activity in NIH-3T3 cells co-transfected with control or MSI2 OE vectors and the CYP1B1 or HSP90 wild type or TCC mutant 3'UTR luciferase reporter (dotted line indicates no change in renilla activity; n=4 CYP1B1 and n=3 HSP90 experiments). f, Flow plots of co-transduced CD34+ CB cells with MSI2 (GFP) and CYP1B1 (BFP) lentivirus. g, GFP+ BFP+ CFU-GEMMs generated from f were replated in to secondary CFU assays and enumerated for total number of colonies formed. A total of 24 CFU-GEMMs from MSI2- BFP and MSI2-CYP1B1 were replated (n=2 experiments). Data presented as mean SEM. ***$p<0.001$, **$p<0.01$. h, A model for AHR pathway attenuation through MSI2 post-transcriptional processing. MSI2 mediates the post-transcriptional down regulation of HSP90 at the outset of culture and continuously represses the prominent AHR pathway effector CYP1B1 to facilitate HSPC expansion. The resultant MSI2-mediated repression of AHR signaling enforces a self-renewal program and allows HSPC expansion ex vivo.

Chapter 5

Discussion and Future Directions

The study of RNA binding proteins, like the study of transcription factors before it has opened up a brand new field in genomics. We have used this power to understand the effects of single RBPs on human health and disease, and are just now understanding how to integrate multiple RBPs to gain a better picture of that is going on in the cell.

5.1 Perspectives on TAF15 project

Single RBP studies can help move the field, even if they dont provide conclusive evidence towards the mechanism of disease. With the TAF15 study we characterized the RBPs binding in the most relevant cell types we could easily gain access to, but conclusions we drew from the results were weak at best. TAF15 appears to at least be loosely associated with ALS, so it may be worthwhile to keep searching for TAF15s true mechanism of action in the disease. It could an unobserved effect on RNA processing, or it may have to do with its aggregation into stress granules, which we were not equipped to study.

5.2 Perspectives on analysis of other RBPs

The value of single RBP studies have been shown in the additional collaborative analysis I performed in the process of writing my dissertation. Thanks CLIP-seq experiments, and integration of results we now better understand the mechanism by which UPF1 functions to degrade NMD targets. Additionally our studies of Musahi 2 have elucidated the mechanism of action drugs used in the process of growing hematopoietic stem cells. I expect in the future this research will directly lead to improved efficiency of bone marrow transplants.

Looking at single RBPs can be hugely informative, and I expect that the majority of studies will continue to be performed utilizing CLIPs on single RBPs.

5.3 Perspectives on eCLIP quality control project

My work on quality control in eCLIP data is the first published suggestions detailing considerations for experimental and computational analysis. I provided pre-sequencing measures of quality control to save money and time for scientists before spending thousands of dollars finishing an experiment. I provided suggestions as to exactly how deeply to sequence to reach an optimal cost to discovery ratio, and most importantly I provided computational quality control metrics post-sequencing so people dont waste time analyzing bad data.

Studies of quality control will continue to be performed. We are just now learning how to effectively scale CLIP-seq assays to understand the effects of hundreds to thousands of experiments at once. In the end quality control for each RBP is not a one size fits all approach. My methods were mostly developed for point source binders, and fail to appreciate the diversity of RBP binding locations and effects. In the future we will need to create more fine grained quality control metrics based on an understanding of RBP function. From the over 2000 datasets that we have in our lab we have just started scratching the surface of binding diversity. At a very basic level we should mark the difference between narrow and broad binders. Pre-mRNA and

mRNA binders should also be looked at differently. Finally RBPs that primary map to repetitive elements need to be looked at in a structured way. We still have trouble knowing exactly what it means to bind to a repetitive element, let alone what level of binding is significant. Repetitive element binders pass QC right now likely due to technical artifacts, but this should be fixed in the future.

In addition to methods refinement to QC researchers could, and should continue to improve the quality of the experimental assay. This could subtlety change the metrics that are used to show quality of data.

Likely the observation with most staying power in my research is the estimate of library complexity. Often times people think that a sequencing library contains infinite molecules. CLIP and iCLIP proved this is not the case. Being able to communicate that sequencing deeper wont actually yield additional information and quantifying that is an obvious, but critical observation for the field.

In the end automated quality control metrics, like automated methods to understand the effect of data can only detect signals you know youre searching for. This is great most of the time, but biology is about the discovery of the unknown, the thing that makes you say hmm, thats unexpected. The best quality control, especially for new RBPs will always be looking at each dataset by eye, and seeing what is unexpected.

5.4 Future needs for CLIP-seq and genomics fields

I see the CLIP-seq field as a whole as just starting, only within the last two years have truly reproducible, scalable methods to understand RBP binding been developed. This will enable a whole new area of basic biology research. With this comes a host of challenges and opportunities.

First and foremost the question we must know who is going to analyze the data. Currently expert experimentalists hand off data to expert computational analysts. This approach worked

when the rate and scale of data generation was small. As high throughput methods get cheaper and faster this will be an inefficient way of doing things. The handoff of data and experimental question between two individuals takes time, not to mention buy in from both parties. There is a need for the data generators to take over primary data analysis, so they can ask the right question, and ask it more quickly than a handoff would allow for. The field isnt there yet, but it is moving in that direction.

On the experimental side understanding the function of an RBP is still an open question. RBPs regulate many more individual pathways than transcription factors, which at the end of the day turn genes on or off, and can be assayed with a simple RNA-seq experiment. The mechanism of action for TFs is still often an open question, but the readout is well known. For RBPs the readout can be any one of tens of different high throughput experiments. RBPs can effect stability, splicing, translation, polyadenylation, capping, one of the 100s of RNA modifications. Going in blind there are too many hypothesis to test to understand the function of a single RBP. We are addressing this somewhat using high throughput screens for RBP function including stability, splicing and translation, but these dont address RBP function in its native context.

In the future we should develop computational approaches, based on RBP binding profiles to automatically provide suggestions as to the function of an RBP. This would help save time and money when trying to assay its function.

Finally we are just now scratching the surface of true integrative analysis between many RBPs binding locations, and other assays to understand their function. We need better methods to understand how RBPs behave cooperatively to regulate cellular state or disease.

To achieve this dream of integrative analysis though more near term issues must be more effectively addressed. It will be difficult or impossible to integrate RBP data without high quality peak calls, and to a lesser extent a better understanding of how to quantify differential binding. In the TF field, integrative finding common co-bound states enabled analysis, based off of high quality peak calls. Truly high quality peak identification algorithms must be developed for RBPs

as well.

The largest current drawback to CLIP-seq peak calling algorithms today is they cant differentiate peak size effectively. Some RBPs bind a specific motif, some bind entire transcripts. Current peak callers use the same model to call point-source peaks for both types of RBPs. A true peak caller would accurately identify both the region of binding for a specific RBP, and provide an estimate of the point-source binding location (if possible) as well.

Important to small scale studies is an understanding of differential binding. Currently peak calling methods have trouble comparing between two conditions. This is primarily because there isnt a solid understanding of the correlation between binding strength and read density. This understanding must be solidified to better understand RBP binding between different conditions.

The future of the RBP field is in basic research. Once RBP binding patterns are better understood machine learning methods can be applied to understand how mutations or changes in gene expression will effect binding and downstream gene regulation. This kind of research and machine learning approach will need to be enabled by vast scale. Currently CLIP-seq is labor intensive. To truly understand effects of all RBPs at scale CLIP-seq will need to be automated, and the price per experiment will need to drop drastically.

These predictions will not only allow the field to predict effects of mutations, but also ideally allow for the prediction of customized RNA (or otherwise) drugs that can treat an individuals specific disease.

In the end through biomedical research we are attempting to cure diseases. Analysis of RNA binding proteins will play their role through basic research, allowing scientists to understand the cause of disease more easily, and providing new therapeutic avenues for treatment.

Bibliography

[AANL12] Aaron Arvey, Phaedra Agius, William Stafford Noble, and Christina Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, 22(9):1723–1734, sep 2012.

[AGK+04] Nadia Amrani, Robin Ganesan, Stephanie Kervestin, David A. Mangus, Shubhendu Ghosh, and Allan Jacobson. A faux 3-UTR promotes aberrant termination and triggers nonsense-mediated mRNA decay. *Nature*, 432(7013):112–118, nov 2004.

[AGVB+11] Sonja Althammer, Juan González-Vallinas, Cecilia Ballaré, Miguel Beato, and Eduardo Eyras. Pyicos: a versatile toolkit for the analysis of high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 27(24):3333–40, dec 2011.

[AHA+06] Tetsuaki Arai, Masato Hasegawa, Haruhiko Akiyama, Kenji Ikeda, Takashi Nonaka, Hiroshi Mori, David Mann, Kuniaki Tsuchiya, Mari Yoshida, Yoshio Hashizume, and Tatsuro Oda. TDP-43 is a component of ubiquitin-positive tau-negative inclusions in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Biochemical and biophysical research communications*, 351(3):602–11, dec 2006.

[ÄN12] Minna Liisa Änkö and Karla M. Neugebauer. RNA-protein interactions in vivo: Global gets specific. *Trends in Biochemical Sciences*, 37(7):255–262, jul 2012.

[AZH+11] Adam Ameur, Ammar Zaghlool, Jonatan Halvardson, Anna Wetterbom, Ulf Gyllensten, Lucia Cavelier, and Lars Feuk. Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nature structural & molecular biology*, 18(12):1435–40, dec 2011.

[Bai11] Timothy L. Bailey. DREME: Motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, jun 2011.

[Bar04] David P Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, jan 2004.

[BB11] Mauricio Budini and Emanuele Buratti. TDP-43 autoregulation: implications for disease. *Journal of molecular neuroscience : MN*, 45(3):473–9, dec 2011.

[BB12] Emanuele Buratti and Francisco E. Baralle. TDP-43: Gumming up neurons through protein-protein and protein-RNA interactions. *Trends in Biochemical Sciences*, 37(6):237–247, 2012.

[BB14] Silpi Banerjee and Pierre Barraud. Functions of double-stranded RNA-binding domains in nucleocytoplasmic transport. *RNA Biology*, 11(10):1226–1232, oct 2014.

[BC16] Amy E. Brinegar and Thomas A. Cooper. Roles for RNA-binding proteins in development and disease. *Brain Research*, 1647:1–8, sep 2016.

[BDH+13] Lukasz S. Borowski, Andrzej Dziembowski, Monika S. Hejnowicz, Piotr P. Stepien, and Roman J. Szczesny. Human mitochondrial RNA decay mediated by PNPasehSuv3 complex takes place in distinct foci. *Nucleic Acids Research*, 41(2):1223–1240, jan 2013.

[BHV+16] Ranjan Batra, Kasey Hutt, Anthony Vu, Stuart J Rabin, Michael W Baughn, Ryan T Libby, Shawn Hoon, John Ravits, and Gene W Yeo. Gene Expression Signatures of Sporadic ALS Motor Neuron Populations. Technical report, feb 2016.

[BLH+96] a Bertolotti, Y Lutz, D J Heard, P Chambon, and L Tora. hTAF(II)68, a novel RNA/ssDNA-binding protein with homology to the pro-oncoproteins TLS/FUS and EWS is associated with both TFIID and RNA polymerase II. *The EMBO journal*, 15(18):5022–5031, 1996.

