UNIVERSITY OF CALIFORNIA SAN DIEGO

High Throughput Discovery of Gene Function in Disease: Image-based CRISPR screens and Modeling Splicing Factor Mutations in Cancer

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

in

Biomedical Sciences

by

Emily Chandler Wheeler

Committee in charge:

Professor Gene Yeo, Chair
Professor Rafael Bejar
Professor Amy Pasquinelli
Professor Pablo Tamayo
Professor Jean Wang

2020

The dissertation of Emily Chandler Wheeler is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To:

To Mom and Dad. My success would not be made possible without your never-ending love, support, and encouragement. Thank you for making this journey possible for me.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

This body of work is the result of my graduate school efforts to become a cancer biologist with a focus on RNA biology and informatics. I have a tremendous amount of gratitude for the mentorship and training from my thesis advisor, Gene Yeo. Gene has carefully guided me along this path, listened to my goals and interests, pushed me to tackle new challenges, provided countless scientific opportunities, and most importantly, trusted me to drive my projects to completion, no matter the pace. I am grateful for his patience and diligence in guiding me to become the scientist I am today. My success has been made possible by the vibrant community of Yeo Lab members, past and present, who have carried me along through all these years. In particular, I want to thank Anthony Vu being a fantastic co-leader of our work and teaching me how to become a better collaborator along the way. I am grateful for the many hours of training I received from Olga Botvinnik and Eric Van Nostrand, both of whom played an instrumental role in educating me on the nuances of programming and analyzing high throughput sequencing data. I cannot speak enough to the incredible environment I have been a part of at UC San Diego for the last 6 years. I am grateful to all the professors, colleagues, and friends who have supported me along the way and are forever a part of my community. And most importantly, I am thankful for the unwavering support of my family. Mom, Dad, and John have always been first to celebrate my successes, and the impact of their love in support cannot be understated in the completion of this PhD. To everyone who has been with me on this journey, thank you.

Chapter 2, in full, is a reprint of material as it appears in: **Wheeler, EC.***, Vu AQ*, Einstein JM, DiSalvo M, Ahmed N, Van Nostrand EL, Shishkin AA, Jin W, Allbritton NA, Yeo GW. Pooled CRISPR screens with imaging on microraft arrays reveals stress granule regulatory factors. Nat

Methods, doi:10.1038/s41592-020-0826-8 (2020). The dissertation author was first author of this paper.

Chapter 3, in full, is a reprint of material as it appears in: **Wheeler, EC.***, Van Nostrand EL, Yeo GW. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley Interdiscip Rev RNA 9, doi:10.1002/wrna.1436 (2018). The dissertation author was first author of this paper.

Chapter 4 in part is a reprint of material submitted to be published in: **Wheeler, EC.***,Vora S.*, Kotini A, Mayer D, Olszewaska M, Park SS, Sunahara RK, Yeo GW*, Papapetrou EP*. *SRSF2* and *U2AF1* mutations drive MDS through alternate *GNAS* isoform usage and cooperate with *GNAS* hotspot mutations. *in preparation*. The dissertation author was the primary author.

VITA

2014	Bachelor of Science in Environmental Science, Indiana University
2020	Doctor of Philosophy in Biomedical Sciences, University of California San Diego

PUBLICATIONS

Hin H, **Wheeler E**, Phillips R. Root-induced changes in nutrient cycling in forests depends on exudation rates. Soil Biology and Biochemistry 78, 213-221 (2014)

Washburn M, Kakaradov B, Sundararaman B, **Wheeler E**, Hoon S, Yeo G, Hundley H. The dsRBP and inactive editor ADR-1 utilizes dsRNA binding to regulate A-to-I RNA editing across the C. elegans transcriptome. Cell Rep 6, 599-607, doi:10.1016/j.celrep.2014.01.011 (2014)

Wheeler EC, Washburn M, Major F, Rusch D, Hundley H. Noncoding regions of C. elegans mRNA undergo selective adenosine to inosine deamination and contain a small number of editing sites per transcript. RNA Biol 12, 162-174, doi:10.1080/15476286.2015.1017220 (2015)

Batra R, Stark TJ, Clark E, Belzile JP, **Wheeler EC**, Yee BA, Huang H, Gelboin-Burkhark C, Huelga SC, Aigner S, Roberts BT, Bod TJ, Sathe S, Donohue JP, Rigo F, Ares M Jr, Spector DH, Yeo GW. RNA-binding protein CPEB1 remodels host and viral RNA landscapes. Nat Struct Mol Biol 23, 1101-1110, doi:10.1038/nsmb.3310 (2016).

Carter H, Marty R, Hofree M, Gross A, Jensen J, Fisch KM, Wu X, DeBoever C, Van Nostrand EL, Song Y, **Wheeler E**, Kreisberg JF, Lippman SF, Yeo G, Gutkind JS, Ideker T. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. Cancer Discov 7, 410-423, doi:10.1158/2159-8290.CD-16-1045 (2017)

Deffit SN, Yee BA, Vadlamani P, Manning AC, **Wheeler EC** Domissy A, Washburn MC, Yeo GW, Hundley HA. The C. elegans neural editome reveals an ADAR target mRNA required for proper chemotaxis. Elife 6, doi:10.7554/eLife.28625 (2017)

Krach F, Batra R, **Wheeler EC**, Vu AQ, Wang R, Hutt K, Rabin SJ, Baughn MW, Libbt RT, Diaz-Garcia S, Stauffer J, Pirie E, Saberi S, Rodriguez M, Madrigal AA, Kohl Z, Winner B, Yeo GW, Ravits J. Transcriptome-pathology correlation identifies interplay between TDP-43 and the expression of its kinase CK1E in sporadic ALS. Acta Neuropathol 136, 405-423, doi:10.1007/s00401-018-1870-7 (2018)

Wheeler EC, Van Nostrand EL, Yeo GW. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley Interdiscip Rev RNA 9, doi:10.1002/wrna.1436 (2018)

Pollyea DA, Harris C, Rabe JL, Hedin BR, De Arras L, Katz S, **Wheeler E**, Bejar R, Walter M, Jordan CT, Pietras EM, Alper S. Myelodysplastic syndrome-associated spliceosome gene mutations enhance innate immune signaling. Haematologica 104, e388-e392, doi:10.3324/haematol.2018.214155 (2019)

Ganem NS, Ben-Asher N, Manning AC, Deffit SN, Washburn MC, **Wheeler EC**, Yeo GW, Zgayer OB, Mantsur E, Hundley HA, Lamm AT. Disruption in A-to-I Editing Levels Affects C. elegans Development More Than a Complete Lack of Editing. Cell Rep 27, 1244-1253 e1244, doi:10.1016/j.celrep.2019.03.095 (2019)

Tan FE, Sathe S, **Wheeler EC**, Nussbacher JK, Peter S, Yeo GW. A Transcriptome-wide Translational Program Defined by LIN28B Expression Level. Mol Cell 73, 304-313 e303, doi:10.1016/j.molcel.2018.10.041 (2019)

Van Nostrand EL, Pratt GA, Yee BA, **Wheeler EC**, Blue SM, Mueller J, Park SS, Garcia KE, Gelboin-Burkhart C, Nguyen TB, Rabano I, Stanton R, Sundararaman B, Wang R, Fu XD, Graveley BR, Yeo GW. Principles of RNA Processing From Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins. Genome Biology. 21(1):90, doi:10.1186/s13059-020-01982-9 (2020)

Wheeler, EC.*, Vu AQ*, Einstein JM, DiSalvo M, Ahmed N, Van Nostrand EL, Shishkin AA, Jin W, Allbritton NA, Yeo GW. Pooled CRISPR screens with imaging on microraft arrays reveals stress granule regulatory factors. Nat Methods, doi:10.1038/s41592-020-0826-8 (2020)

Wheeler, EC.*,Vora S.*, Kotini A, Mayer D, Olszewaska M, Park SS, Sunahara RK, Yeo GW*, Papapetrou EP*. *SRSF2* and *U2AF1* mutations drive MDS through alternate *GNAS* isoform usage and cooperate with *GNAS* hotspot mutations. *in preparation*.

ABSTRACT OF THE DISSERTATION

High Throughput Discovery of Gene Function in Disease: Image-based CRISPR screens and Modeling Splicing Factor Mutations in Cancer

by

Emily Chandler Wheeler

Doctor of Philosophy in Biomedical Sciences

University of California San Diego, 2020

Professor Gene Yeo, Chair

Cellular behavior is dictated by numerous factors including cell-type and environmental-dependent gene expression as well as somatically acquired mutations in cancer. The goals of my thesis research are to employ high throughput techniques and informatics to assay gene function in the regulation of disease-relevant phenotypes. First, we developed a screening platform to combine image-based readouts with the easy-to-use pooled CRISPR knockout approach. Pooled CRISPR methods allow for the genetic knock-out of tens of thousands of genes in a single pool of cells, but are limited thus far to bulk-selected phenotypes including prolif-

eration rates and FACS-based screening. In this work, we adapted a microraft array platform and developed automated imaging and analysis programs to screen over 1,000 RNA binding proteins (RBPs) for their role in regulating the abundance of cytoplasmic stress granules, a cellular phenotype with implications in neurodegeneration. Second, to understand the role of somatic mutations in splicing factor (SF) proteins to drive cancer phenotypes, we developed a human isogenic induced pluripotent stem cell (iPSC) model of the canonical *U2AF1 S34F* and *SRSF2 P95L* mutations. We found that iPSC-derived hematopoietic stem/progenitor cells (iPSC-HSPCs) have a reduced ability to terminally differentiate into mature blood cells, thus reproducing the cancer phenotype found in patients with Myelodysplastic Syndromes (MDS). An integrative RNA-Seq and eCLIP analysis these cells revealed an that alternative splicing of *GNAS* is a direct target of mutant SFs. Functional studies confirmed that expression of the cancer-associated isoform, *GNAS-L* reproduced our disease phenotype in normal cells, and correction of SF-mutant cells to *GNAS-S* rescued the disease phenotype back to normal levels. *GNAS* is a major regulator of cell signaling, and also contains point mutations in a subset of MDS patients. Mutational analysis of patient data and biochemical work on purified forms of this protein suggest usage of *GNAS-L* in addition to the *GNAS R201H* point mutation cooperate to induce overactivation of G-protein signaling in cancer cells. This work has identified a novel cellular function of *GNAS* isoform regulation, and opened up the door for G protein signaling as a potential therapeutic avenue for SF-mutant MDS.

Chapter 1

Introduction

1.1 Genetic Screening of Cellular Phenotypes

While all cells in an organism share the same genetic material, there are highly regulated mechanisms in place to control which genes encoded in the DNA are expressed as proteins in various cellular contexts. Current estimates suggest that the human genome encodes for 20,000 protein coding genes, and it is the cell-type specific expression of those genes that allow for unique functionalities (International Human Genome Sequencing Consortium 2004). Cellular behavior is incredibly complex, in which thousands of expressed genes control both the basics of cell survival as well as the specific behaviors that allow cells to carry out a unique function. Oftentimes, those unique functions are coopted in human disease, allowing cells to confer unique properties that result in unwanted disease phenotypes in humans. It is the goal of our medical community to develop a deep understanding of these abnormalities, so that appropriate therapeutics can be used and developed to offer treatments when such abnormalities arise.

One of the most powerful tools the research community has to determine which genes

are regulating specific behaviors are genetic screens. Forward genetic screens are methods used to induce genetic variation in search of specific conditions that regulate a phenotype of interest. These screens are particularly valuable in instances where the phenotype driving disease is well-defined and easy to assay in a laboratory setting. One example of this is cell growth which is known to be more rapid in cancer cells. Therefore, researchers have used genetic screening to evaluate which genes are responsible for driving cell division and growth. Other disease phenotypes are more complex, including those that can only be observed with imaging techniques to look inside the cell and measure the spatial distribution of protein, RNA, or organelle localization. In chapter 2 of my thesis, I will describe my work to develop a screening platform that allows researchers to easily perform a genetic screen, in high throughput, to assay for cellular behavior detected with imaging-based readouts.

1.2 CRISPR as a tool to induce genetic variation

Equally as important as our ability to evaluate cellular behavior in a laboratory setting is the ability to induce genetic mutation in a programmable fashion. Historically, screening was performed using toxins to randomly induce mutations in DNA (Bauer et al., 2015). Once a cell or organism with a desired behavior was identified, genetic mapping was used to find the specific location of alteration in the genome. With the advent of high throughput sequencing technologies and completion of the human genome project came a much greater ability to introduce and identify genetic variation in a systematic manner. The tool that has come to the forefront of programmable genetic mutation is CRISPR/Cas9, a DNA editor complex identified and adapted from bacterial cells (Jinek et al., 2012). With this tool, cells can be infected with a short guide RNA (gRNA) containing a specific sequence in the genome, where the Cas9 machinery will be

directed to make a cut in the DNA.

The utility of CRIPSR tools is constantly expanding and proving to be incredibly diverse. One of the most impactful adaptations of this tool is the invention of pooled screening approaches (Wang et al., 2014). Computational advances in the design of gRNAs and the use of array-based synthesis to produce hundreds of thousands of gRNA, targeting the genetic knockout of thousands of genes, made pooled CRISPR screening possible and relatively simple to implement in any laboratory setting. This tool has become standard practice to systematically deplete genes in high throughput to survey the role of each in driving cellular phenotypes.

1.3 Splicing Factor Mutations in Cancer

In addition to screening technologies to determine gene function, cells naturally provide evidence of gene function when protein products are commonly mutated in disease. Cells acquire many mutations over time, most of which are not harmful and do not alter the normal cellular functions. However, sometimes such mutations lead to a selective growth advantage or disease phenotype that presents in an individual seeking medical attention. With DNA sequencing technologies, we are now able to easily look at a sample of cancer cells, for instance, and identify the specific DNA changes that are present in those cells compared to normal.

In 2011 it was revealed that many cancer cells found in patients with myelodysplastic syndromes (MDS) contain specific mutations in splicing factor proteins (Yoshida et al., 2011). Splicing factors are the components of the large assembly, known as the spliceosome, that is a very tightly regulated protein-RNA complex that controls the removal of intronic sequences from pre-mRNAs. Surprisingly, sequencing efforts revealed that more than 50% of patients with this disease contained a mutation in a splicing factor protein (Bejar R and Steensma DP 2014).

It is the goal of my thesis work, to determine how those mutations are altering cell behavior. Chapter 4 of my thesis will describe the results of this work that have been completed in close collaboration with Eirini Papapetrou's lab.

One very useful tool to understand the behavior of RNA binding proteins (RBPs) is CLIP-seq (crosslinking followed by immunoprecipitation and sequencing) (Ule et al., 2003). This method has been developed over many years as an approach to identify all of the locations in the transcriptome where RBPs bind in a cell. In analyzing this type of data, we can determine the genes that are direct targets of these splicing factors of interest, and importantly, we can identify different genes that are differentially bound by the wild-type and mutant factors. Chapter 3 of this thesis will be dedicated to a review of CLIP technologies that have been dramatically improved upon over the last ten years.

1.4 Outline of the dissertation

I will begin this thesis with a description of the technology platform I developed to systematically screen gene function of an imaging-based phenotype (Chapter 2). This platform has widespread utility in the genetic screening field, and the results have provided insight into the regulation of stress granule formation, a cellular phenotype with implications in neurodegeneration. Next, I will review the details of technological advances, not made by myself, in another molecular technique, CLIP-seq (Chapter 3). These technical advances laid the groundwork for the second half of my thesis work that is aimed at understanding the role of splicing factor mutations in cancer. In collaboration with Eirini Papapetrou's lab, we developed an isogenic stem cell model of blood cells that contain the cancer-associated point mutations in splicing factor proteins. My analysis of alternative splicing and differential binding in these cells led to the new

finding that expression of a *GNAS-L* in cancer cells is driving the disease phenotype (Chapter 4). Each section contains it's own detailed introduction and conclusion on the specific subject matter.

Chapter 2

Pooled CRISPR screens with imaging on microraft arrays reveals stress granule-regulatory factors

Genetic screens using pooled CRISPR-based approaches are scalable and inexpensive, but restricted to standard readouts including survival, proliferation and sortable markers. However, many biologically relevant cell states involve cellular and subcellular changes that are only accessible by microscopic visualization, and are currently impossible to screen with pooled methods. Here we combine pooled CRISPR/Cas9 screening with microraft array technology and high-content imaging to screen image-based phenotypes (CRaft-ID; CRISPR-based microraft, followed by gRNA Identification). By isolating microrafts that contain genetic clones harboring individual guide RNAs, we identify RNA binding proteins (RBPs) that influence the formation of stress granules, punctate protein-RNA assemblies, that form during stress. To automate hit

identification, we developed a machine-learning model trained on nuclear morphology to remove unhealthy cells or imaging artifacts. In doing so, we identified and validated previously uncharacterized RBPs that modulate stress granule abundance, highlighting the applicability of our approach to facilitate image-based pooled CRISPR screens.

2.1 Introduction

CRISPR tools in genetic screening

Pooled genetic knockout screens are widely used by the functional genomics community to identify genes responsible for cellular phenotypes. However, these screens have been limited to bulk selection methods including growth rate (Shalem et al., 2014), synthetic lethality (Blomen et al. 2015) and reporter-based fluorescent sorting (DeJesus et al. 2016, Parnas et al., 2015). Recently, pooled methods combined with single-cell sequencing (Jaitin et al., 2016, Adamson et al., 2016, Dixit et al., 2016, Datlinger et al., 2017) allow for whole-transcriptome quantification following perturbation, enabling multi-dimensional analyses of molecular pathways associated with genetic alterations. While these methods have dramatically increased the throughput in genetic knock-out studies, they cannot assay subcellular phenotypes with the spatiotemporal resolution detected by imaging.

Subcellular phenotypes account for both physiological and pathological changes in cell identity and function, such as transcription factor translocation into the nucleus (Link et al., 2009), protein localization to cellular sub-structures (de Groot et al., 2018), or mis-localization of proteins into disease-associated aggregates (Maharana et al., 2018). More broadly, high-throughput imaging unbiasedly captures functional and morphological cell states (Caicedo et

al., 2016) that dictate response to various stimuli (Kiger et al., 2003, Liu et al., 2009). However, screening for regulators of these phenotypes is currently limited to arrayed methods that often require expensive robotic platforms. Technologies to integrate pooled screening with cellular and subcellular imaging readouts are critical to improve the throughput of image-based genetic knock-out studies. (Figure 2-1A). Recently, studies using sequencing in-situ with fluorescently-labeled nucleotides with pooled CRISPR libraries, in combination with image-based phenotyping, identify genetic regulators of transcription factor localization (Feldman et al., 2019) and long-noncoding RNA localization (Wang et al., 2019).

Here, we present a new method for pooled CRISPR screens (¿12,000 sgRNAs) on microraft arrays (Wang et al., 2010), followed by automated high-resolution confocal imaging to identify regulators of stress granules, which are cytoplasmic protein aggregates that form during cellular stress. microraft arrays are an attractive platform to screen bulk-infected cells because thousands of clonal cell colonies (5-20 cells per colony) can be cultured in isolation from one another after plating cells in limiting-dilution (Wang et al., 2010, DiSalvo et al., 2019, Gach et al., 2011). Though the micro-scale cell carriers ("rafts") are physically separated from one another on-array, they share a common media reservoir, eliminating artifacts that arise from manipulating hundreds or thousands of cell culture wells individually. And finally, single microrafts can be removed from the array allowing for extended culture or genomic analyses.

Stress Granules play a role in neurodegeneration

Stress granules are protein-RNA cytoplasmic foci that form transiently during cellular perturbations including oxidative stress, heat shock and immune activation (Kedersha et al., 2007). Aberrant stress granule dynamics have been linked to the pathobiology of human dis-

eases including cancer (Anderson et al., 2015, Grabocka et al., 2016) and neurodegeneration (Wolozin et al., 2019). To illustrate, mutations present in amyotrophic lateral sclerosis (ALS), a form of neurodegenerative disease, have been shown to alter stress granule dynamics and composition (Lee et al., 2016, Martinez et al., 2016). Proteomics approaches have identified proteins that localize to stress granules (Markmiller et al., 2018, Jain et al., 2016, Youn et al., 2018); however, many genes that affect stress granule abundance remain unidentified. Therefore, the identification of genetic modulators that regulate stress granule biology could lead to novel, disease-relevant therapies.