[BLH+16] James P. Broughton, Michael T. Lovci, Jessica L. Huang, Gene W. Yeo, and Amy E. Pasquinelli. Pairing beyond the Seed Supports MicroRNA Targeting Specificity. *Molecular Cell*, 64(2):320–333, oct 2016.

[Boe16] Valentina Boeva. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Frontiers in Genetics*, 7(FEB):24, feb 2016.

[BSD+07] Ewan Birney, John A. Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R. Gingeras, Elliott H. Margulies, Zhiping Weng, Michael Snyder, Emmanouil T. Dermitzakis, Robert E. Thurman, Michael S. Kuehn, Christopher M. Taylor, Shane Neph, Christoph M. Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A. Greenbaum, Robert M. Andrews, Paul Flicek, Patrick J. Boyle, Hua Cao, Nigel P. Carter, Gayle K. Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C. Dillon, Michael O. Dorschner, Heike Fiegler, Paul G. Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D. James, Brett E. Johnson, Ericka M. Johnson, Tristan T. Frum, Elizabeth R. Rosenzweig,

Neerja Karnani, Kirsten Lee, Gregory C. Lefebvre, Patrick A. Navas, Fidencio Neri, Stephen C.J. Parker, Peter J. Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S. Collins, Job Dekker, Jason D. Lieb, Thomas D. Tullius, Gregory E. Crawford, Shamil Sunyaev, William S. Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L. Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A. Hirsch, Edward A. Sekinger, Julien Lagarde, Josep F. Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S. Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C. Dickson, Daryl J. Thomas, Matthew T. Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, Xiaodong Zhao, K. G. Srinivasan, Wing Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L. Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G. Clark, James B. Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N. Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S. Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M. Myers, Jane Rogers, Peter F. Stadler, Todd M. Lowe, Chia Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E. Antonarakis, Yutao Fu, Eric D. Green, Ula Karaöz, Adam Siepel, James Taylor, Laura A. Liefer, Kris A. Wetterstrand, Peter J. Good, Elise A. Feingold, Mark S. Guyer, Gregory M. Cooper, George Asimenos, Colin N. Dewey, Minmei Hou, Sergey Nikolaev, Juan I. Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R. Zhang, Ian Holmes, James C. Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W. James Kent, Eric A. Stone, Serafim Batzoglou, Nick Goldman, Ross C. Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D. Trinklein, Zhengdong D. Zhang, Leah Barrera, Rhona Stuart, David C. King, Adam Ameur, Stefan Enroth, Mark C. Bieda, Jonghwan Kim, Akshay A. Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B. Vega, Charlie W.H. Lee, Patrick Ng, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J. Oberley, David Inman, Michael A. Singer, Todd A. Richmond, Kyle J. Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C. Fowler, Phillippe Couttet, Alexander W. Bruce, Oliver M. Dovey, Peter D. Ellis, Cordelia F. Langford, David A. Nix, Ghia Euskirchen, Stephen Hartman, Alexander E. Urban, Peter Kraus, Sara Van Calcar, Nate Heintzman, Tae Hoon Kim, Kun Wang, Chunxu Qu, Gary Hon, Rosa Luna, Christopher K. Glass, M. Geoff Rosenfeld, Shelley Force Aldred, Sara J. Cooper, Anason Halees, Jane M. Lin, Hennady P. Shulha, Xiaoling Zhang, Mousheng Xu, Jaafar N.S. Haidar, Yong Yu, Vishwanath R. Iyer, Roland D. Green, Claes Wadelius, Peggy J. Farnham, Bing

Ren, Rachel A. Harte, Angie S. Hinrichs, Heather Trumbower, Hiram Clawson, Jennifer Hillman-Jackson, Ann S. Zweig, Kayla Smith, Archana Thakkapallayil, Galt Barber, Robert M. Kuhn, Donna Karolchik, Lluis Armengol, Christine P. Bird, Paul I.W. De Bakker, Andrew D. Kern, Nuria Lopez-Bigas, Joel D. Martin, Barbara E. Stranger, Abigail Woodroffe, Eugene Davydov, Antigone Dimas, Eduardo Eyras, Ingileif B. Hallgrímsdóttir, Julian Huppert, Michael C. Zody, Gonçalo R. Abecasis, Xavier Estivill, Gerard G. Bouffard, Xiaobin Guan, Nancy F. Hansen, Jacquelyn R. Idol, Valerie V.B. Maduro, Baishali Maskeri, Jennifer C. McDowell, Morgan Park, Pamela J. Thomas, Alice C. Young, Robert W. Blakesley, Donna M. Muzny, Erica Sodergren, David A. Wheeler, Kim C. Worley, Huaiyang Jiang, George M. Weinstock, Richard A. Gibbs, Tina Graves, Robert Fulton, Elaine R. Mardis, Richard K. Wilson, Michele Clamp, James Cuff, Sante Gnerre, David B. Jaffe, Jean L. Chang, Kerstin Lindblad-Toh, Eric S. Lander, Maxim Koriabine, Mikhail Nefedov, Kazutoyo Osoegawa, Yuko Yoshinaga, Baoli Zhu, and Pieter J. De Jong. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, jun 2007.

[BSPSU15] Emad Bahrami-Samani, Luiz O.F. Penalva, Andrew D. Smith, and Philip J. Uren. Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, 43(1):95–103, jan 2015.

[BWR+10] A. E. Boitano, J. Wang, R. Romeo, L. C. Bouchez, A. E. Parker, S. E. Sutton, J. R. Walker, C. A. Flaveny, G. H. Perdew, M. S. Denison, P. G. Schultz, and M. P. Cooke. Aryl Hydrocarbon Receptor Antagonists Promote the Expansion of Human Hematopoietic Stem Cells. *Science*, 329(5997):1345–1348, sep 2010.

[CBC+02] Bertrand Cosson, Nadia Berkova, Anne Couturier, Svetlana Chabelskaya, Michel Philippe, and Galina Zhouravleva. Poly(A)-binding protein and eRF3 are associated in vivo in human and Xenopus cells. *Biology of the Cell*, 94(4-5):205–216, sep 2002.

[CBS+14] Sutapa Chakrabarti, Fabien Bonneau, Steffen Schüssler, Elfriede Eppinger, and Elena Conti. Phospho-dependent and phospho-independent interactions of the helicase UPF1 with the NMD factors SMG5-SMG7 and SMG6. *Nucleic Acids Research*, 42(14):9447–9460, aug 2014.

[CCR+11] William A. Cantara, Pamela F. Crain, Jef Rozenski, James A. McCloskey, Kimberly A. Harris, Xiaonong Zhang, Franck A.P. Vendeix, Daniele Fabris, and Paul F. Agris. The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research*, 39(SUPPL. 1):D195–D201, jan 2011.

[CFP+09] Stuart M Chambers, Christopher A Fasano, Eirini P Papapetrou, Mark Tomishima, Michel Sadelain, and Lorenz Studer. Highly efficient neural conversion of human ES and iPS cells by dual inhibition of SMAD signaling. *Nature biotechnology*, 27(3):275–80, mar 2009.

[CGM+11] David L Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L Skalsky, Jack D Keene, and Uwe Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome biology*, 12(8):R79, jan 2011.

[CHC+13] Hana Cho, Sisu Han, Junho Choe, Seung Gu Park, Sun Shim Choi, and Yoon Ki Kim. SMG5-PNRC2 is functionally dominant compared with SMG5-SMG7 in mammalian nonsense-mediated mRNA decay. *Nucleic Acids Research*, 41(2):1319–1328, jan 2013.

[CHS+11] Julien Couthouis, Michael P Hart, James Shorter, Mariely DeJesus-Hernandez, Renske Erion, Rachel Oristano, Annie X Liu, Daniel Ramos, Niti Jethava, Divya Hosangadi, James Epstein, Ashley Chiang, Zamia Diaz, Tadashi Nakaya, Fadia Ibrahim, Hyung-Jun Kim, Jennifer A Solski, Kelly L Williams, Jelena Mojsilovic-Petrovic, Caroline Ingre, Kevin Boylan, Neill R Graff-Radford, Dennis W Dickson, Dana Clay-Falcone, Lauren Elman, Leo McCluskey, Robert Greene, Robert G Kalb, Virginia M-Y Lee, John Q Trojanowski, Albert Ludolph, Wim Robberecht, Peter M Andersen, Garth A Nicholson, Ian P Blair, Oliver D King, Nancy M Bonini, Vivianna Van Deerlin, Rosa Rademakers, Zissimos Mourelatos, and Aaron D Gitler. A yeast functional screen predicts new candidate ALS disease genes. *Proceedings of the National Academy of Sciences of the United States of America*, 108(52):20881–90, dec 2011.

[CKZ+11] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research*, 39(Database):D301–D308, jan 2011.

[CREP+98] Kevin Czaplinski, Maria J. Ruiz-Echevarria, Sergey V. Paushkin, Xia Han, Youmin Weng, Haley A. Perlick, Harry C. Dietz, Michael D. Ter-Avanesyan, and Stuart W. Peltz. The surveillance complex interacts with the translation release factors to enhance termination and degrade aberrant mRNAs. *Genes and Development*, 12(11):1665–1677, jun 1998.

[CRG+09] Lucia Corrado, A Ratti, C Gellera, E Buratti, B Castellotti, Y Carlomagno, N Ticozzi, L Mazzini, L Testa, F Taroni, F E Baralle, V Silani, and S D'Alfonso. High frequency of TARDBP gene mutations in Italian patients with amyotrophic lateral sclerosis. *Human mutation*, 30(4):688–94, apr 2009.

[CVP+16] Anne E. E Conway, Eric L. Van Nostrand, Gabriel A. A Pratt, Stefan Aigner, Melissa L. L Wilbert, Balaji Sundararaman, Peter Freese, Nicole J. J Lambert, Shashank Sathe, Tiffany Y. Y Liang, Anthony Essex, Severine Landais, Christopher B. B Burge, D. Leanne Leanne Jones, Gene W. W Yeo, Eric L Van Nostrand, Gabriel A. A Pratt, Stefan Aigner, Melissa L. L Wilbert, Balaji Sundararaman, Peter Freese, Nicole J. J Lambert, Shashank Sathe, Tiffany Y. Y Liang, Anthony Essex, Severine Landais, Christopher B. B Burge, D. Leanne Leanne Jones, and

Gene W. W Yeo. Enhanced CLIP Uncovers IMP Protein-RNA Targets in Human Pluripotent Stem Cells Important for Cell Adhesion and Survival. *Cell reports*, 15(3):666–679, apr 2016.

[CYK+14] Beibei Chen, Jonghyun Yun, Min Soo Kim, Joshua T Mendell, and Yang Xie. PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis. *Genome biology*, 15(1):R18, jan 2014.

[Dar12] Robert Darnell. CLIP (Cross-linking and immunoprecipitation) identification of RNAs bound by a specific protein. *Cold Spring Harbor Protocols*, 7(11):1146–1160, nov 2012.

[DDS+13] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, jan 2013.

[DFA+17] Daniel Dominguez, Peter Freese, Maria S. Alexis, Amanda Su, Myles Hochman, Tsultrim Palden, Cassandra Bazile, Nicole J Lambert, Eric L. Van Nostrand, Gabriel A Pratt, Gene W Yeo, Brenton Graveley, and Christopher B Burge. Sequence, Structure and Context Preferences of Human RNA Binding Proteins. *Doi.Org*, page 201996, oct 2017.