In this work, we developed CRaft-ID (CRISPR-based microRaft, followed by gRNA identification) to couple the power of image-based phenotyping of stress granules with an easy-to-use pooled CRISPR screening workflow on microraft arrays. (Figure 2-1B). We performed a bulk-infection of cells with a gRNA library targeting over 1,000 annotated RBPs (¿12,000 sgRNAs) followed by single-cell plating on 20 microraft arrays to screen 119,050 genetic knock-out clones for stress granule abundance. Notably, our gRNA library is the same design as those traditionally used for pooled-CRISPR screens and requires no library modifications, making this workflow amenable to existing CRISPR sgRNA libraries. We performed high-content confocal microscopy and developed machine learning tools to identify genetic clones with reduced stress granule abundance following CRISPR knock-out. Our screen identified and validated six previously known stress granule modulators, along with 17 new RBPs that, when depleted, reduce sodium arsenite-induced stress granules in human cells. This work illustrates the power of combining broadly applicable pooled CRISPR methods with microraft-enabled high-content imaging analysis to identify genetic factors that affect subcellular phenotypes.

2.2 Results and Discussion

CRaft-ID Screening Platform for CRISPR-infected Cells

The microraft arrays used in this work consist of 40,000 separable magnetic polystyrene 100 μm square tiles ("Rafts") as cell growth surfaces, embedded in a polydimethylsiloxane (PDMS) microwell array substrate. The PDMS barriers between microrafts physically separate cell colonies, which grow adhered to microrafts within a shared media reservoir. This design allows for standard cell culture and subsequent isolation of individual colonies expanded from single cells. EIF2AK1 is a kinase that initiates stress granule formation by phosphorylating EIF2α in response to oxidative stress induced by sodium arsenite (McEwen et al., 2005). As a proof-of-concept that variability in stress granule abundance can be reliably measured on these arrays, we pooled cells stably expressing a sgRNA targeting EIF2AK1 with cells that contain a non-targeting sgRNA (sg-NTC) on the same array, each expressing a distinguishing nuclear fluorescent marker (Figure 2-2A-C). Cells were plated at clonal density (0.3 cells per raft). We observed a doublet rate of 9% by counting the number of wells that contain signal from both fluorescent markers (Figure 2-2D). Cells were then treated with 500 μM sodium arsenite for an hour to robustly induce stress granules that stained positive for the canonical stress granule marker, G3BP1. As expected, mCherry-positive (sg-NTC) colonies contained stress granules, and mCitrine-positive (sg-EIF2AK1) colonies did not form stress granules (Figure 2-2E-F). We then isolated 56 colonies of each genotype from the array using the respective fluorescent marker to distinguish genotypes. DNA was extracted and prepared into libraries for sequencing using a targeted-PCR approach. Of the 112 colonies sequenced, we identified 2 colonies with both guides, and 110 were properly assigned to the predicted genotype. These results support

the utility of the microraft array platform to quantify variability in stress granule formation and accurately assign the proper sgRNA after cell retrieval.

To screen for RBPs that modulate stress granule abundance, we developed a CRISPR/Cas9 sgRNA library targeting 1,078 RBPs (Gerstberger et al., 2014) with 10 sgRNAs targeting each gene (ADDGENE 141438). We included 628 control sgRNAs targeting essential genes and 1,070 non-targeting control sgRNAs (including 12 sgRNAs targeting fluorescent proteins). Of the sgRNAs targeting RBPs, 2,210 are annotated as essential to the survival of HEK293T cells (Hart et al., 2017). HEK293T cells were transduced in bulk with the RBP library at low MOI (0.15 viral particles per cell) and cultured for 7 days to allow sgRNAs targeting lethal genes to deplete from the pool (Figure 2-3A). On day 7 post-selection, cells were plated at clonal density on the microraft arrays and cultured for 3 days to form small colonies (5-20 cells per colony) (Figure 2-3B). As stress granule quantification required high resolution, automated confocal imaging previously infeasible on microrafts, we fixed a 1mm thick glass slide to the bottom of microraft arrays using a water-soluble adhesive that is removed prior to isolation of target-colonies. This eliminated depth-of-focus variability caused by the soft PDMS material (Figure 2-4A) and allowed for automated imaging across the entire array. After quantifying cell abundance with nuclei staining, we determined that 6,000 colonies formed on each array (15% of wells on the array). To achieve 10X representation of each sgRNA in our library (Figure 2-4B), we plated a total of 20 arrays, one of which was used as a negative (no stress) control totaling 124,312 plated colonies. As each RBP has 10 unique sgRNAs in the library, this results in 100X coverage per-RBP in this experiment. With commercially available microraft arrays, users can scale their screening needs at 6,000 cell increments with additional arrays (currently $200 per array). Additional cost associated with PCR-based sequencing is negligible, as sequencing

depth required is low ($<$100,000 reads per microraft) and standard confocal imaging equipment is readily available.

Automating confocal microscopy for improved resolution and facilitation of high-content screening

To image the microRafts at high-resolution, we used automated confocal microscopy to acquire 324 tile scan images in 5 z-planes and 3 color channels across each array (Figure 2-5). We developed computational tools to process these images to isolate individual rafts for analysis (https://github.com/YeoLab/CRaftID). To extract stress granules and nuclei features, we collapsed z-stacks using the maximum intensity projection of the red (G3BP1) and blue (nuclei) channels. Brightfield images were used to segment each field of view for individual rafts using the defined grid-pattern. Individual raft images were then filtered for those containing signal in both the nuclei and G3BP1 channels, and raft images with no cells were discarded. In total, we identified 119,050 colonies from CRISPR-infected, stressed cells along with 5,262 unstressed control colonies (Figure 2-5).

To remove common aberrations that compromise image quality in high-content screening, we developed a machine learning tool automating the removal of artifacts prior to stress granule quantification. A consortium of binary classifiers using convolutional neural networks (LeCun et al., 2015) was trained on images of cells stained for nuclei, making this quality control step a generalizable approach to other phenotypes screened on this platform. We manually curated a total of 1,477 images of nuclei-stained cells (70% for training, 10% for training validation, and 20% for a test set) into five different phenotypic categories: overgrown colonies, debris, lysed cells, autofluorescence, and out-of-focus images (Figure 2-6A). Images classified

with at least 99% confidence to any category using a one-vs-all classification approach were eliminated from the dataset, while remaining images were categorized as healthy and retained in the dataset. Overall multi-class precision was 98% and multi-class recall rate was 0.98% when testing this approach on an independent test set of 365 nuclei-stained images. This filtering approach removed 9% (471 of 5,262) of unstressed colony images and 15% (17,462 of 119,050) of stressed colony images (Figure 2-6B).

The remaining 101,588 colony images were quantified for nuclei and stress granule area using a custom pipeline developed in CellProfiler (https://github.com/YeoLab/CRaftID). Images were ranked by the lowest ratio of stress granule to nuclei area and further manually inspected to identify high-confidence colonies containing RBP knockouts that reduce stress granule formation. In total, 101 colonies were isolated from the microRaft arrays for targeted-sequencing preparation to identify the infected sgRNA in each colony (Figure 2-6C-D).

Identification of sgRNA in selected colonies

To sequence the sgRNA associated with reduced stress granule abundances, we isolated target colonies adhered to microRafts from the array. A motorized microneedle, fitted over the microscope objective, was actuated to pierce the PDMS microarray substrate and dislodge individual magnetic microRafts from the array. Released microRafts and their cargo were collected with a magnetic wand into a strip-tube containing lysis buffer for a targeted 2-step PCR with in-line barcodes followed by high throughput sequencing (Figure 2-7A). We identified the minimum number of PCR cycles required to amplify enough material for sequencing (22 for PCR1, 10 for PCR2) and size-selected the final library from a 3% agarose gel (Figure 2-7B). Despite isolating gDNA from small colonies fixed on the rafts, this method generated a sequenc-

ing library for 60% (61) of the picked colonies. The sequencing rate here is lower than what was achieved in the proof-of-concept experiment, likely due to the low MOI required for pooled screening. Notably, the population of knock-out cells generated for the proof-of-concept were generated with a high MOI and clonally selected to ensure protein depletion and therefore resulted in a higher sequencing success rate than was observed in the screen. Failure to generate a library could also arise due to incomplete DNA extraction, or cells peeling off the raft during microRaft isolation from the array. Of the successfully sequenced colonies, 49% contained a single gRNA, and 51% contained 2 or more gRNAs. Importantly, as our sequencing method is able to deconvolute multi-sgRNA colonies, there is no information loss due to doublets, and candidates that occur in multiple rafts are prioritized for independent validation.

Independent Validation of Candidate Stress Granule Regulators

Reassuringly, the most frequently detected RBP candidate is our positive control EIF2AK1, for which we retrieved 5 different sgRNAs from a total of 12 rafts (Figure 2-8A). Images collected from EIF2AK1-identified rafts contain the lowest stress granule area among all sodium arsenite treated cells in our screen. In addition to EIF2AK1, we detected sgRNAs targeting five other known modulators of stress granule assembly: EEF2 (Schneider-Poetsch et al., 2010), G3BP1 (Tourriere et al., 2003), UBAP2L (Markmiller et al., 2018, Youn et al., 2018), EIF5 (Ohn et al., 2008), and ADARB2 (Ohn et al., 2008), along with 65 additional candidates. To independently confirm that candidate RBPs play a role in stress granule abundance, we used a pool of four siRNAs per target to deplete each RBP in HEK293T cells expressing GFP-tagged G3BP1. We identified a total of 17 new RBPs that reduce stress granule abundance by 10% to 50% when depleted with siRNAs in at least two of three independent replicate experiments

relative to a non-targeting control (Figure 2-8B). When possible, we confirmed protein loss by western blot analysis (Figure 2-8B-C); however, due to limited antibody availability we could not confirm knock-down for all targets and therefore cannot rule out the possibility that some hits were not validated due to insufficient protein knock-down. In total, 62% (38) of sequenced rafts contained an RBP target that met our validation criteria in siRNA experiments. The top three hits not previously known to regulate stress granules were SNRNP200, ZCCHC14, and PUF60, each of which reduced stress granule abundance by 50% when depleted with siRNAs (Figure 2-8C). Interestingly, loss-of-function mutations in each of these proteins have been implicated in neurological diseases. PUF60 loss-of-function is associated with microcephaly and intellectual disability (Low et al., 2017), SNRNP200 is lost in patients with retinitis pigmentosa (Zhang et al., 2013), and ZCCHC14 mutations are found in patients with cerebral small vessel disease and autism spectrum disorder (Handrigan et al., 2013, Chung et al., 2019). While no direct link has been made between these diseases and stress granule biology, our results raise the possibility that cytoplasmic granules may play a role given their strong association with neurodegenerative disease. To determine if the identified factors influence recruitment of other known stress granule components, we co-stained cells from siRNA knockdown experiments for UBAP2L and quantified the abundance of UBAP2L-positive granules. We found that 12 of the 22 candidates tested contained a significant reduction in UBAP2L positive granules in at least three of four replicate experiments (Figure 2-9A-B). Interestingly, we observed an abundance of UBAP2L positive granules in cells depleted for G3BP1. This supports a recently published model of distinct UBAP2L-positive stress granule cores that act independently of G3BP1 and can nucleate G3BP1-positive stress granule assemblies (Cirillo et al. 2020). To reveal the direct, or indirect role these candidate RBPs have in regulating stress granule abundance, we

generated a protein-protein interaction (PPI) network by curating the interactomes of UBAP2L and G3BP1 (Youn et al., 2018), as well as the BioPlex project of experimentally validated and predicted PPIs (Huttlin et al., 2015). Many of the modulators identified in this work localize to stress granules (Figure 2-10, solid lines) and interact with G3BP1. Importantly, some RBPs do not colocalize in stress granules or interact with other stress granule modulators identified in this screen (Figure 2-10, dashed outlines). These RBPs represent regulatory nodes of stress granule assembly that exist independently from direct granule formation.

Discussion

The CRaft-ID screening platform presented here represents an advance and expansion of the application of pooled screening approaches to encompass cellular and subcellular phenotypic readouts. This method is widely accessible as it employs standard cell culture practices, CRISPR libraries, off-the-shelf confocal imaging techniques and PCR-based DNA sequencing. In this work, we screened for genetic modulators of stress granule abundance in human cells by analyzing 120,000 lentiCRISPR infected cell colonies. This resulted in the identification of both known and unexpected RBPs that regulate stress granule assembly. MicroRaft arrays behave like standard cell culture dishes, and therefore can accommodate a variety of cell-types, including stem cells (DiSalvo et al., 2019) and cells derived from differentiation protocols. However, this method may not be easily amenable to nondividing cells plated directly on-array. CRaft-ID is best suited to screen for regulators of rare phenotypes among a population of genetic knockouts as individual microRafts must be isolated and PCR amplified to determine the infected sgRNA. In instances where a phenotype is affected by a large fraction of sgRNAs, an alternative strategy is to isolate many microRafts into a single tube and prepare

one library to sequence all sgRNAs in the combined pool. As manual isolation of microRafts takes 2 minutes per raft, it is feasible to pick hundreds of rafts. Two alternative approaches have been used to perform imaging-based pooled genetic screening with varying degrees of throughput and resolution. The first was optimized for detection of high-resolution phenotypes (RNA localization) albeit low-throughput (54 genes targeted, 30,000 cells imaged) (Wang et al., 2019), while the second performed high-throughput measurement (963 genes targeted, 3 million cells imaged) of low-resolution phenotypes (nuclear/cytoplasmic localization) (Feldman et al., 2019). The CRaft-ID platform, by comparison is intermediate with regards to throughput (1,078 genes targeted, 120,000 colonies imaged) and resolution of phenotype (cytoplasmic protein-RNA puncta). In contrast to these methods that use sequencing-by-synthesis of fluorescently labeled nucleotides, CRaft-ID uses traditional PCR-based sequencing to identify the infected sgRNA. While the image-based sequencing approaches have the benefit of identifying the sgRNA barcode present in all screened cells, it requires a customized microscope setup and data analysis engine. Additionally, CRaft-ID is the only platform to our knowledge that is compatible with live-cell recovery upon phenotypic selection. In our work, sodium arsenite treatment is lethal to cells after wash-out and stress granule formation is transient; therefore, colonies were fixed to preserve morphology prior to imaging. However, in experiments where perturbation is nonlethal to cells, imaging can be performed on live colonies to capture dynamic localization patterns of endogenously-tagged proteins and cell morphology followed by live-cell colony isolation from the array for further cell-based studies (Wang et al., 2010). In conclusion, CRaft-ID expands the utility of CRISPR-screening to high-content imaging allowing for the interrogation of genetic modulators of subcellular and cell-morphological phenotypes that have previously been inaccessible with bulk-infection methods. We have shown here that CRaft-ID

can robustly identify both previously described, and newly validated modulators of stress granule abundance, providing novel insight into stress granule biology. This platform is accessible and flexible to countless imaging-based phenotypes, creating a significant advancement in the field of functional genomics screening.

2.3 Concluding remarks

Pooled CRISPR approaches to perform high throughput gene knockout have revolutionized the field of functional genomics. However, by design, they have thus far been limited in the scope of phenotypes available to measure. Our work has expanded the opportunity to evaluate subcellular phenotypes from a heterogeneous group of cells. This method is easy to use and can be widely applied in any laboratory setting. However, it is still limited in throughput as imaging of each microRaft array takes approximately 6 hours. Additionally, a major caviat of this work is that the sgRNA is not assigned for every colony imaged, only the ones isolated from the microRaft. Many technological advances will continue to improve upon this work to address these limiations. Robotic selection and library preparation can assist in increasing throughput, however will then again limit the number of labs who can use this technology. Additionally, in situ sequencing can be used to assign sgRNAs to all cells imaged and as those technologies improved, they can be applied in combination with microRaft arrays for clonal selection of heterogeneous cell populations.

2.4 Materials and Methods

Generating mCitrine and mCherry fluorescently labeled HEK293T cells

PiggyBAC shuttle vectors expressing mCitrine or mCherry were stably integrated via transient transfection. Plasmids were delivered by transfecting 70% confluent HEK293T in a 6 well using 2 μg PiggyBAC shuttle vector, 0.5 μg Super PiggyBAC transposase and 12 μL Lipofectamine 2000. Media was replaced 24 hours after transfection to DMEM supplemented with 10% FBS.

Generating HEK293T cells stably expressing EIF2AK1 or NTC sgRNA

sgRNAs targeting EIF2AK1 and a non-targeting control were cloned into the Lenti-CRISPRv2 backbone. sgRNAs and backbone were digested with BsmBI restriction digest, and ligated with T4 DNA ligase. Complete plasmids were confirmed with sanger sequencing. sgRNA EIF2AK1: TTTAACACCTGGATTTGTGC sgRNA NTC: TCCCAAGGGTTTAAGTCGGG Lentiviral particles were packaged in a 10 cm plate by transfecting HEK293xT at 70% confluency with 10 μg of complete LentiCRISPRv2 plasmid, 5 μg PMD2.G, 7.5 μg psPAX2, 45 μL P3000 reagent and 45 μL Lipofectamine 3000. Media was replaced the next day with 7 mL DMEM supplemented with 10% FBS. 72 hours post transfection, media containing lentivirus was collected, centrifuged at 400 x g for 5 mins, and 2 mL of supernatant was used to transduce a 6 well of fluorescently labeled HEK293T cells. After 48 hours, transduced cells were treated with 2 μg/mL puromycin for 3 days to select for infected cells. To identify cells edited by sgRNA targeting EIF2AK1, mCitrine-positive cells were single cell FACS sorted into a 96 well plate. Cells were grown for 7-9 days before being re-plated onto two replicate 96 well plates. For the

plate used to phenotype cell clones, cells were stressed and stained as described in cell stress treatment and antibody staining method, below. Hit clone from replicate well was expanded for pilot experiment and protein knock-out validated with western blotting.

MicroRaft array microfabrication

MicroRaft array microfabrication followed previously reported techniques. The elastomeric microarray substrates were fabricated via soft-lithographic molding of polydimethylsiloxane (PDMS) microwells from photoresist negative templates. Specifically, a 300 μm PDMS pre-cursor layer was sandwiched between an octyltrichlorosilane-treated template and a 1 mm thick glass slide cleaned with acetone and coated with 30 kDa poly(acrylic-acid). Solid PDMS microwells were formed by curing the sandwich for 40 minutes at 95°C and demolding the template. The glass-laminated PDMS microwell substrate was dipcoated in poly(styrene-co-acrylic-acid) with iron oxide nanoparticles in gamma butyrolactone solvent. As the substrate was retracted from the solution, beads of polymer were formed within each PDMS microwell via discontinuous dewetting of the solution. The polymer microRafts were solidified overnight at 95°C. Injection-molded polystyrene media chambers were cleaned by a 1 hour sonication and overnight incubation in detergent (Alcanox), followed by isopropanol and water rinses. The media chambers were dried and then attached to the microRaft arrays using PDMS that was cured for 3 hours at 70°C. The glass laminate was retained on the microRaft array during confocal microscopy imaging. *NOTE – MicroRaft arrays are commercially available through Cell Microsystems

MicroRaft cell culture

MicroRaft arrays were plasma treated (Harrick Plasma, Ithaca NY) for 5 minutes and sterilized using a 30 minute incubation in 75% ethanol. After three serial 5 minute rinses in 1x PBS to remove traces of ethanol, the microRaft arrays were coated in 0.001% w/v PDL for 1 hour at 37°C. The PDL coated array was washed twice with 1x PBS and stored in cell growth media until cell culture. For the pilot experiment, mCitrine-positive (with sgRNA targeting EIF2AK1) and mCherry-positive (with non-targeting sgRNA) were dissociated and passed through a 40 μm mesh filter to remove clumps. Cells were then combined at a 1:1 ratio and a total of 1.2x104 cells were plated on the array in 1 ml DMEM supplemented with 10% FBS and 1% PenStrep (Gibco 15140122). For the screen, dissociated CRISPR-infected HEK293T cells were filtered through a 40 μm mesh filter to remove clumps and 1.2x104 cells were plated onto each array in 1 mL DMEM with 10% FBS and 1% PenStrep. Arrays were spun at 400 x g for 4 minutes in a swinging-bucket centrifuge to settle cells onto microRaft and stored at 37°C, 5% CO2. After 24 hours, 2 mL media was added to each microarray.