[DGC+09] R Del Bo, S Ghezzi, S Corti, M Pandolfo, M Ranieri, D Santoro, I Ghione, A Prelle, V Orsetti, M Mancuso, G Sorarù, C Briani, C Angelini, G Siciliano, N Bresolin, and G P Comi. TARDBP (TDP-43) sequence analysis in patients with familial and sporadic ALS: identification of two novel mutations. *European journal of neurology : the official journal of the European Federation of Neurological Societies*, 16(6):727–32, jun 2009.

[DHMB+11] Mariely DeJesus-Hernandez, Ian R. Mackenzie, Bradley F. Boeve, Adam L. Boxer, Matt Baker, Nicola J. Rutherford, Alexandra M. Nicholson, Ni Cole A. Finch, Heather Flynn, Jennifer Adamson, Naomi Kouri, Aleksandra Wojtas, Pheth Sengdy, Ging Yuek R. Hsiung, Anna Karydas, William W. Seeley, Keith A. Josephs, Giovanni Coppola, Daniel H. Geschwind, Zbigniew K. Wszolek, Howard Feldman, David S. Knopman, Ronald C. Petersen, Bruce L. Miller, Dennis W. Dickson, Kevin B. Boylan, Neill R. Graff-Radford, and Rosa Rademakers. Expanded GGGGCC Hexanucleotide Repeat in Noncoding Region of C9ORF72 Causes Chromosome 9p-Linked FTD and ALS. *Neuron*, 72(2):245–256, oct 2011.

[DKA+12] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun

Dong, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Pouya Kheradpour, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C.J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A.L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Peter J. Good, Elise A. Feingold, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Morgan C. Giddings, Thomas R. Gingeras, Roderic Guigó, Timothy J. Hubbard, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Matthew L. Eaton, Alex Dobin, Andrea Tanzer, Julien Lagarde, Wei Lin, Chenghai Xue, Brian A. Williams, Chris Zaleski, Maik Röder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Cédric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, Yoshihide Hayashizaki, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Yijun Ruan, Piero Carninci, Cricket A. Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Linda L. Grasfeder, Paul G. Giresi, Anna Battenhouse, Nathan C. Sheffield, Kimberly A. Showers, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Vishwanath R. Iyer, Meizhen Zheng, Ping Wang, Jason Gertz, Jost Vielmetter, E. Christopher Partridge, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Michael A. Muratet, Kimberly M. New-

berry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Wing H. Wong, Matthew J. Blow, Axel Visel, Len A. Pennachio, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Claire Davidson, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Electra Tapanari, Michael L. Tress, Marijke J. Van Baren, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Alfonso Valencia, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J. Mu, Henriette O'Geen, Zhengqing Ouyang, Dorrelyn Patacsil, Debasish Raha, Lucia Ramirez, Brian Reed, Minyi Shi, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon Kiu Yan, Xinqiong Yang, Kevin Struhl, Sherman M. Weissman, Luiz O. Penalva, Subhradip Karmakar, Raj R. Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Alec Victorsen, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Gaurav Jain, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Audra K. Johnson, Ericka M. Johnson, Tattyana V. Kutyavin, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Minerva E. Sanchez, Richard S. Sandstrom, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Kathryn Beal, Alvis Brazma, Paul Flicek, Nathan Johnson, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia

Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Webb Miller, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, Haiyan Huang, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Bilmes, Orion J. Buske, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, and Lucas Lochovsky. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012.

[DPQ11] Ryan K. Dale, Brent S. Pedersen, and Aaron R. Quinlan. Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, 27(24):3423–3424, dec 2011.

[DVK+09] H Daoud, P N Valdmanis, E Kabashi, P Dion, N Dupré, W Camu, V Meininger, and G A Rouleau. Contribution of TARDBP mutations to sporadic amyotrophic lateral sclerosis. *Journal of medical genetics*, 46(2):112–4, mar 2009.

[DVZ+11] Jennifer C Darnell, Sarah J Van Driesche, Chaolin Zhang, Ka Ying Sharon Hung, Aldo Mele, Claire E Fraser, Elizabeth F Stone, Cynthia Chen, John J Fak, Sung Wook Chi, Donny D Licatalosi, Joel D Richter, and Robert B Darnell. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2):247–61, jul 2011.

[EK12a] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–6, jan 2012.

[EK12b] Jason Ernst and Manolis Kellis. ChromHMM: automating chromatin-state discovery and characterization. *Nature methods*, 9(3):215–6, jan 2012.

[EK15] Jason Ernst and Manolis Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology*, 33(4):364–376, feb 2015.

[ELAMJ09] Andrea B. Eberle, Søren Lykke-Andersen, Oliver Mühlemann, and Torben Heick Jensen. SMG6 promotes endonucleolytic cleavage of nonsense mRNA in human cells. *Nature Structural and Molecular Biology*, 16(1):49–55, jan 2009.

[EUA13] Ran Elkon, Alejandro P. Ugalde, and Reuven Agami. Alternative cleavage and polyadenylation: Extent, regulation and function. *Nature Reviews Genetics*, 14(7):496–506, jun 2013.

[FA14] Xiang Dong Fu and Manuel Ares. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701, aug 2014.

[FCG+14] I. Fares, J. Chagraoui, Y. Gareau, S. Gingras, R. Ruel, N. Mayotte, E. Csaszar, D. J. H. F. Knapp, P. Miller, M. Ngom, S. Imren, D.-C. Roy, K. L. Watts, H.-P. Kiem, R. Herrington, N. N. Iscove, R. K. Humphries, C. J. Eaves, S. Cohen, A. Marinier,

P. W. Zandstra, and G. Sauvageau. Pyrimidoindole derivatives are agonists of human hematopoietic stem cell self-renewal. *Science*, 345(6203):1509–1512, sep 2014.

[FGG+04] E. A. Feingold, P. J. Good, M. S. Guyer, S. Kamholz, L. Liefer, K. Wetterstrand, F. S. Collins, T. R. Gingeras, D. Kampa, E. A. Sekinger, J. Cheng, H. Hirsch, S. Ghosh, Z. Zhu, S. Patel, A. Piccolboni, A. Yang, H. Tammana, S. Bekiranov, P. Kapranov, R. Harrison, G. Church, K. Struhl, B. Ren, T. H. Kim, L. O. Barrera, C. Qu, S. van Calcar, R. Luna, C. K. Glass, M. G. Rosenfeld, R. Guigo, S. E. Antonarakis, E. Birney, M. Brent, L. Pachter, A. Reymond, E. T. Dermitzakis, C. Dewey, D. Keefe, F. Denoeud, J. Lagarde, J. Ashurst, T. Hubbard, J. J. Wesselink, R. Castelo, E. Eyras, R. M. Myers, A. Sidow, S. Batzoglou, N. D. Trinklein, S. J. Hartman, S. F. Aldred, E. Anton, D. I. Schroeder, S. S. Marticke, L. Nguyen, J. Schmutz, J. Grimwood, M. Dickson, G. M. Cooper, E. A. Stone, G. Asimenos, M. Brudno, A. Dutta, N. Karnani, C. M. Taylor, H. K. Kim, G. Robins, G. Stamatoyannopoulos, J. A. Stamatoyannopoulos, M. Dorschner, P. Sabo, M. Hawrylycz, R. Humbert, J. Wallace, M. Yu, P. A. Navas, M. McArthur, W. S. Noble, I. Dunham, C. M. Koch, R. M. Andrews, G. K. Clelland, S. Wilcox, J. C. Fowler, K. D. James, P. Groth, O. M. Dovey, P. D. Ellis, V. L. Wraight, A. J. Mungall, P. Dhami, H. Fiegler, C. F. Langford, N. P. Carter, D. Vetrie, M. Snyder, G. Euskirchen, A. E. Urban, U. Nagalakshmi, J. Rinn, G. Popescu, P. Bertone, S. Hartman, J. Rozowsky, O. Emanuelsson, T. Royce, S. Chung, M. Gerstein, Z. Lian, J. Lian, Y. Nakayama, S. Weissman, V. Stolc, W. Tongprasit, H. Sethi, S. Jones, M. Marra, H. Shin, J. Schein, M. Clamp, K. Lindblad-Toh, J. Chang, D. B. Jaffe, M. Kamal, E. S. Lander, T. S. Mikkelsen, J. Vinson, M. C. Zody, P. J. de Jong, K. Osoegawa, M. Nefedov, B. Zhu, A. D. Baxevanis, T. G. Wolfsberg, G. E. Crawford, J. Whittle, I. E. Holt, T. J. Vasicek, D. Zhou, S. Luo, E. D. Green, G. G. Bouffard, E. H. Margulies, M. E. Portnoy, N. F. Hansen, P. J. Thomas, J. C. McDowell, B. Maskeri, A. C. Young, J. R. Idol, R. W. Blakesley, G. Schuler, W. Miller, R. Hardison, L. Elnitski, P. Shah, S. L. Salzberg, M. Pertea, W. H. Majoros, D. Haussler, D. Thomas, K. R. Rosenbloom, H. Clawson, A. Siepel, W. J. Kent, Z. Weng, S. Jin, A. Halees, H. Burden, U. Karaoz, Y. Fu, Y. Yu, C. Ding, C. R. Cantor, R. E. Kingston, J. Dennis, R. D. Green, M. A. Singer, T. A. Richmond, J. E. Norton, P. J. Farnham, M. J. Oberley, D. R. Inman, M. R. McCormick, H. Kim, C. L. Middle, M. C. Pirrung, X. D. Fu, Y. S. Kwon, Z. Ye, J. Dekker, T. M. Tabuchi, N. Gheldof, J. Dostie, and S. C. Harvey. The ENCODE (ENCyclopedia of DNA Elements) Project. *Science*, 306(5696):636–640, oct 2004.

[FSLA10] Tobias M. Franks, Guramrit Singh, and Jens Lykke-Andersen. Upf1 ATPase-dependent mRNP disassembly is required for completion of nonsense-mediated mRNA decay. *Cell*, 143(6):938–950, 2010.

[FWGJ10] Margaret E. Fairman-Williams, Ulf Peter Guenther, and Eckhard Jankowsky.

SF1 and SF2 helicases: Family matters. *Current Opinion in Structural Biology*, 20(3):313–324, jun 2010.

[FZS+16] Ryan A. Flynn, Qiangfeng Cliff Zhang, Robert C. Spitale, Byron Lee, Maxwell R. Mumbach, and Howard Y. Chang. Transcriptome-wide interrogation of RNA secondary structure in living cells with icSHAPE. *Nature Protocols*, 11(2):273–290, 2016.

[GBYD08] Tina Glisovic, Jennifer L. Bachorik, Jeongsik Yong, and Gideon Dreyfuss. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14):1977–1986, jun 2008.

[GCZ+11] Weirui Guo, Yanbo Chen, Xiaohong Zhou, Amar Kar, Payal Ray, Xiaoping Chen, Elizabeth J Rao, Mengxue Yang, Haihong Ye, Li Zhu, Jianghong Liu, Meng Xu, Yanlian Yang, Chen Wang, David Zhang, Eileen H Bigio, Marsel Mesulam, Yan Shen, Qi Xu, Kazuo Fushimi, and Jane Y Wu. An ALS-associated mutation affecting TDP-43 enhances protein aggregation, fibril formation and neurotoxicity. *Nature structural & molecular biology*, 18(7):822–30, jul 2011.

[GFM+14] Stefanie Grosswendt, Andrei Filipchyk, Mark Manzano, Filippos Klironomos, Marcel Schilling, Margareta Herzog, Eva Gottwein, and Nikolaus Rajewsky. Unambiguous Identification of miRNA: Target site interactions by different types of ligation reactions. *Molecular Cell*, 54(6):1042–1054, jun 2014.

[GHT14] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews Genetics*, 15(12):829–845, nov 2014.

[GPA+10] Yuchun Guo, Georgios Papachristoudis, Robert C. Altshuler, Georg K. Gerber, Tommi S. Jaakkola, David K. Gifford, and Shaun Mahony. Discovering homotypic binding events at high spatial resolution. *Bioinformatics*, 26(24):3028–3034, 2010.

[Gre79] Jay R. Greenberg. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Research*, 6(2):715–732, feb 1979.

[GSM+14] Lea H. Gregersen, Markus Schueler, Mathias Munschauer, Guido Mastrobuoni, Wei Chen, Stefan Kempa, Christoph Dieterich, and Markus Landthaler. MOV10 Is a 5' to 3' RNA Helicase Contributing to UPF1 mRNA Target Degradation by Translocation along 3' UTRs. *Molecular Cell*, 54(4):573–585, 2014.

[HB94] Peter E. Hodges and Jean D. Beggs. U2 fulfils a commitment Genetic and biochemical studies of pre-mRNA splicing have recently. *Current Biology*, 4(3):264–267, mar 1994.

[HBS+10] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C Lin, Peter Laslo, Jason X Cheng, Cornelis Murre, Harinder Singh, and Christopher K Glass. Simple combinations of lineage-determining transcription factors prime

cis-regulatory elements required for macrophage and B cell identities. *Molecular cell*, 38(4):576–89, may 2010.

[HBW+12] Michael M Hoffman, Orion J Buske, Jie Wang, Zhiping Weng, Jeff A Bilmes, and William Stafford Noble. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature methods*, 9(5):473–476, mar 2012.

[HDF+06] Jennifer Harrow, France Denoeud, Adam Frankish, Alexandre Reymond, Chao-Kung Chen, Jacqueline Chrast, Julien Lagarde, James GR Gilbert, Roy Storey, David Swarbreck, Colette Rossier, Catherine Ucla, Tim Hubbard, Stylianos E Antonarakis, Roderic Guigo, JS Mattick, DP Bartel, R Guigo, P Flicek, J Abril, A Reymond, J Lagarde, F Denoeud, S Antonarakis, M Ashburner, VB Bajic, E Birney, P Deloukas, LH Matthews, J Ashurst, J Burton, JG Gilbert, M Jones, G Stavrides, JP Almeida, AK Babbage, CL Bagguley, CL Will, R Luhrmann, G Parra, E Blanco, R Guigo, C Burge, S Karlin, M Wang, J Buhler, MR Brent, T Wiehe, S Gebauer-Jung, T Mitchell-Olds, R Guigo, AA Salamov, VV Solovyev, E Eyras, M Caccamo, V Curwen, M Clamp, E Birney, TD Andrews, P Bevan, M Caccamo, Y Chen, L Clarke, G Coates, J Cuff, V Curwen, T Cutts, M Kozak, BP Lewis, RE Green, SE Brenner, U Ohler, N Shomron, CB Burge, P Kapranov, J Drenkow, J Cheng, J Long, G Helt, S Dike, TR Gingeras, T Shiraki, S Kondo, S Katayama, K Waki, T Kasukawa, H Kawaji, R Kodzius, A Watahiki, M Nakamura, T Arakawa, P Ng, CL Wei, WK Sung, KP Chiu, L Lipovich, CC Ang, S Gupta, A Shahab, A Ridwan, CH Wong, SC Potter, L Clarke, V Curwen, S Keenan, E Mongin, SM Searle, A Stabenau, R Storey, M Clamp, P Rice, I Longden, A Bleasby, G Benson, R Mott, E Birney, M Clamp, R Durbin, TM Lowe, SR Eddy, TA Down, TJ Hubbard, SM Searle, J Gilbert, V Iyer, M Clamp, EL Sonnhammer, JC Wootton, A Reymond, M Friedli, CN Henrichsen, F Chapot, S Deutsch, C Ucla, C Rossier, R Lyle, M Guipponi, SE Antonarakis, A Reymond, AA Camargo, S Deutsch, BJ Stevenson, RB Parmigiani, C Ucla, F Bettoni, C Rossier, R Lyle, M Guipponi, R Guigo, ET Dermitzakis, P Agarwal, CP Ponting, G Parra, A Reymond, JF Abril, E Keibler, R Lyle, and C Ucla. GENCODE: producing a reference annotation for ENCODE. *Genome Biology*, 7(Suppl 1):S4, 2006.

[HEW+13] Michael M. Hoffman, Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, Paul M. Ellenbogen, Jeffrey A. Bilmes, Ewan Birney, Ross C. Hardison, Ian Dunham, Manolis Kellis, and William Stafford Noble. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Research*, 41(2):827–841, jan 2013.

[HFG+12] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, If Barnes, Alexandra Bignell, Veronika Boychenko, Toby Hunt,

Mike Kay, Gaurab Mukherjee, Jeena Rajan, Gloria Despacio-Reyes, Gary Saunders, Charles Steward, Rachel Harte, Michael Lin, Cédric Howald, Andrea Tanzer, Thomas Derrien, Jacqueline Chrast, Nathalie Walters, Suganthi Balasubramanian, Baikang Pei, Michael Tress, Jose Manuel Rodriguez, Iakes Ezkurdia, Jeltje van Baren, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Alexandre Reymond, Mark Gerstein, Roderic Guigó, and Tim J Hubbard. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*, 22(9):1760–1774, sep 2012.

[HG10] J. Robert Hogg and Stephen P. Goff. Upf1 senses 3'UTR length to potentiate mRNA decay. *Cell*, 143(3):379–389, 2010.

[HKDT13] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, apr 2013.

[HKT+16] David Hendrickson, David R. Kelley, Danielle Tenen, Bradley Bernstein, and John L. Rinn. Widespread RNA binding by chromatin-associated proteins. *Genome Biology*, 17(1):28, dec 2016.

[HLB+10] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcritpome wide identification of RNA binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2010.

[HLR+11] Jessica I Hoell, Erik Larsson, Simon Runge, Jeffrey D Nusbaum, Sujitha Duggimpudi, Thalia A Farazi, Markus Hafner, Arndt Borkhardt, Chris Sander, and Thomas Tuschl. RNA targets of wild-type and mutant FET family proteins. *Nature Structural & Molecular Biology*, 18(12):1428–1431, nov 2011.

[HPCO+15] Anthony K. Henras, Célia Plisson-Chastang, Marie Françoise O'Donohue, Anirban Chakraborty, and Pierre Emmanuel Gleizes. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 6(2):225–242, 2015.

[HPS+15] T. Hung, G. A. Pratt, B. Sundararaman, M. J. Townsend, C. Chaivorapol, T. Bhangale, R. R. Graham, W. Ortmann, L. A. Criswell, G. W. Yeo, and T. W. Behrens. The Ro60 autoantigen binds endogenous retroelements and regulates inflammatory gene expression. *Science*, 350(6259):455–459, 2015.

[HRB13] Jessica A. Hurt, Alex D. Robertson, and Christopher B. Burge. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Research*, 23(10):1636–1650, oct 2013.

[HSH+07] Nathaniel D Heintzman, Rhona K Stuart, Gary Hon, Yutao Fu, Christina W Ching, R David Hawkins, Leah O Barrera, Sara Van Calcar, Chunxu Qu, Keith a Ching, Wei Wang, Zhiping Weng, Roland D Green, Gregory E Crawford, and Bing Ren. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature genetics*, 39(3):311–8, mar 2007.

[HVA+12] Stephanie C. Huelga, Anthony Q. Vu, Justin D. Arnold, Tiffany D. Liang, Patrick P. Liu, Bernice Y. Yan, John Paul Donohue, Lily Shiue, Shawn Hoon, Sydney Brenner, Manuel Ares, and Gene W. Yeo. Integrative Genome-wide Analysis Reveals Cooperative Regulation of Alternative Splicing by hnRNP Proteins. *Cell Reports*, 1(2):167–178, feb 2012.

[IGK+08] Pavel V. Ivanov, Niels H. Gehring, Joachim B. Kunz, Matthias W. Hentze, and Andreas E. Kulozik. Interactions between UPF1, eRFs, PABP and the exon junction complex suggest an integrated model for mammalian NMD pathways. *EMBO Journal*, 27(5):736–747, mar 2008.

[IMA+13] Fadia Ibrahim, Manolis Maragkakis, Panagiotis Alexiou, Margaret a. Maronski, Marc a. Dichter, and Zissimos Mourelatos. Identification of In Vivo, Conserved, TAF15 RNA Binding Sites Reveals the Impact of TAF15 on the Neuronal Transcriptome. *Cell Reports*, 3(2):301–308, 2013.

[IZJ+14] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter L??nnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, dec 2014.

[Jeo17] Sunjoo Jeong. SR Proteins: Binders, Regulators, and Connectors of RNA. *Molecules and Cells*, 40(1):1–9, jan 2017.

[JLH+14] Youngsook L. Jung, Lovelace J. Luquette, Joshua Ho, Francesco Ferrari, Michael Tolstorukov, Aki Minoda, Robbyn Issner, Charles B. Epstein, Gary H. Karpen, Mitzi I. Kuroda, and Peter J. Park. Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Research*, 42(9):e74–e74, may 2014.

[JPV+09] Laure Jobert, Natalia Pinzón, Elodie Van Herreweghe, Beáta E Jády, Apostolia Guialis, Tamás Kiss, and László Tora. Human U1 snRNA forms a new chromatin-associated snRNP with TAF15. *EMBO reports*, 10(5):494–500, 2009.

[JT06] L. Joshua-Tor. The argonautes. *Cold Spring Harbor Symposia on Quantitative Biology*, 71(0):67–72, jan 2006.

[KAS+07] Tae Hoon Kim, Ziedulla K. Abdullaev, Andrew D. Smith, Keith A. Ching, Dmitri I. Loukinov, RolandD D. Green, Michael Q. Zhang, Victor V. Lobanenkov, and Bing Ren. Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome. *Cell*, 128(6):1231–1245, mar 2007.

[KBL+09] Jr Kwiatkowski, T J, D A Bosco, A L Leclerc, E Tamrazian, C R Vanderburg, C Russ, A Davis, J Gilchrist, E J Kasarskis, T Munsat, P Valdmanis, G A Rouleau, B A Hosler, P Cortelli, P J de Jong, Y Yoshinaga, J L Haines, M A Pericak-Vance, J Yan, N Ticozzi, T Siddique, D McKenna-Yasek, P C Sapp, H R Horvitz, J E Landers, and Jr Brown, R H. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science (New York, N.Y.)*, 323(5918):1205–1208, feb 2009.

[Kim16] D Y Kim. Post-transcriptional regulation of gene expression in neural stem cells. *Cell Biochem Funct*, 34(4):197–208, jun 2016.

[KJ12] Stephanie Kervestin and Allan Jacobson. NMD: A multifaceted response to premature translational termination. *Nature Reviews Molecular Cell Biology*, 13(11):700–712, oct 2012.