Cell stress treatment and antibody staining

After 72 hours in culture on the microRaft, cells were treated with 500 uM sodium arsenite in DMEM supplemented with 5% FBS for 1 hour at 37°C. To fix the cells, 12% paraformaldehyde (PFA) was added to a final concentration of 4% PFA and incubated for 15 minutes at room temperature. Three washes with Wash Buffer (0.01% Triton-X in 1x PBS) were performed to remove the PFA. Blocking and cell permeabilization were performed with 1 hour incubation in 0.1% Triton-X (Sigma-Aldrich X100) and 5% goat serum (Sigma-Aldrich G9023) diluted in 1x PBS. The cells were then washed with Wash Buffer and incubated in the primary antibody

solution overnight at 4°C (Wash Buffer with 5% goat serum, 1:1000 rabbit anti-G3BP1 [MBL RN048PW, RRID:AB 10794608]). Samples were washed three times in Wash buffer prior to being incubated in secondary antibody solution (Wash Buffer with 5% goat serum, 1:1000 Alexa Fluor 555 goat anti-rabbit [Invitrogen A21429, RRID:AB 141761] or 1:1000 Alexa Fluor 633 goat anti-rabbit [Invitrogen A21070, RRID:AB 2535731]) for 1 hour at room temperature. Cells were then washed three times with Wash Buffer and incubated with Hoechst 33342 (Thermo Scientific 62249) diluted to 1 μg/mL in Wash buffer for 30 minutes. MicroRafts were stored in 1x PBS containing 1% PenStrep for imaging and colony isolation.

Confocal imaging of microRaft array

MicroRaft array plated with fluorescently labeled sg-EIF2AK1 and sg-NTC CRISPR-infected 293T was imaged using a Crest X-Light V2 LFOV spinning disk confocal with a Lumencor Celesta laser engine and mounted on a Nikon Eclipse Ti2 microscope. Images were acquired using a Hamamatsu ORCA Fusion sCMOS camera. The system was operated with NIS Elements High Content (HC) software (Nikon). A 10x (0.45 NA) objective was used to collect a 42 micron thick stack with 7 z-slices for each of the 324 views acquired to capture the entire raft. The 405 nm, 520 nm, 546 nm, and 638 nm lines of the Celesta were used to capture DAPI, mCitrine, and mCherry, Alexa647-Cy5 respectively. A Semrock penta primary dichroic and 450-40, 535-30, 595-31, and 694-72 bandpass emission filters were used to separate excitation and emission light.

To image fixed RBP CRISPR-infected cells on microRafts for the screen, laser scanning confocal microscopy was performed using an Olympus FluoView 3000 microscope operated with Olympus software. An Olympus 10x objective (Olympus UPLSAPO 10) with 0.4 NA, 3.1

mm w.d. was used with 1x digital zoom and 0.62 micron per pixel digital resolution. Images were acquired in a 35 micron focal range, 5 optical slices with 7 microns between slices, through colonies adhered to the microRaft surface. An 18 x 18 paneling of FOVs was used to scan a 1 square inch microRaft array with 14 percent, or 180 micron, overlap between adjacent images. Excitation and emission of Hoechst 33342 and Alexa Fluor 555 were performed simultaneously with combined 405 and 561 nm lasers and a multispectral detector set for wavelengths of 430-470 nm and 615-715 nm, respectively. A confocal aperture of 118 μm was used for all imaging. Sample tilt correction was utilized.

Image processing

A total of 324 images containing 3 channels and 5 slices per channel were collected per microRaft array. The 5 slices were merged with the maximum intensity projection function in FIJI and saved as a single image per channel. Bright-field images were used to segment each image on individual microRafts for further analysis using custom Python scripts (https://github.com/YeoLab/CRaftID). Individual rafts were filtered based on Hoechst 33342 and Alexa Fluor 555 signal to select rafts that contained cells and remove empty wells.

CRISPR plasmid library preparation

A comprehensive list of sgRNA sequences projected to efficiently direct Cas9 cleavage at their target sites was generated using the sequence model using CIRSPR-FOCUS and ordered as a pool of equal molar oligos. The lentiCRISPR RBP plasmid library was cloned using previously reported methods (Shalem et al., 2014). Briefly, the lentiCRISPR v2 backbone50 was digested using BsmBI restriction sites and sgRNA oligonucleotide inserts were PCR amplified

and Gibson-assembled using 36 parallel electroporations to maintain a 300X library complexity. Transformations were spread on fourteen 24.5 x 24.5 carbenicillin selection agar plates. Colonies were grown for 16-18 hours at 32°C. The next day, colonies were scraped off the plates and the cell pellet was maxiprepped (0.9 g cells/column). Plasmid library was stored at -20°C.

CRISPR library virus preparation

HEK293xT cells were seeded on twelve 15 cm plates cells seeded at 40% confluency the day before transfection. One hour prior to transfection the media was removed and replaced with 8mL of pre-warmed OptiMEM. Transfections were performed using 62.5 μL Lipofectamine 2000, 125 μL Plus reagent, 12.5 μg lentiCRISPR plasmid library, 6.25 μg of pMD.2g, and 9.375 μg psPAX2. Media was changed 6 hours after transfection to DMEM supplemented with 10% FBS. After 48 hours, the supernatant was filtered through a 0.45 μm low protein binding membrane. The virus was then ultracentrifuged at 24,000 rpm for 2 hours at 4°C and resuspended overnight at 4°C in PBS. Virus aliquots were stored at -80°C.

Multiplicity of Infection

The volume of virus to achieve a multiplicity of infection (MOI) of 0.15 was determined by titrating virus in each well of a 6-well plate (tested volumes ranged from 0.5μL to 6 μL per well). 7x10^5 cells per well of a 6-well plate were transduced in medium supplemented with 8 μg/mL polybrene for 24 hours. Media (without polybrene) was then replaced and half the cells were split into replicate wells, one of which was treated with 2 μg/mL puromycin. Cells were counted after 3-4 days and MOI was determined by the volume of virus that allows 15% of the cells to

survive.

Viral infection of 293T cells with RBP CRISPR/Cas9 Library

4.2x10^6 HEK293T cells were seeded per 10 cm plate on 4 plates. 18 μL of lentivirus was added per plate the next day in medium supplemented with 8 μg/mL polybrene. Lentivirus was removed 24 hours later, and transduced cells were treated with 2 μg/mL puromycin for 3 days. Plates were then combined and seeded onto microrafts.

Classification of healthy colony images

To infer the classification of an image as a healthy colony or as one of several unwanted classes of colony images, it was fed through a series of binary classifiers with convolutional neural networks (CNN). The models were trained on a manually curated set of 1,477 DAPI images using the Keras-Tensorflow framework running on P6000 GPU instances on Paperspace. During inference, if any classifier in the model gave positive classification to an image above a 99% confidence threshold, the image was classified as such. If none of the classifiers positively called the image, the image was retained for quantification of stress granule abundance.

Stress granule image segmentation and quantification

Nuclei and stress granule images were segmented and quantified using a custom pipeline developed in CellProfiler (v3.0.0, available at https://github.com/YeoLab/CRaftID). Nuclei were identified in the DAPI channel using an object diameter threshold of 9-80 pixel units. To eliminate autofluorescent signal from artifacts detected outside of cell boundaries, a cell body mask was generated by overlaying the granule channel (G3BP1 or UBAP2L) and propagating a

trace from nuclei to the edge of the cytoplasmic protein fluorescent signal. Punctate structures in cell bodies were image processed to enhance speckles with a maximum feature size of 6 pixel units. Stress granules were identified from the enhanced image using an object diameter of 2-10 pixel units. Total stress granule and nuclei area were measured from each image.

MicroRaft Cell Isolation

MicroRafts containing hit cells were isolated from the microRaft array using a motorized microneedle device and previously reported methods (Gach et al., 2011) . Prior to isolation, the glass laminate was removed from microRaft array using an overnight incubation in water to dissolve the PAA adhesive. The position of target rafts was calculated in micron distance relative to the upper-left corner of the array using image processing scripts available on GitHub (https://github.com/YeoLab/CRaftID). The motorized microneedle device and microRaft array were placed on an Olympus IX81 microscope. A custom software graphical user interface was used to automatically position and actuate the microneedle to puncture the elastomeric PDMS array substrate, thus dislodging the target microRaft and its cellular cargo. This method can be adopted to any motorized stage using the coordinates calculated for each target microRaft and manual calibration of the microneedle device (5 min setup per picking session). A hand-manipulated magnetic wand was utilized to transfer the floating microRaft into 6 μl of QuickExtract (Lucigen QE09050) buffer in a strip tube. Time required to isolate and collect individual microRafts is approximately 2 minutes. *NOTE – microneedle and motorized device are commercially available through Cell Microsystems

Guide Identification of target wells

Isolated rafts were stored at -20°C in QuickExtract buffer until library preparation. Samples were thawed and DNA was isolated following manufacturer's protocol: 15 s vortex, 65°C for 6 minutes, 15 s vortex, 98°C for 2 minutes. All PCR reactions were carried out with Q5 High-Fidelity DNA polymerase (NEB M0492L). A first round containing 22 PCR cycles was performed using indexed PCR primers targeting the common regions flanking the CRISPR guides that contain a handle for sequencing barcodes to bind in a subsequent reaction. 4 individual reactions with unique index sequences from the first PCR were combined and purified with a Qiagen PCR purification kit to generate the template for a second round of PCR with 10 cycles. A second round of PCR was performed using the purified template and Illumina sequencing primers to generate a sequencing library. Gel extraction was used to specifically isolate the desired product for sequencing (260 bp). Libraries were sequenced on Illumina HiSeq4000 SE75. For more details, see Supplementary Experimental Protocol.

(NNNNNN is reserved for a unique, 6bp index sequence).

PCR1 Fwd: CCTACACGACGCTCTTCCGATCTNNNNNNTTCTTGTGGAAAGGACGAAACACC PCR1 Rev: GTTCAGACGTGTGCTCTTCCGATCTCCACTTTTTCAAGTTGATAACGGACTAGCC

PCR2 Fwd: AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCT PCR2 Rev: CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

Bulk CRISPR gRNA library preparation

DNA Preparation

DNA libraries were prepared using a targeted-enrichment approach. gDNA was extracted from pellets of 4 million cells using DNeasy Blood and Tissue kit (Qiagen 69504) eluted in 130 μL, with typical yields of 150 ng/μL. gDNA samples were sonicated to 1000 bp by Bioruptor. Average fragment size was determined with genomic DNA ScreenTapes on the Agilent Tapestation (Agilent 5067-5365).