[KKF06] Jack D. Keene, Jordan M. Komisarow, and Matthew B. Friedersdorf. RIP-Chip: The isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Protocols*, 1(1):302–307, jun 2006.

[KKX+13] Ilmin Kwon, Masato Kato, Siheng Xiang, Leeju Wu, Pano Theodoropoulos, Hamid Mirzaei, Tina Han, Shanhai Xie, Jeffry L Corden, and Steven L McKnight. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, 155(5):1049–60, nov 2013.

[KLH+14] Tatsuaki Kurosaki, Wencheng Li, Mainul Hoque, Maximilian W L Popp, Dmitri N. Ermolenko, Bin Tian, and Lynne E. Maquat. A Post-Translational regulatory switch on UPF1 controls targeted mRNA degradation. *Genes and Development*, 28(17):1900–1916, sep 2014.

[KLL+14] Yarden Katz, Feifei Li, Nicole J. Lambert, Ethan S. Sokol, Wai Leong Tam, Albert W. Cheng, Edoardo M. Airoldi, Christopher J. Lengner, Piyush B. Gupta, Zhengquan Yu, Rudolf Jaenisch, and Christopher B. Burge. Musashi proteins are post-transcriptional regulators of the epithelial-luminal cell state. *eLife*, 3:e03915, nov 2014.

[KM13] Tatsuaki Kurosaki and Lynne E. Maquat. Rules that govern UPF1 binding to mRNA 3 UTRs. *Proceedings of the National Academy of Sciences*, 110(9):3357–3362, feb 2013.

[KPV+16] Katannya Kapeli, Gabriel A Pratt, Anthony Q Vu, Kasey R Hutt, Fernando J Martinez, Balaji Sundararaman, Ranjan Batra, Peter Freese, Nicole J Lambert, Stephanie C Huelga, Seung J Chun, Tiffany Y Liang, Jeremy Chang, John P Donohue, Lily Shiue, Jiayu Zhang, Haining Zhu, Franca Cambi, Edward Kasarskis, Shawn Hoon, Manuel Ares Jr, Christopher B Burge, John Ravits, Frank Rigo, and Gene W Yeo. Distinct and shared functions of ALS-associated proteins TDP-43,

FUS and TAF15 revealed by multisystem analyses. *Nature Communications*, 7:1–14, 2016.

[KRC+10] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS computational biology*, 6(7):e1000832, jan 2010.

[KRZ11] Mohsen Khorshid, Christoph Rodak, and Mihaela Zavolan. CLIPZ: A database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Research*, 39(SUPPL. 1):1–8, jan 2011.

[KSK14] Panagiota Kafasla, Antonis Skliris, and Dimitris L. Kontoyiannis. Post-transcriptional coordination of immunological responses by RNA-binding proteins. *Nature Immunology*, 15(6):492–502, may 2014.

[KSV+08] Peter Kühnlein, Anne-Dorte Sperfeld, Ben Vanmassenhove, Vivianna Van Deerlin, Virginia M-Y Lee, John Q Trojanowski, Hans A Kretzschmar, Albert C Ludolph, and Manuela Neumann. Two German kindreds with familial amyotrophic lateral sclerosis due to TARDBP mutations. *Archives of neurology*, 65(9):1185–9, sep 2008.

[KTP08] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12):1351–1359, dec 2008.

[KYI+06] Isao Kashima, Akio Yamashita, Natsuko Izumi, Naoyuki Kataoka, Ryo Morishita, Shinichi Hoshino, Mutsuhito Ohno, Gideon Dreyfuss, and Shigeo Ohno. Binding of a novel SMG-1-Upf1-eRF1-eRF3 complex (SURF) to the exon junction complex triggers Upf1 phosphorylation and nonsense-mediated mRNA decay. *Genes and Development*, 20(3):355–367, jan 2006.

[KZR+10] Julian König, Kathi Zarnack, Gregor Rot, Tomaž Tomaz Curk, Melis Kayikci, Blaž Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature structural & molecular biology*, 17(7):909–915, 2010.

[Lad16] A. N. Ladd. New Insights Into the Role of RNA-Binding Proteins in the Regulation of Heart Development. *International Review of Cell and Molecular Biology*, 324:125–185, 2016.

[LBHB11] Qunhua Li, James B. Brown, Haiyan Huang, and Peter J. Bickel. Measuring reproducibility of high-throughput experiments. *Annals of Applied Statistics*, 5(3):1752–1779, sep 2011.

[LDZ+13] Elisa Laurenti, Sergei Doulatov, Sasan Zandi, Ian Plumb, Jing Chen, Craig April, Jian Bing Fan, and John E. Dick. The transcriptional architecture of early human hematopoiesis identifies multilevel control of lymphoid commitment. *Nature Immunology*, 14(7):756–763, jul 2013.

[LGM+13] Michael T. Lovci, Dana Ghanem, Henry Marr, Justin Arnold, Sherry Gee, Marilyn Parra, Tiffany Y. Liang, Thomas J. Stark, Lauren T. Gehman, Shawn Hoon, Katlin B. Massirer, Gabriel A. Pratt, Douglas L. Black, Joe W. Gray, John G. Conboy, and Gene W. Yeo. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nature Structural and Molecular Biology*, 20(12):1434–1442, dec 2013.

[LHA14] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, pages 1–21, 2014.

[Li11] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, nov 2011.

[LJI13] Belinda Loh, Stefanie Jonas, and Elisa Izaurralde. The SMG5-SMG7 heterodimer directly recruits the CCR4-NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2. *Genes and Development*, 27(19):2125–2138, oct 2013.

[LKC+12] Gabriel B. Loeb, Aly a. Khan, David Canner, Joseph B. Hiatt, Jay Shendure, Robert B. Darnell, Christina S. Leslie, and Alexander Y. Rudensky. Transcriptome-wide miR-155 Binding Map Reveals Widespread Noncanonical MicroRNA Targeting. *Molecular Cell*, 48(5):760–770, 2012.

[LL79] Regine Losson and Francois Lacroute. Interference of nonsense mutations with eukaryotic messenger RNA stability [mapping/mRNA synthesis/oligo(dT)-cellulose/RNA-DNA hybridization]. *Cell Biology*, 76(10):5134–5137, oct 1979.

[LLZ+14] Jun Hao Li, Shun Liu, Hui Zhou, Liang Hu Qu, and Jian Hua Yang. StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Research*, 42(D1):D92–D97, jan 2014.

[LMF+08] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, Jennifer C Darnell, and Robert B Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, nov 2008.

[LMK+12] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B

Brown, Philip Cayting, Yiwen Chen, Gilberto DeSalvo, Charles Epstein, Katherine I Fisher-Aylor, Ghia Euskirchen, Mark Gerstein, Jason Gertz, Alexander J Hartemink, Michael M Hoffman, Vishwanath R Iyer, Youngsook L Jung, Subhradip Karmakar, Manolis Kellis, Peter V Kharchenko, Qunhua Li, Tao Liu, X Shirley Liu, Lijia Ma, Aleksandar Milosavljevic, Richard M Myers, Peter J Park, Michael J Pazin, Marc D Perry, Debasish Raha, Timothy E Reddy, Joel Rozowsky, Noam Shoresh, Arend Sidow, Matthew Slattery, John A Stamatoyannopoulos, Michael Y Tolstorukov, Kevin P White, Simon Xi, Peggy J Farnham, Jason D Lieb, Barbara J Wold, and Michael Snyder. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9):1813–31, sep 2012.

[LMT+11] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq). *Proceedings of the National Academy of Sciences*, 108(27):11063–11068, jul 2011.

[LPJC91] P. Leeds, S. W. Peltz, A. Jacobson, and M. R. Culbertson. The product of the yeast UPF1 gene is required for rapid turnover of mRNAs containing a premature translational termination codon. *Genes and Development*, 5(12 A):2303–2314, dec 1991.

[LPK17] Cole J.T. Lewis, Tao Pan, and Auinash Kalsotra. RNA modifications and structures cooperate to guide RNA-protein interactions. *Nature Reviews Molecular Cell Biology*, 18(3):202–210, feb 2017.

[LPM+15] Suzanne R. Lee, Gabriel A. Pratt, Fernando J. Martinez, Gene W. Yeo, and Jens Lykke-Andersen. Target Discrimination in Nonsense-Mediated mRNA Decay Requires Upf1 ATPase Activity. *Molecular Cell*, 59(3):413–425, aug 2015.

[LPTW15] Jonathan P. Ling, Olga Pletnikova, Juan C. Troncoso, and Philip C. Wong. TDP-43 repression of nonconserved cryptic exons is compromised in ALS-FTD. *Science*, 349(6248):650–655, aug 2015.

[LQLM10] X. Li, G. Quon, H. D. Lipshitz, and Q. Morris. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *Rna*, 16(6):1096–1107, jun 2010.

[LRH+09] R Lemmens, V Race, N Hersmus, G Matthijs, L Van Den Bosch, P Van Damme, B Dubois, S Boonen, A Goris, and W Robberecht. TDP-43 M311V mutation in familial amyotrophic lateral sclerosis. *Journal of neurology, neurosurgery, and psychiatry*, 80(3):354–5, mar 2009.

[LRJ+14] Nicole Lambert, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip a. Sharp, and Christopher B. Burge. RNA Bind-n-Seq: Quantitative Assessment of the

Sequence and Structural Binding Specificity of RNA Binding Proteins. *Molecular Cell*, 54(5):887–900, jun 2014.

[LSS14] Yang Liao, Gordon K Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–30, apr 2014.

[LTPC10] Clotilde Lagier-Tourenne, Magdalini Polymenidou, and Don W Cleveland. TDP-43 and FUS/TLS: emerging roles in RNA processing and neurodegeneration. *Human molecular genetics*, 19(R1):R46–64, apr 2010.

[LTPH+12] Clotilde Lagier-Tourenne, Magdalini Polymenidou, Kasey R Hutt, Anthony Q Vu, Michael Baughn, Stephanie C Huelga, Kevin M Clutario, Shuo-Chien Ling, Tiffany Y Liang, Curt Mazur, Edward Wancewicz, Aneeza S Kim, Andy Watt, Sue Freier, Geoffrey G Hicks, John Paul Donohue, Lily Shiue, C Frank Bennett, John Ravits, Don W Cleveland, and Gene W Yeo. Divergent roles of ALS-linked proteins FUS/TLS and TDP-43 intersect in processing long pre-mRNAs. *Nature Neuroscience*, 15(11):1488–1497, 2012.

[LTPS09] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, jan 2009.

[LZM+15] Qi Liu, Xue Zhong, Blair B. Madison, Anil K. Rustgi, and Yu Shyr. Assessing Computational Steps for CLIP-Seq Data Analysis. *BioMed Research International*, 2015:1–10, oct 2015.

[MANM16] Manolis Maragkakis, Panagiotis Alexiou, Tadashi Nakaya, and Zissimos Mourelatos. CLIPSeqToolsa novel bioinformatics CLIP-seq analysis suite. *Rna*, 22(1):1–9, jan 2016.

[Mar11] Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads, feb 2011.

[MFK03] Junsei Mimura and Yoshiaki Fujii-Kuriyama. Functional role of AhR in the expression of toxic effects by TCDD. *Biochimica et Biophysica Acta - General Subjects*, 1619(3):263–268, feb 2003.

[MHDN+00] H Müller-Hartmann, H Deissler, F Naumann, B Schmitz, J Schröer, and W Doerfler. The human 20-kDa 5'-(CGG)(n)-3'-binding protein is targeted to the nucleus and affects the activity of the FMR1 promoter. *The Journal of biological chemistry*, 275(9):6447–52, mar 2000.

[MKE13] Paul H. Miller, David J.H.F. Knapp, and Connie J. Eaves. Heterogeneity in hematopoietic stem cell populations. *Current Opinion in Hematology*, 20(4):257–264, jul 2013.

[MKPW14] Georgi K Marinov, Anshul Kundaje, Peter J Park, and Barbara J Wold. Large-Scale Quality Analysis of Published ChIP-seq Data. *G3: Genes—Genomes—Genetics*, 4(2):209–223, 2014.

[MKRR81] Lynne E. Maquat, Alan J. Kinniburgh, Eliezer A. Rachmilewitz, and Jeffrey Ross. Unstable β-globin mRNA in mRNA-deficient β0 thalassemia. *Cell*, 27(3 PART 2):543–553, 1981.

[MLCB14] Daniel Maticzka, Sita J. Lange, Fabrizio Costa, and Rolf Backofen. GraphProt: Modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1):R17, jan 2014.

[Moo05] M. J. Moore. From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science*, 309(5740):1514–1518, sep 2005.

[MPV+16] Fernando J. Martinez, Gabriel A. Pratt, Eric L. Van Nostrand, Ranjan Batra, Stephanie C. Huelga, Katannya Kapeli, Peter Freese, Seung J. Chun, Karen Ling, Chelsea Gelboin-Burkhart, Layla Fijany, Harrison C. Wang, Julia K. Nussbacher, Sara M. Broski, Hong Joo Kim, Rea Lardelli, Balaji Sundararaman, John P. Donohue, Ashkan Javaherian, Jens Lykke-Andersen, Steven Finkbeiner, C. Frank Bennett, Manuel Ares, Christopher B. Burge, J. Paul Taylor, Frank Rigo, and Gene W. Yeo. Protein-RNA Networks Regulated by Normal and ALS-Associated Mutant HNRNPA2B1 in the Nervous System. *Neuron*, 92(4):780–795, 2016.

[MPZ+15] Kate D. Meyer, Deepak P. Patil, Jun Zhou, Alexandra Zinoviev, Maxim A. Skabkin, Olivier Elemento, Tatyana V. Pestova, Shu Bing Qian, and Samie R. Jaffrey. 5 UTR m6A Promotes Cap-Independent Translation. *Cell*, 163(4):999–1010, nov 2015.

[MRM+17] Rafael G. Miranda, Margarita Rojas, Michael P. Montgomery, Kyle P. Gribbin, and Alice Barkan. RNA-binding specificity landscape of the pentatricopeptide repeat protein PPR10. *Rna*, 23(4):586–599, apr 2017.

[MS04] Stavroula Mili and Joan A Steitz. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA (New York, N.Y.)*, 10(11):1692–4, nov 2004.

[MSL+15] Michael J. Moore, Troels K.H. Scheel, Joseph M. Luna, Christopher Y. Park, John J. Fak, Eiko Nishiuchi, Charles M. Rice, and Robert B. Darnell. MiRNA-target chimeras reveal miRNA 3-end pairing as a major determinant of Argonaute target specificity. *Nature Communications*, 6:8864, nov 2015.

[NBD+11] Manuela Neumann, Eva Bentmann, Dorothee Dormann, Ali Jawaid, Mariely Dejesus-Hernandez, Olaf Ansorge, Sigrun Roeber, Hans a. Kretzschmar, David G. Munoz, Hirofumi Kusaka, Osamu Yokota, Lee Cyn Ang, Juan Bilbao, Rosa Rademakers, Christian Haass, and Ian R a MacKenzie. FET proteins TAF15

and EWS are selective markers that distinguish FTLD with FUS pathology from amyotrophic lateral sclerosis with FUS mutations. *Brain*, 134(9):2595–2609, 2011.

[NBLTY15] Julia K. Nussbacher, Ranjan Batra, Clotilde Lagier-Tourenne, and Gene W. Yeo. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends in Neurosciences*, 38(4):226–236, apr 2015.

[NFBK17] Cindo O. Nicholson, Matthew B. Friedersdorf, Laura S. Bisogno, and Jack D. Keene. DO-RIP-seq to quantify RNA binding sites transcriptome-wide. *Methods*, 118-119(1):16–23, jan 2017.

[Nis16] Kazuko Nishikura. A-to-I editing of coding and non-coding RNAs by ADARs. *Nature Reviews Molecular Cell Biology*, 17(2):83–96, dec 2016.

[NJF+16] Abhinav Nellore, Andrew E. Jaffe, Jean Philippe Fortin, José Alquicira-Hernández, Leonardo Collado-Torres, Siruo Wang, Robert A. Phillips, Nishika Karbhari, Kasper D. Hansen, Ben Langmead, and Jeffrey T. Leek. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biology*, 17(1):266, dec 2016.

[NJK+14] Pamela Nicholson, Christoph Josi, Hitomi Kurosawa, Akio Yamashita, and Oliver Mühlemann. A novel phosphorylation-independent interaction between SMG6 and UPF1 is essential for human NMD. *Nucleic Acids Research*, 42(14):9217–9235, aug 2014.

[NSK+06] Manuela Neumann, Deepak M. Sampathu, Linda K. Kwong, Adam C. Truax, Matthew C. Micsenyi, Thomas T. Chou, Jennifer Bruce, Theresa Schuck, Murray Grossman, Christopher M. Clark, Leo F. McCluskey, Bruce L. Miller, Eliezer Masliah, Ian R. Mackenzie, Howard Feldman, Wolfgang Feiden, Hans A. Kretzschmar, John Q. Trojanowski, and Virginia M-Y M.-Y. Lee. Ubiquitinated TDP-43 in frontotemporal lobar degeneration and amyotrophic lateral sclerosis. *Science*, 314(5796):130–3, oct 2006.

[NSL+11] Noa Novershtern, Aravind Subramanian, Lee N. Lawton, Raymond H. Mak, W. Nicholas Haining, Marie E. McConkey, Naomi Habib, Nir Yosef, Cindy Y. Chang, Tal Shay, Garrett M. Frampton, Adam C B Drake, Ilya Leskov, Bjorn Nilsson, Fred Preffer, David Dombkowski, John W. Evans, Ted Liefeld, John S. Smutko, Jianzhu Chen, Nir Friedman, Richard A. Young, Todd R. Golub, Aviv Regev, and Benjamin L. Ebert. Densely interconnected transcriptional circuits control cell states in human hematopoiesis. *Cell*, 144(2):296–309, jan 2011.

[OKYK+12] Yukiko Okada-Katsuhata, Akio Yamashita, Kei Kutsuzawa, Natsuko Izumi, Fumiki Hirahara, and Shigeo Ohno. N-and C-terminal Upf1 phosphorylations create binding platforms for SMG-6 and SMG-5:SMG-7 during NMD. *Nucleic Acids Research*, 40(3):1251–1266, feb 2012.

[ONT+12] Takako Ohyama, Takashi Nagata, Kengo Tsuda, Naohiro Kobayashi, Takao Imai, Hideyuki Okano, Toshio Yamazaki, and Masato Katahira. Structure of Musashi1 in a complex with target RNA: The role of aromatic stacking interactions. *Nucleic Acids Research*, 40(7):3218–3231, apr 2012.

[ORJ+12] Pawel K Olszewski, Jan Rozman, Josefin A Jacobsson, Birgit Rathkolb, Siv Strömberg, Wolfgang Hans, Anica Klockars, Johan Alsiö, Ulf Risérus, Lore Becker, Sabine M Hölter, Ralf Elvert, Nicole Ehrhardt, Valérie Gailus-Durner, Helmut Fuchs, Robert Fredriksson, Eckhard Wolf, Thomas Klopstock, Wolfgang Wurst, Allen S Levine, Claude Marcus, Martin Hrabě de Angelis, Martin Klingenspor, Helgi B Schiöth, and Manfred W Kilimann. Neurobeachin, a regulator of synaptic protein targeting, is associated with body fat mass and feeding behavior in mice and body-mass index in humans. *PLoS genetics*, 8(3):e1002568, jan 2012.

[PA93] R. Pulak and P. Anderson. mRNA Surveillance by the Caenorhabditis elegans smg genes. *Genes and Development*, 7(10):1885–1897, oct 1993.

[PBA17] Bruno Pereira, Marc Billaud, and Raquel Almeida. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer*, 3(7):506–528, jul 2017.

[PIB+12] Isabel Peixeiro, Ângela Inácio, Cristina Barbosa, Ana Luísa Silva, Stephen A. Liebhaber, and Luísa Romão. Interaction of PABPC1 with the translation initiation complex is critical to the NMD resistance of AUG-proximal nonsense mutations. *Nucleic Acids Research*, 40(3):1160–1173, feb 2012.

[PL13] Odil Porrua and Domenico Libri. RNA quality control in the nucleus: The Angels' share of RNA. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829(6-7):604–611, 2013.

[PLTH+11] Magdalini Polymenidou, Clotilde Lagier-Tourenne, Kasey R Hutt, Stephanie C Huelga, Jacqueline Moran, Tiffany Y Liang, Shuo-Chien Ling, Eveline Sun, Edward Wancewicz, Curt Mazur, Holly Kordasiewicz, Yalda Sedaghat, John Paul Donohue, Lily Shiue, C Frank Bennett, Gene W Yeo, and Don W Cleveland. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. *Nature neuroscience*, 14(4):459–468, 2011.

[PLTH+12] Magdalini Polymenidou, Clotilde Lagier-Tourenne, Kasey R. Hutt, C. Frank Bennett, Don W. Cleveland, and Gene W. Yeo. Misregulated RNA processing in amyotrophic lateral sclerosis. *Brain Research*, 1462:3–15, 2012.

[PTSC+12] Sharon L Paige, Sean Thomas, Cristi L Stoick-Cooper, Hao Wang, Lisa Maves, Richard Sandstrom, Lil Pabon, Hans Reinecke, Gabriel Pratt, Gordon Keller, Randall T Moon, John Stamatoyannopoulos, and Charles E Murry. A temporal chromatin signature in human embryonic stem cells identifies regulators of cardiac development. *Cell*, 151(1):221–32, sep 2012.

[PWA+12] Trevor J. Pugh, Shyamal Dilhan Weeraratne, Tenley C. Archer, Daniel A. Pomeranz Krummel, Daniel Auclair, James Bochicchio, Mauricio O. Carneiro, Scott L. Carter, Kristian Cibulskis, Rachel L. Erlich, Heidi Greulich, Heidi Greulich, Niall J. Lennon, Aaron Mc Kenna, James Meldrim, Alex H. Ramos, Michael G. Ross, Carsten Russ, Erica Shefler, Andrey Sivachenko, Brian Sogoloff, Petar Stojanov, Pablo Tamayo, Jill P. Mesirov, Vladimir Amani, Natalia Teider, Soma Sengupta, Jessica Pierre Francois, Paul A. Northcott, Michael D. Taylor, Furong Yu, Gerald R. Crabtree, Amanda G. Kautzman, Stacey B. Gabriel, Gad Getz, Natalie Jäger, David T W Jones, Peter Lichter, Stefan M. Pfister, Thomas M. Roberts, Jeffery L. Dangl, Scott L. Pomeroy, and Yoon Jae Cho. Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations, jul 2012.