Probe Generation

To selectively enrich sgRNA-containing regions in the genomic DNA, we generated two antisense probes by PCR amplification of a 500 nt constant region flanking the sgRNA sequence. Corresponding 592nt and 574nt biotinylated RNA probes were generated using HiScribe T7 High Yield RNA Synthesis Kit (NEB E2040S) with bio-CTP (Thermo 19519016), and bio-UTP (Sigma/Roche 11388908910) nucleotides. PCR Primer set 1: Fwd: GGGATATTCACCATTATCGTTTCAGACC Rev: GGATTCTAATACGACTCACTATAGGGTGTTTCGTCCTTTCCACAAGA PCR Primer set 2: Fwd: GGTGTATCTTCTTCTGGCGGTTC Rev: GGATTCTAATACGACTCACTATAGGGCAAGTTAAAATAAGGCTAGTCCGTTATCA

Probe Capture

1% of 1 M DTT was added to genomic DNA for a final concentration of 10 mM. Concentration of probes was determined for each sample as 10% of the total DNA yield (in micrograms), diluted in water to a final volume of 10 μL. Samples were placed in a pre-heated thermomixer set at 95°C with interval mixing (1200 rpm, 30 second on/1 min off). Immediately after adding the samples, the temperature was changed to 65°C to begin cooling. When cooled to 65°C, 10 μL of probes were added, followed by 73.5 μL of 3X Hybridization buffer (75 mM Tris, 15 mM EDTA, 1.2 M LiCl, 3 M Urea, 0.3% NP-40, 0.3% SDS, 0.3% DOC). Incubation was performed with interval mixing as follows: 65μC 5 min, 64μC 5 min, 63μC 5 min, 62μC 5 min, 61μC 150

min.

Streptavidin Capture

30 μL of streptavidin beads (Invitrogen 11205D) per 1 μg of probes were used for each sample, washed with 500 μL of 1X Hybridization buffer and resuspended in 20 μL of 1X Hybridization buffer. Following probe capture, 14 μL of beads (75%) were added to each sample and incubated for 15 mins at 62°C with interval mixing. Supernatant was removed and transferred back into the tube with the remaining 25% of the beads for a second round of hybridization (62°C with interval mixing for 15 mins). Meanwhile, the collected 75% of the beads (on magnet) were resuspended in 200 μL pre-warmed 1X Hybridization buffer and incubated for 5 mins at 37°C. Supernatant was discarded and tubes were kept on ice. Following the second 15 min incubation, supernatant was discarded from the tube containing the remaining 25% of the beads and beads were resuspended in 200 μL pre-warmed 1X Hybridization buffer for 5 mins at 37°C. Samples were combined by resuspending all beads in 54 μL of LoTE (100 mM NaCl, 0.25% NP-40) + 6 μL of RNase Cocktail (Thermo AM2286) and incubated at 37°C for 10 mins. 6 μL of 1M NaOH was added followed by incubation at 70°C for 10 minutes. Supernatant was transferred to a fresh tube and a second elution was performed by resuspending beads in 30 μL of 100 mM NaOH and incubated at 70°C for 2 mins with shaking. Supernatant was combined with first transfer. 9 μL of 1 M HCl was added to the final sample and DNA cleanup was performed with Zymo DNA concentrator-5 kit (Zymo D4014) following manufacturer's instructions, eluted in 40 μL of pre-warmed water.

PCR amplification

First PCR. 100 μL per sample and split into 2x 50 μL samples in strip tubes: 40 μL DNA, 50 μL 2X Q5 PCR mix (NEB M0492L), 5 μL of each primer at a concentration of 20

μM. PCR program: 98°C 30 sec, 98°C 15 sec, 68°C 1 min, 72°C 1 min, GOTO step2 9 times, 72°C 2 min, HOLD 4°C. Primers: Fwd: CCTACACGACGCTCTTCCGATCTTGTGGAAAGGAC-GAAACACCG Rev: GTTCAGACGTGTGCTCTTCCGATCTCCACTTTTTCAAGTTGATAACG-GACTAGCC Cleanup was performed with 1.8X AmpureXP beads (Beckman Coulter A63881) according to manufacturer's instructions, eluted in 40 μL water for second PCR input. Second PCR. 100 μL per sample was split into 2x 50 μL reactions in strip tubes: 40 μL DNA elution from 1st bead cleanup, 50 μL 2X Q5 mix, 5 μL each of 20 μM Illumina sequencing primers. PCR program: 98°C 30 sec, 98°C 15 sec, 68°C 1 min, 72°C 1 min, GOTO step2 6 times, 72°C 2 min, HOLD 4°C. Performed bead cleanup with 1.4X AmpureXP beads according to manufacturer's instructions. Elution was performed in 20 μL water. Library size (260 bp) and concentration were calculated using D1000 Tapestation (Agilent 5067-5582) and sequenced to 2M reads per library on the Hi-Seq4000 in single-end 75 bp mode.

siRNA transfections

Well plates were coated with 0.001% w/v PDL and incubated overnight at 37°C. Immediately before cell plating, the PDL was aspirated and the wells were washed twice with 1x PBS. For imaging in 384-well plates (PerkinElmer 6057300), 4.5x10^3 HEK293T-G3BP1-GFP31 or HEK293T cells were reverse transfected using 10 nM of siRNA (Dharmacon On-TARGETplus SMARTpool) and Lipofectamine RNAiMax (Invitrogen 13778), according to manufacturer's protocol, per well. Similarly, 2.25x10^5 HEK293T cells were reverse transfected with siRNAs in 12-well plates for protein lysate collection. After 48 hours, stress granules were induced by adding sodium arsenite diluted in DMEM with 10% FBS to a final concentration of 500 μM and incubated for 1 hour at 37°C. Cells intended for imaging were then fixed by adding 24% PFA

to each well to a final concentration of 4% and incubated for 90 minutes at room temperature. Three washes with 1x PBS were performed to remove PFA. HEK293T cells were immunostained for G3BP1 as described above for cell stress treatment and antibody staining protocol. For UBAP2L co-stainings, primary antibodies 1:1000 mouse anti-G3BP1 [Millipore 05-1938, RRID:AB 11214423] with 1:500 rabbit anti-UBAP2L [Bethyl A300-533A, RRID:AB 477953] and secondary antibodies 1:1000 Alexa Fluor 488 goat anti-mouse [Invitrogen A11029, RRID:AB 138404], 1:1000 Alexa Fluor 555 goat anti-rabbit [Invitrogen A21429, RRID: AB 141761] were used. Nuclei were stained with DAPI (1:5000 v/v in PBS) for 30 minutes at room temperature. Cells were washed once with 1x PBS before being preserved in 50% v/v glycerol diluted in 1x PBS.

Imaging of siRNA KD in 384-well plate

Plates were imaged using a Nikon Eclipse Ti2 microscope system operated with NIS Elements High Content (HC) software (Nikon). A 20x (0.75 NA) objective was used to collect an 8 micron focal range with 8 z-slices for each of the four views acquired per well. The lasers used were 395 nm, 470 nm, and 555 nm for DAPI, GFP, and RFP, respectively.

Western Blot Analysis for RBP KD

Cells were collected from 12-well plates and pelleted in ice-cold PBS. Pellets were resuspended in 100 μL of RIPA buffer (Sigma R0278) supplemented with Protease Inhibitor (Sigma 539134). Total protein was quantified with BCA assay (Thermo 23225) and 20 μg was run on 4%-12% BisTris gel (Thermo NP0322). Primary antibodies were diluted at 1:1000 (1:4000 for GAPDH) in 5% milk-TBST and probed overnight at 4 °C. EIF2AK1 (Bethyl A302-685A, RRID:AB

10754970), PUF60 (Bethyl A302-817A, RRID:AB 10631036), ZCCHC14 (Bethyl A303-096A, RRID:AB 10895018), SNRNP200 (Bethyl A303453A, RRID:AB 10949362), GAPDH (Abcam ab8245, RRID:AB 2107448). Secondary antibodies were diluted at 1:4000 in 5% milk TBST and probed for 2 hours at room temperature [Rabbit Secondary (Rockland 18-8816-31, RRID:AB 2610847), Mouse Secondary (Rockland 18-8817-30, RRID:AB 2610849)]. Visualization was performed with ECL and film.

PPI Interaction

The human PPI data was retrieved from BioPlex project (BioPlex2.0), Mentha dataset (version 2018-01-08) and proximity-based proteomic studies of stress granule components. The Local PPI network of the protein of interest were presented as an undirected and unweighted graph with each protein as a node and each interaction as an edge. The RBP annotations were collected from several RBP discovery studies and GO database, where we retrieved the proteins under the GO term of “RNA-binding” (GO:0003723) and its descendent terms from AmiGO 2 database (version released in July 2016).

2.5 Acknowledgements

The authors thank Yeo lab members S. Markmiller for the HEK293T-G3BP1-GFP cell line and F. Tan for the PiggyBAC shuttle vectors. The authors acknowledge Yeo lab members S. Markmiller, M. Perelis, J. Nussbacher, A. Smargon, M. Corley and E. Boyle for critical reading of the manuscript. We thank the members of the Nikon Imaging Center at UC San Diego for help with imaging experiments. E.C.W and A.Q.V were supported by the National Science Foundation Graduate Research Fellowship. E.C.W and N.A. were supported in part by a Ruth L.

Kirschstein Institutional National Research Award from the National Institute for General Medical Sciences, T32 GM008666. J.M.E. is supported by the Ruth L. Kirschstein F31 National Research Service Award (F31 CA217173) and Cancer Systems Biology Training Program (P50 GM085764 and U54 CA209891). M.D. is supported by the Ruth L. Kirschstein F31 National Research Service Award (F31 CA206233). E.L.V. is supported by the National Human Genome Research Institute (K99HG009530). This work is partially supported by NIH grants HG004659 and NS103172 to G.W.Y and NIH grant EY024556 to N.L.A.

Chapter 2, in full, is a reprint of material as it appears in: **Wheeler, EC.***, Vu AQ*, Einstein JM, DiSalvo M, Ahmed N, Van Nostrand EL, Shishkin AA, Jin W, Allbritton NA, Yeo GW. Pooled CRISPR screens with imaging on microraft arrays reveals stress granule regulatory factors. Nat Methods, doi:10.1038/s41592-020-0826-8 (2020). The dissertation author was first author of this paper.

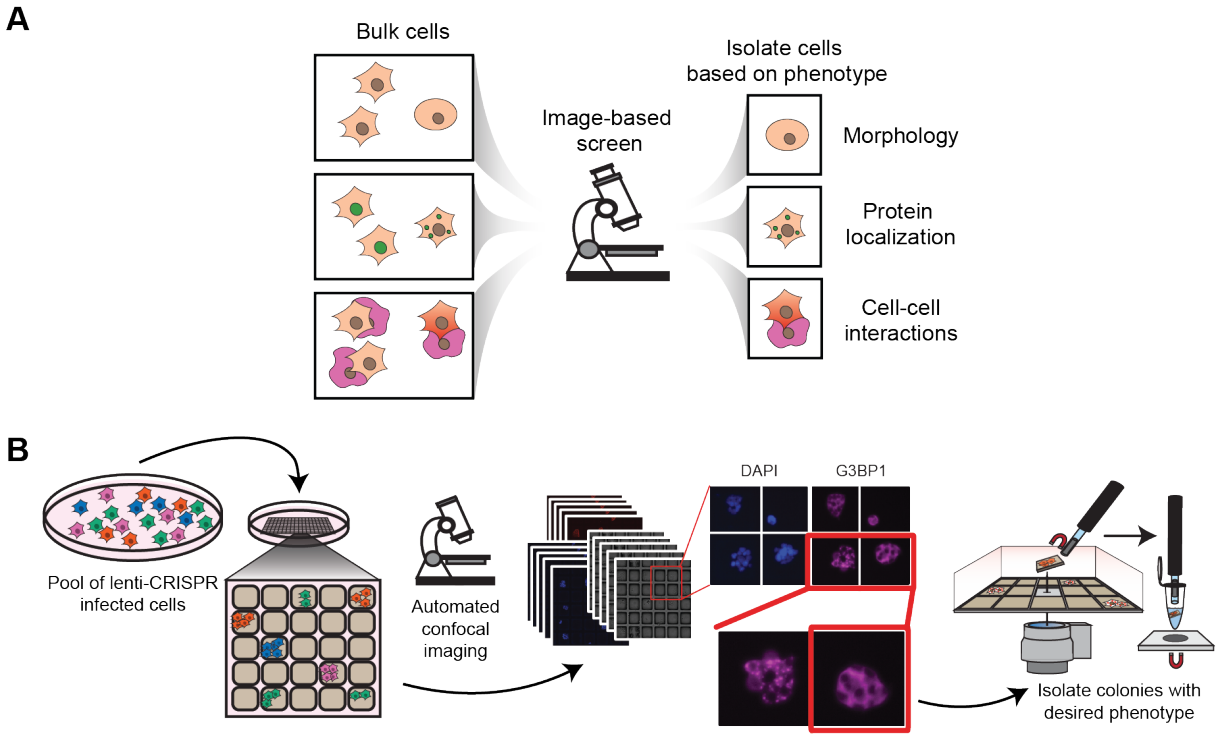


Figure 2.1: Schematic of imaging-based screening platform. (A) Cartoon of imaging-based phenotypes that can be assayed in bulk (left side) with an image-based screening platform to select clones with phenotypes of interest (right side). (B) CRaft-ID workflow. Cells are infected and cultured in bulk with a lentiCRISPR library before plating on microraft arrays. Automated confocal imaging is used to image the entire array and identify colonies with a phenotype of interest. A microneedle fitted over a microscope objective is used to isolate rafts out of the array, and collected with a magnetic wand for PCR.

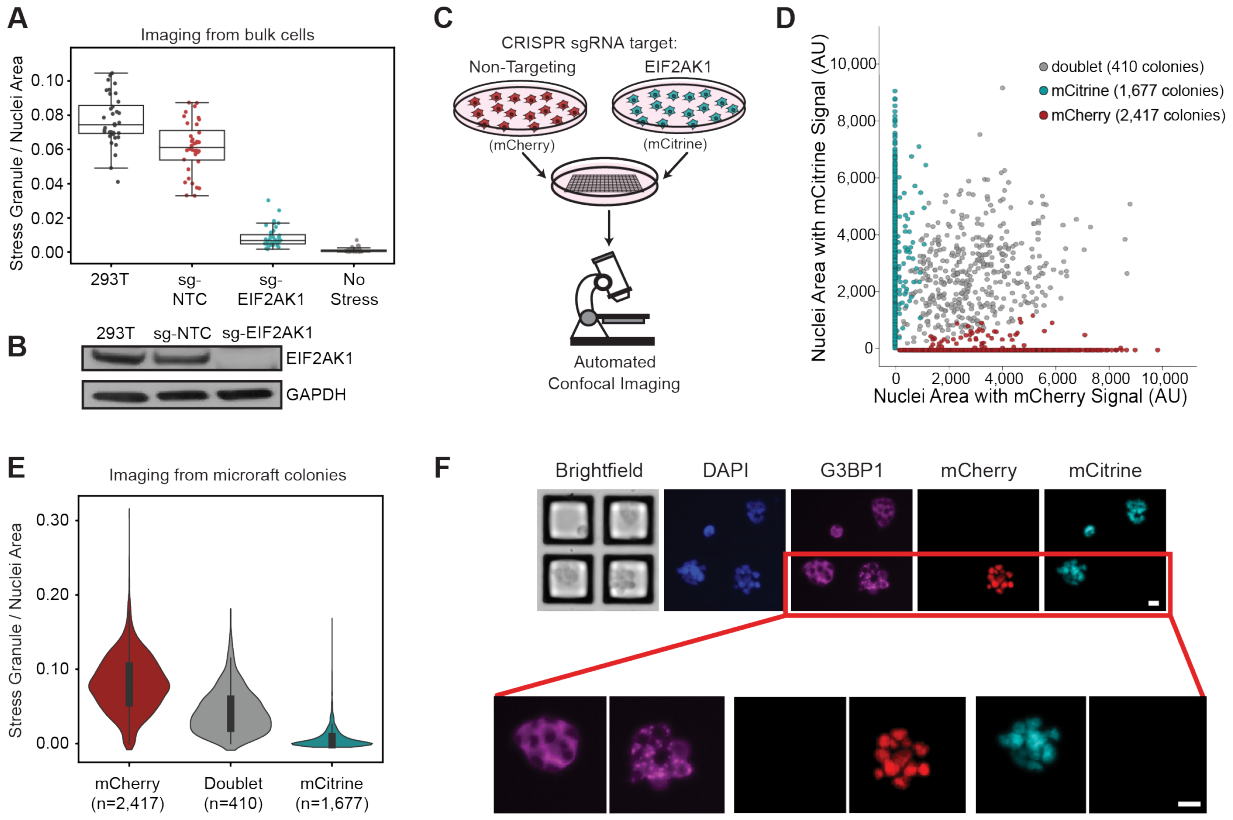


Figure 2.2: Proof-of-concept study of stress granule quantification on microRaft arrays. (A) Quantification of the ratio of stress granule to nuclei area of cell lines used in proof-of-concept study along with an unstressed control. Each data point is one image taken among 6 different wells plated (6 images per well, total 36 images per condition). Overlaid boxplots represent the interquartile range (IQR, 25th, 50th, and 75th percentiles), while whiskers represent 1.5 times the IQR from the 25th (lower) and 75th (upper) percentiles. (B) Western blot of EIF2AK1 and loading control (GAPDH) in sg-EIF2AK1, sg-NTC (nontargeting control), and uninfected (293T) cells. (n = 1). (C) Schematic of pooled experiment used in proof-of-concept study. sg-NTC cells are labeled with mCherry and sg-EIF2AK1 cells are labeled with mCitrine. Cells were pooled and plated on a microRaft array. (D) Scatterplot of mCherry and mCitrine area measured in the nuclei of all colonies detected on a microRaft array. Colonies that contain fluorescent signal from both channels in more than 10% of the total nuclei area are determined as doublets (gray). (E) Quantification of the ratio of stress granule to nuclei area for all colonies, binned by fluorescent signal. Thick line in middle of violin represents the IQR, middle dot shows the median, and thin lines extend 1.5 times the IQR. (F) Representative images showing 4 colonies in the field of view with zoomed images, below. Scale bar = 20 μm.

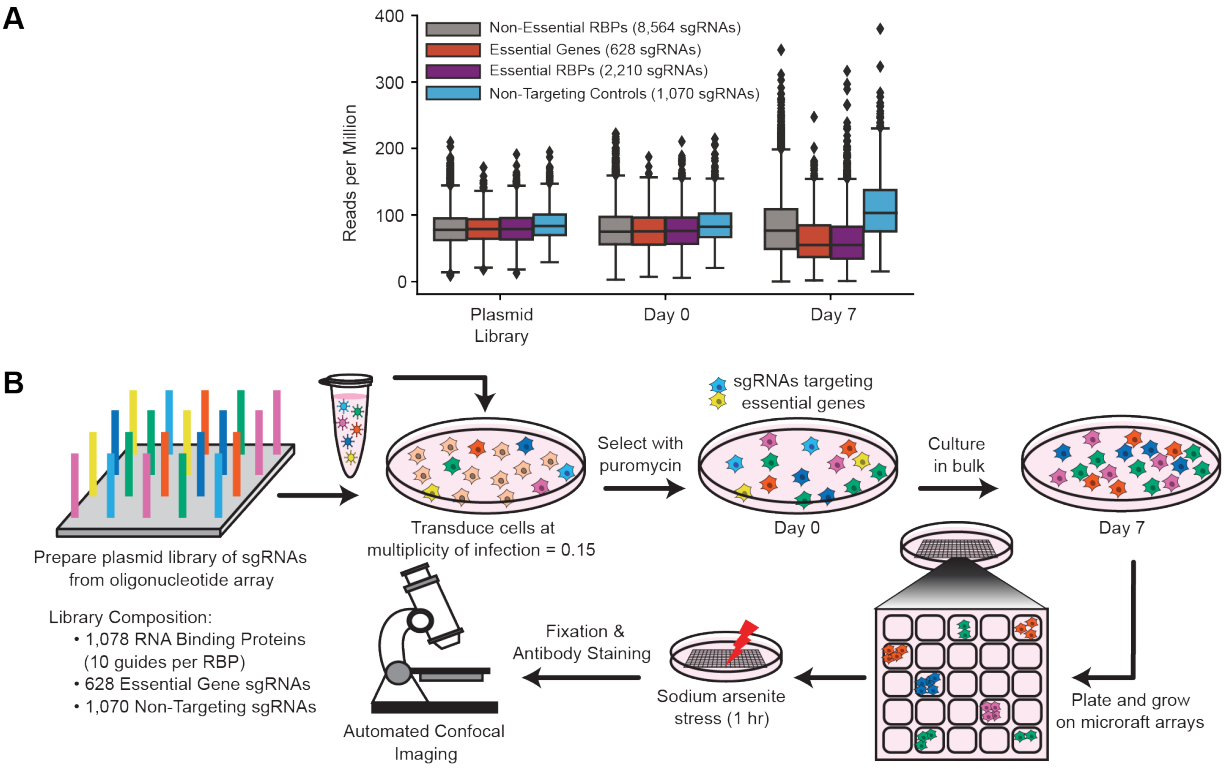


Figure 2.3: CRISPR screening workflow of bulk infection combined with image-based phenotyping. (A) Relative abundances of sgRNAs in the plasmid library, cells at Day 0 (after puromycin selection), and Day 7. Boxes represent the interquartile range (IQR, 25th, 50th, and 75th percentiles), while whiskers represent 1.5 times the IQR from the 25th and 75th percentiles. (B) Cell culture workflow for CRaft-ID. A CRISPR/Cas9-sgRNA library was generated from an array of sgRNA oligos cloned into the lentiCRISPRv2 backbone. HEK293T cells were infected at low multiplicity of infection and cultured in bulk for 7 days after selection allowing lethal guides (light blue, yellow) to drop out of the pool. Cells were plated on microRaft arrays and grown at low density, stressed with sodium arsenite (500 μM), fixed and stained for G3BP1 and nuclei prior to confocal imaging.