[QH10] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, mar 2010.

[RCN+14] Frank Rigo, Seung J Chun, Daniel A Norris, Gene Hung, Sam Lee, John Matson, Robert A Fey, Hans Gaus, Yimin Hua, John S Grundy, Adrian R Krainer, Scott P Henry, and C Frank Bennett. Pharmacology of a central nervous system delivered 2'-O-methoxyethyl-modified survival of motor neuron splicing oligonucleotide in mice and nonhuman primates. *The Journal of pharmacology and experimental therapeutics*, 350(1):46–55, jul 2014.

[REB+12] Boris Rogelj, Laura E Easton, Gireesh K Bogu, Lawrence W Stanton, Gregor Rot, Tomaž Curk, Blaž Zupan, Yoichiro Sugimoto, Miha Modic, Nejc Haberman, James Tollervey, Ritsuko Fujii, Toru Takumi, Christopher E Shaw, and Jernej Ule. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Scientific reports*, 2:603, jan 2012.

[RHK+14] Oliver Rossbach, Lee-Hsueh Hung, Ekaterina Khrameeva, Silke Schreiner, Julian König, Tomaž Curk, Blaž Zupan, Jernej Ule, Mikhail S Gelfand, and Albrecht Bindereif. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biology*, 11(2):146–155, 2014.

[RKC+09] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, 27(7):667–670, jul 2009.

[RKC+13] Debashish Ray, Hilal Kazan, Kate B. Cook, Matthew T. Weirauch, Hamed S. Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, Hong Na, Manuel Irimia, Leah H. Matzat, Ryan K. Dale, Sarah A. Smith, Christopher A. Yarosh, Seth M. Kelly, Behnam Nabet, Desirea Mecenas, Weimin Li, Rakesh S. Laishram, Mei Qiao, Howard D. Lipshitz, Fabio Piano, Anita H. Corbett, Russ P. Carstens, Brendan J. Frey, Richard A. Anderson, Kristen W. Lynch, Luiz O.F. Penalva, Elissa P. Lei, Andrew G. Fraser, Benjamin J. Blencowe, Quaid D. Morris,

and Timothy R. Hughes. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177, jul 2013.

[RMW+11] Alan E. Renton, Elisa Majounie, Adrian Waite, Javier Simón-Sánchez, Sara Rollinson, J. Raphael Gibbs, Jennifer C. Schymick, Hannu Laaksovirta, John C. van Swieten, Liisa Myllykangas, Hannu Kalimo, Anders Paetau, Yevgeniya Abramzon, Anne M. Remes, Alice Kaganovich, Sonja W. Scholz, Jamie Duckworth, Jinhui Ding, Daniel W. Harmer, Dena G. Hernandez, Janel O. Johnson, Kin Mok, Mina Ryten, Danyah Trabzuni, Rita J. Guerreiro, Richard W. Orrell, James Neal, Alex Murray, Justin Pearson, Iris E. Jansen, David Sondervan, Harro Seelaar, Derek Blake, Kate Young, Nicola Halliwell, Janis Bennion Callister, Greg Toulson, Anna Richardson, Alex Gerhard, Julie Snowden, David Mann, David Neary, Michael A. Nalls, Terhi Peuralinna, Lilja Jansson, Veli Matti Isoviita, Anna Lotta Kaivorinne, Maarit Hölttä-Vuori, Elina Ikonen, Raimo Sulkava, Michael Benatar, Joanne Wuu, Adriano Chiò, Gabriella Restagno, Giuseppe Borghero, Mario Sabatelli, David Heckerman, Ekaterina Rogaeva, Lorne Zinman, Jeffrey D. Rothstein, Michael Sendtner, Carsten Drepper, Evan E. Eichler, Can Alkan, Ziedulla Abdullaev, Svetlana D. Pack, Amalia Dutra, Evgenia Pak, John Hardy, Andrew Singleton, Nigel M. Williams, Peter Heutink, Stuart Pickering-Brown, Huw R. Morris, Pentti J. Tienari, and Bryan J. Traynor. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron*, 72(2):257–268, oct 2011.

[RZB+08] Nicola J Rutherford, Yong-Jie Zhang, Matt Baker, Jennifer M Gass, Nicole A Finch, Ya-Fei Xu, Heather Stewart, Brendan J Kelley, Karen Kuntz, Richard J P Crook, Jemeen Sreedharan, Caroline Vance, Eric Sorenson, Carol Lippa, Eileen H Bigio, Daniel H Geschwind, David S Knopman, Hiroshi Mitsumoto, Ronald C Petersen, Neil R Cashman, Mike Hutton, Christopher E Shaw, Kevin B Boylan, Bradley Boeve, Neill R Graff-Radford, Zbigniew K Wszolek, Richard J Caselli, Dennis W Dickson, Ian R Mackenzie, Leonard Petrucelli, and Rosa Rademakers. Novel mutations in TARDBP (TDP-43) in patients with familial amyotrophic lateral sclerosis. *PLoS genetics*, 4(9):e1000193, jan 2008.

[SBT+08] Jemeen Sreedharan, Ian P Blair, Vineeta B Tripathi, Xun Hu, Caroline Vance, Boris Rogelj, Steven Ackerley, Jennifer C Durnall, Kelly L Williams, Emanuele Buratti, Francisco Baralle, Jacqueline de Belleroche, J Douglas Mitchell, P Nigel Leigh, Ammar Al-Chalabi, Christopher C Miller, Garth Nicholson, and Christopher E Shaw. TDP-43 mutations in familial and sporadic amyotrophic lateral sclerosis. *Science (New York, N.Y.)*, 319(5870):1668–72, mar 2008.

[SCP14] Jacob C. Schwartz, Thomas R. Cech, and Roy R. Parker. Biochemical Properties and Biological Functions of FET Proteins. *Annual Review of Biochemistry*, 84(1):141210135300003, 2014.

[SEP+12] Jacob C. Schwartz, Christopher C. Ebmeier, Elaine R. Podell, Joseph Heimiller, Dylan J. Taatjes, and Thomas R. Cech. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes and Development*, 26(24):2690–2695, 2012.

[SHS17] Tom Smith, Andreas Heger, and Ian Sudbery. UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, 27(3):491–499, mar 2017.

[SLA+14] Ian M. Silverman, Fan Li, Anissa Alexander, Loyal Goff, Cole Trapnell, John L. Rinn, and Brian D. Gregory. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Genome Biology*, 15(1):R3, jan 2014.

[SLL+16] Aarti Sharma, Alexander K Lyashchenko, Lei Lu, Sara Ebrahimi Nasrabady, Margot Elmaleh, Monica Mendelsohn, Adriana Nemes, Juan Carlos Tapia, George Z Mentis, and Neil A Shneider. ALS-associated mutant FUS induces selective motor neuron degeneration through toxic gain of function. *Nature communications*, 7:10465, jan 2016.

[SLQ+15] Shuying Sun, Shuo-Chien Ling, Jinsong Qiu, Claudio P Albuquerque, Yu Zhou, Seiya Tokunaga, Hairi Li, Haiyan Qiu, Anh Bui, Gene W Yeo, Eric J Huang, Kevin Eggan, Huilin Zhou, Xiang-Dong Fu, Clotilde Lagier-Tourenne, and Don W Cleveland. ALS-causative mutations in FUS/TLS confer gain and loss of function by altered association with SMN and U1-snRNP. *Nature communications*, 6:6171, 2015.

[SM00] Aaron J. Shatkin and James L. Manley. The ends of the affair: Capping and polyadenylation. *Nature Structural Biology*, 7(10):838–842, oct 2000.

[SQWVZ17] Ankeeta Shah, Yingzhi Qian, Sebastien M. Weyn-Vanhentenryck, and Chaolin Zhang. CLIP Tool Kit (CTK): A flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, 33(4):566–567, oct 2017.

[SRLA08] Guramrit Singh, Indrani Rebbapragada, and Jens Lykke-Andersen. A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay. *PLoS Biology*, 6(4):860–871, apr 2008.

[SRZ+13] Christoph Schweingruber, Simone C. Rufener, David Zünd, Akio Yamashita, and Oliver Mühlemann. Nonsense-mediated mRNA decay - Mechanisms of substrate mRNA recognition and degradation in mammalian cells. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1829(6-7):612–623, feb 2013.

[SS87] M B Shapiro and P Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic acids research*, 15(17):7155–74, sep 1987.

[SSS+12] Cem Sievers, Tommy Schlumpf, Ritwick Sawarkar, Federico Comoglio, and Renato Paro. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Research*, 40(20), 2012.

[STK+14] Chantelle F Sephton, Amy A Tang, Ashwinikumar Kulkarni, James West, Mieu Brooks, Jeremy J Stubblefield, Yun Liu, Michael Q Zhang, Carla B Green, Kimberly M Huber, Eric J Huang, Joachim Herz, and Gang Yu. Activity-dependent FUS dysregulation disrupts synaptic homeostasis. *Proceedings of the National Academy of Sciences of the United States of America*, 111(44):E4769–78, nov 2014.

[SZB+16] Balaji Sundararaman, Lijun Zhan, Steven M. Blue, Rebecca Stanton, Keri Elkins, Sara Olson, Xintao Wei, Eric L. Van Nostrand, Gabriel A. Pratt, Stephanie C. Huelga, Brendan M. Smalec, Xiaofeng Wang, Eurie L. Hong, Jean M. Davidson, Eric Lécuyer, Brenton R. Graveley, and Gene W. Yeo. Resources for the Comprehensive Discovery of Functional RNA Elements. *Molecular Cell*, 61(6):903–913, mar 2016.

[TCLK00] S a Tenenbaum, C C Carson, P J Lager, and J D Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26):14085–90, dec 2000.

[TCR+11] James R Tollervey, Tomaž Curk, Boris Rogelj, Michael Briese, Matteo Cereda, Melis Kayikci, Julian König, Tibor Hortobágyi, Agnes L Nishimura, Vera Zupunski, Rickie Patani, Siddharthan Chandran, Gregor Rot, Blaž Zupan, Christopher E Shaw, and Jernej Ule. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature neuroscience*, 14(4):452–8, apr 2011.

[TDH+09] Derek J Taylor, Batsal Devkota, Andrew D Huang, Maya Topf, Eswar Narayanan, Andrej Sali, Stephen C Harvey, and Joachim Frank. Comprehensive Molecular Structure of the Eukaryotic Ribosome. *Structure*, 17(12):1591–1604, dec 2009.

[TGE+13] Christer Thomsen, Pernilla Grundevik, Per Elias, Anders Ståhlberg, and Pierre Aman. A conserved N-terminal motif is required for complex formation between FUS, EWSR1, TAF15 and their oncogenic fusion proteins. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 27(12):4965–74, dec 2013.

[Tij05] N. Tijet. The aryl hydrocarbon receptor regulates distinct dioxin-dependent and dioxin-independent gene batteries. *Molecular Pharmacology*, 69(1):140–53, jan 2005.

[TTPMW06] W H Davin Townley-Tilson, Sarah A Pendergrass, William F Marzluff, and Michael L Whitfield. Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. *RNA*, 12(10):1853–1867, oct 2006.

[TVL+11] N Ticozzi, C Vance, A L Leclerc, P Keagle, J D Glass, D McKenna-Yasek, P C Sapp, V Silani, D A Bosco, C E Shaw, R H Brown, and J E Landers. Mutational analysis reveals the FUS homolog TAF15 as a candidate gene for familial amyotrophic lateral sclerosis. *American journal of medical genetics. Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics*, 156B(3):285–90, apr 2011.

[UBSB+12] Philip J. Uren, Emad Bahrami-Samani, Suzanne C. Burns, Mei Qiao, Fedor V. Karginov, Emily Hodges, Gregory J. Hannon, Jeremy R. Sanford, Luiz O.F. Penalva, and Andrew D. Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–3020, dec 2012.

[UCK+12] Andreas Untergasser, Ioana Cutcutache, Triinu Koressaar, Jian Ye, Brant C Faircloth, Maido Remm, and Steven G Rozen. Primer3–new capabilities and interfaces. *Nucleic acids research*, 40(15):e115, aug 2012.

[UiHI+02] Naoyuki Uchida, Shin ichi Hoshino, Hiroaki Imataka, Nahum Sonenberg, and Toshiaki Katada. A novel role of the mammalian GSPT/eRF3 associating with poly(A)-binding protein in cap/poly(A)-dependent translation. *Journal of Biological Chemistry*, 277(52):50286–50292, dec 2002.

[UJR+03] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, 2003.

[VGBW+17] Eric L. Van Nostrand, Chelsea Gelboin-Burkhart, Ruth Wang, Gabriel A. Pratt, Steven M. Blue, and Gene W. Yeo. CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods*, 118-119:50–59, 2017.

[VNGB+17] Eric L. Van Nostrand, Thai B. Nguyen, Chelsea Gelboin-Burkhart, Ruth Wang, Steven M. Blue, Gabriel A. Pratt, Ashley L. Louie, and Gene W. Yeo. Robust, cost-effective profiling of RNA binding protein targets with single-end enhanced crosslinking and immunoprecipitation (SeCLIP). *Methods in Molecular Biology*, 1648:177–200, 2017.

[VPS+91] Annemiske J Verkerk, Maura Pieretti, James S. Sutcliffe, Ying Hui Fu, Derek P.A. Kuhl, Antonio Pizzuti, Orly Reiner, Stephen Richards, Maureen F. Victoria, Fuping Zhang, Bert E. Eussen, Gert Jan B. van Ommen, Lau A.J. Blonden, Gregory J. Riggins, Jane L. Chastain, Catherine B. Kunst, Hans Galjaard, C. Thomas Caskey, David L. Nelson, Ben A. Oostra, and Stephen T. Warran. Identification of a gene

(FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65(5):905–914, may 1991.

[VPS+16] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, Rebecca Stanton, Frank Rigo, Mitchell Guttman, and Gene W Yeo. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514, jun 2016.

[VRH+09] Caroline Vance, Boris Rogelj, Tibor Hortobágyi, Kurt J De Vos, Agnes Lumi Nishimura, Jemeen Sreedharan, Xun Hu, Bradley Smith, Deborah Ruddy, Paul Wright, Jeban Ganesalingam, Kelly L Williams, Vineeta Tripathi, Safa Al-Saraj, Ammar Al-Chalabi, P Nigel Leigh, Ian P Blair, Garth Nicholson, Jackie de Belleroche, Jean-Marc Gallo, Christopher C Miller, and Christopher E Shaw. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science (New York, N.Y.)*, 323(5918):1208–1211, feb 2009.

[VSN+13] Caroline Vance, Emma L Scotter, Agnes L Nishimura, Claire Troakes, Jacqueline C Mitchell, Claudia Kathe, Hazel Urwin, Catherine Manser, Christopher C Miller, Tibor Hortobágyi, Mike Dragunow, Boris Rogelj, and Christopher E Shaw. ALS mutant FUS disrupts nuclear localization and sequesters wild-type FUS within cytoplasmic stress granules. *Human molecular genetics*, 22(13):2676–88, jul 2013.

[VW11] Ambro Van Hoof and Eric J. Wagner. A brief survey of mRNA surveillance. *Trends in Biochemical Sciences*, 36(11):585–592, 2011.

[WCK+14] Tao Wang, Beibei Chen, MinSoo Kim, Yang Xie, and Guanghua Xiao. A model-based approach to identify binding sites in CLIP-seq data. *PLoS ONE*, 9(4), 2014.

[WGB+14] Charles Wang, Binsheng Gong, Pierre R Bushel, Jean Thierry-Mieg, Danielle Thierry-Mieg, Joshua Xu, Hong Fang, Huixiao Hong, Jie Shen, Zhenqiang Su, Joe Meehan, Xiaojin Li, Lu Yang, Haiqing Li, Paweł P Łabaj, David P Kreil, Dalila Megherbi, Stan Gaj, Florian Caiment, Joost van Delft, Jos Kleinjans, Andreas Scherer, Viswanath Devanarayan, Jian Wang, Yong Yang, Hui-Rong Qian, Lee J Lancashire, Marina Bessarabova, Yuri Nikolsky, Cesare Furlanello, Marco Chierici, Davide Albanese, Giuseppe Jurman, Samantha Riccadonna, Michele Filosi, Roberto Visintainer, Ke K Zhang, Jianying Li, Jui-Hua Hsieh, Daniel L Svoboda, James C Fuscoe, Youping Deng, Leming Shi, Richard S Paules, Scott S Auerbach, and Weida Tong. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nature Biotechnology*, 32(9):926–932, aug 2014.

[WHK+12] Melissa L Wilbert, Stephanie C Huelga, Katannya Kapeli, Thomas J Stark, Tiffany Y Liang, Stella X Chen, Bernice Y Yan, Jason L Nathanson, Kasey R Hutt, Michael T Lovci, Hilal Kazan, Anthony Q Vu, Katlin B Massirer, Quaid Morris, Shawn Hoon, and Gene W Yeo. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Molecular cell*, 48(2):195–206, oct 2012.

[WK12] Lucas D. Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature Biotechnology*, 30(11):1095–1106, nov 2012.

[WLW+11] Lili Wang, Michael S. Lawrence, Youzhong Wan, Petar Stojanov, Carrie Sougnez, Kristen Stevenson, Lillian Werner, Andrey Sivachenko, David S. DeLuca, Li Zhang, Wandi Zhang, Alexander R. Vartanov, Stacey M. Fernandes, Natalie R. Goldstein, Eric G. Folco, Kristian Cibulskis, Bethany Tesar, Quinlan L. Sievers, Erica Shefler, Stacey Gabriel, Nir Hacohen, Robin Reed, Matthew Meyerson, Todd R. Golub, Eric S. Lander, Donna Neuberg, Jennifer R. Brown, Gad Getz, and Catherine J. Wu. SF3B1 and Other Novel Cancer Genes in Chronic Lymphocytic Leukemia. *New England Journal of Medicine*, 365(26):2497–2506, dec 2011.

[WRV80] Anton J.M. Wagenmakers, Rita J. Reinders, and Walther J. Van Venrooij. Crosslinking of mRNA to Proteins by Irradiation of Intact Cells with Ultraviolet Light. *European Journal of Biochemistry*, 112(2):323–330, nov 1980.

[WSL+08] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–6, nov 2008.

[WSWS12] Xiaowei Wang, Athanasia Spandidos, Huajun Wang, and Brian Seed. PrimerBank: a PCR primer database for quantitative gene expression analysis, 2012 update. *Nucleic acids research*, 40(Database issue):D1144–9, jan 2012.

[WU11] Joshua T. Witten and Jernej Ule. Understanding splicing regulation through RNA splicing maps, mar 2011.

[WVMY+14] Sebastien M. Weyn-Vanhentenryck, Aldo Mele, Qinghong Yan, Shuying Sun, Natalie Farny, Zuo Zhang, Chenghai Xue, Margaret Herre, Pamela A. Silver, Michael Q. Zhang, Adrian R. Krainer, Robert B. Darnell, and Chaolin Zhang. HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Reports*, 6(6):1139–1152, 2014.

[WWC+12] Troy W. Whiteld, Jie Wang, Patrick J. Collins, E. Christopher Partridge, Shelley F. Aldred, Nathan D. Trinklein, Richard M. Myers, and Zhiping Weng. Functional

analysis of transcription factor binding sites in human promoters. *Genome Biology*, 13(9):R50, sep 2012.

[WXC+15] Tao Wang, Guanghua Xiao, Yongjun Chu, Michael Q. Zhang, David R. Corey, and Yang Xie. Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Research*, 43(11):5263–5274, jun 2015.

[WXX14] Tao Wang, Yang Xie, and Guanghua Xiao. dCLIP: a computational approach for comparative CLIP-seq analyses. *Genome Biology*, 15(1):R11, 2014.

[WZI+12] Jie Wang, Jiali Zhuang, Sowmya Iyer, Xin Ying Lin, Troy W. Whitfield, Melissa C. Greven, Brian G. Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, Oliver J. Rando, Ewan Birney, Richard M. Myers, William S. Noble, Michael Snyder, and Zhiping Weng. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, 22(9):1798–1812, sep 2012.

[YCL+09] Gene W. Yeo, Nicole G. Coufal, Tiffany Y. Liang, Grace E. Peng, Xiang Dong Fu, and Fred H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural and Molecular Biology*, 16(2):130–137, feb 2009.

[YCY+12] Tomohiro Yamazaki, Shi Chen, Yong Yu, Biao Yan, Tyler C Haertlein, Monica A Carrasco, Juan C Tapia, Bo Zhai, Rita Das, Melanie Lalancette-Hebert, Aarti Sharma, Siddharthan Chandran, Gareth Sullivan, Agnes Lumi Nishimura, Christopher E Shaw, Steve P Gygi, Neil A Shneider, Tom Maniatis, and Robin Reed. FUS-SMN protein interactions link the motor neuron diseases ALS and SMA. *Cell reports*, 2(4):799–806, oct 2012.

[YLS+11] Jian Hua Yang, Jun Hao Li, Peng Shao, Hui Zhou, Yue Qin Chen, and Liang Hu Qu. StarBase: A database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Research*, 39(SUPPL. 1):D202–D209, jan 2011.

[YZK+15] Liuqing Yang, Jiayu Zhang, Marisa Kamelgarn, Chunyan Niu, Jozsef Gal, Weimin Gong, and Haining Zhu. Subcellular localization and RNAs determine FUS architecture in different cellular compartments. *Human molecular genetics*, 24(18):5174–83, sep 2015.

[ZD11] Chaolin Zhang and Robert B Darnell. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nature biotechnology*, 29(7):607–614, 2011.

[ZFS+16] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. IrCLIP platform for efficient characterization of protein-RNA interactions. *Nature Methods*, 13(6):489–492, apr 2016.

[ZGZM13] David Zünd, Andreas R. Gruber, Mihaela Zavolan, and Oliver Mühlemann. Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3 UTRs. *Nature Structural and Molecular Biology*, 20(8):936–943, jul 2013.

[ZKT+13] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152(3):453–466, jan 2013.

[ZRS+08] Zhengdong D Zhang, Joel Rozowsky, Michael Snyder, Joseph Chang, and Mark Gerstein. Modeling ChIP sequencing in silico with applications. *PLoS computational biology*, 4(8):e1000158, jan 2008.