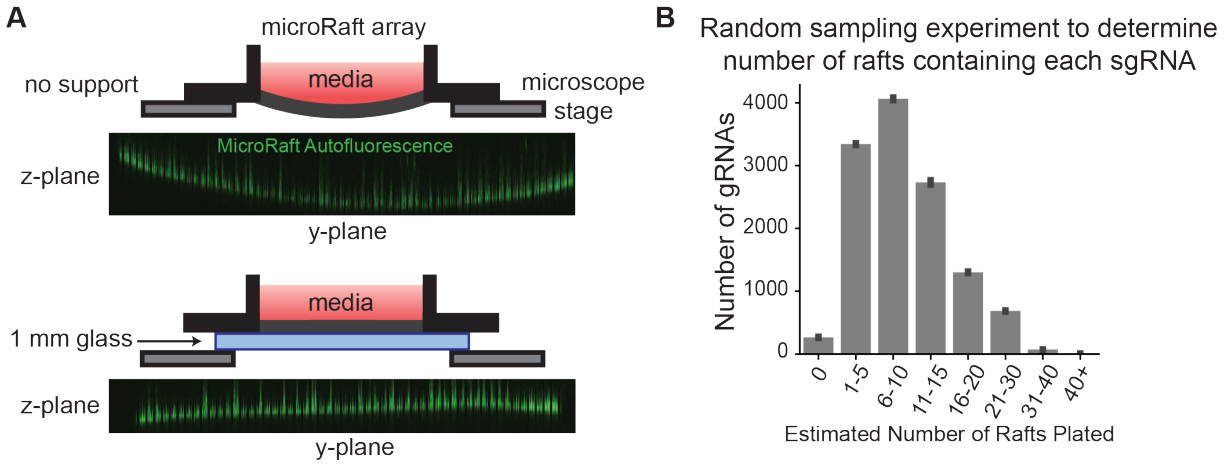


Figure 2.4: MicroRaft array platform for CRISPR screening. (A) Top, schematic of microRaft array without glass-back support. Orthogonal view of autofluorescence (green) in microRafts across PDMS array after imaging with high laser power. Bottom, diagram of microRaft array with 1 mm glass support with orthogonal view of autofluorescent microRafts after imaging with high laser power (green). (B) Random sampling to estimate plating frequency of sgRNAs on rafts in this screen. Given the relative abundances of sgRNAs on day 7 and the total number of colonies plated (120,000), random sampling was used to estimate the number of rafts that contain each sgRNA (x-axis), binned in counts of 5. Bars are the average of n = 10 random samplings with error bars displaying standard deviation.

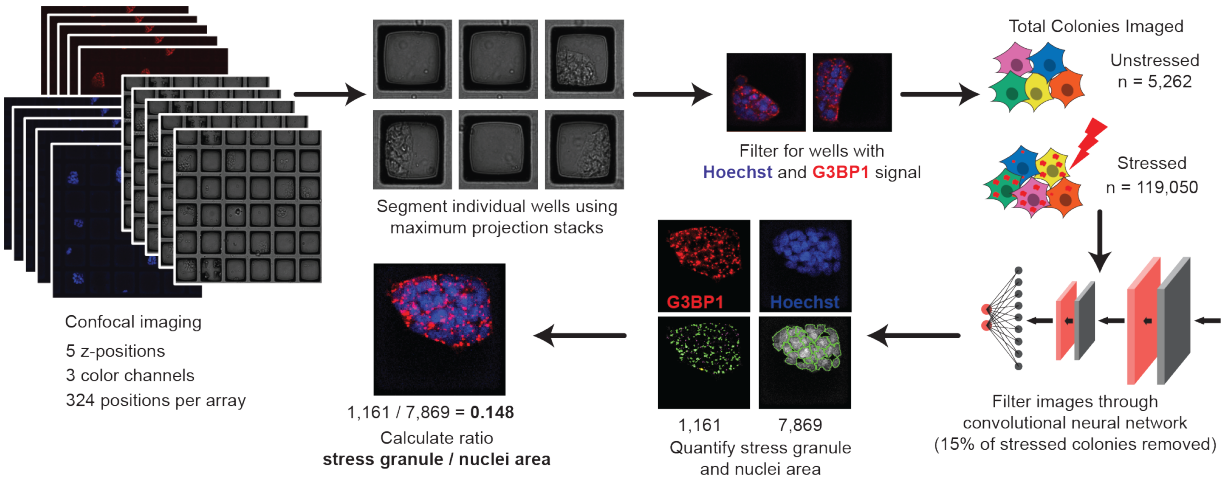


Figure 2.5: Image analysis workflow. Schematic of image processing. Confocal microscopy was performed covering each array in 324 fields of view with 5 z-stacks and 3 channels (Hoechst, G3BP1 staining, and brightfield). Raft coordinates were identified from the brightfield image and filtered for those with signal in Hoechst and G3BP1 channels. Images were processed by a convolutional neural network (CNN) trained on Hoechst staining to remove imaging artifacts and unhealthy colonies. Quantification of stress granule and nuclei area was performed for each colony in CellProfiler.

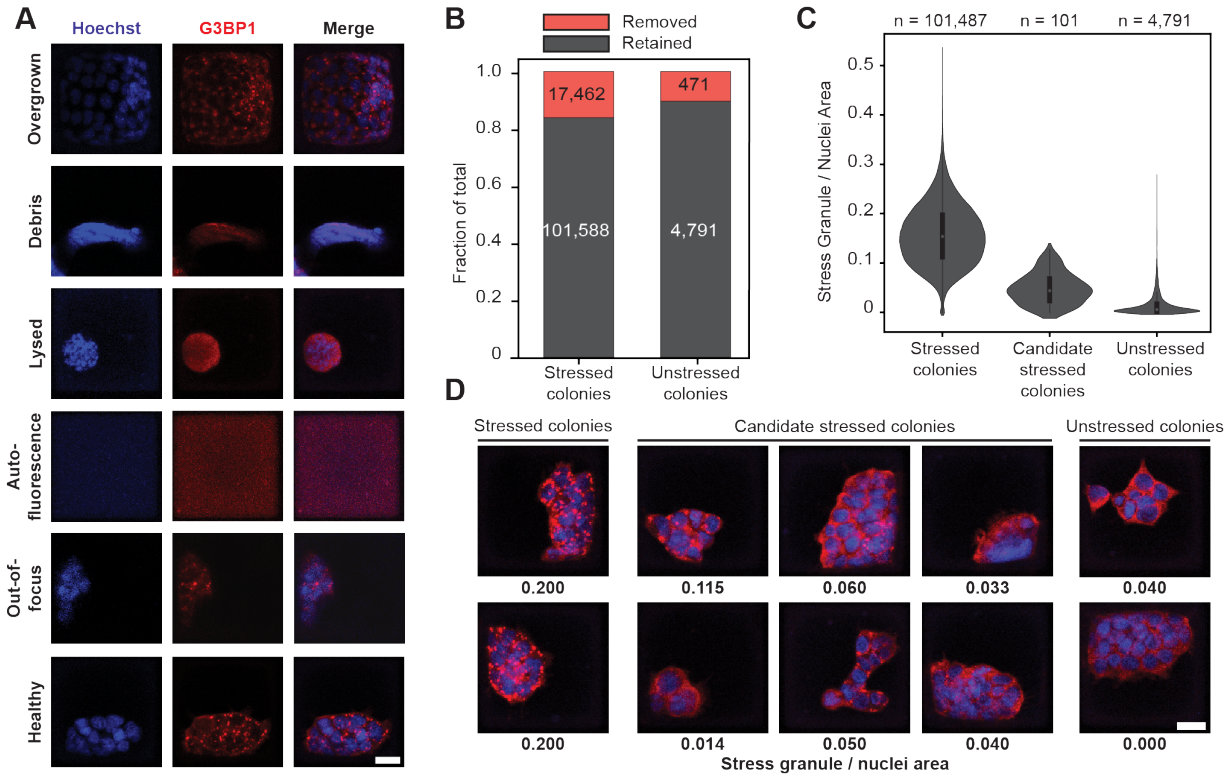


Figure 2.6: Selection of images with low stress granule abundance. (A) Example raft images that were discarded using the CNN classification. Raft with healthy cells and acceptable imaging shown at the bottom of the panel. Scale bar = 20 μm. (B) Fraction of all rafts removed from by CNN. 15% of stressed colonies and 9% of unstressed colonies were removed. (C) Violin plots of stress granule / nuclei area of all healthy raft images. 101 stressed colonies with low stress granule / nuclei area were manually picked as candidates with low stress granule abundance. Thick line in middle of violin represents the IQR, white dot shows the median, and thin lines extend 1.5 times the IQR. (D) Representative images of stressed cells not selected, candidate stressed colonies, and unstressed controls. Numbers below each image represent the quantified stress granule / nuclei area. Scale bar = 20 μm.

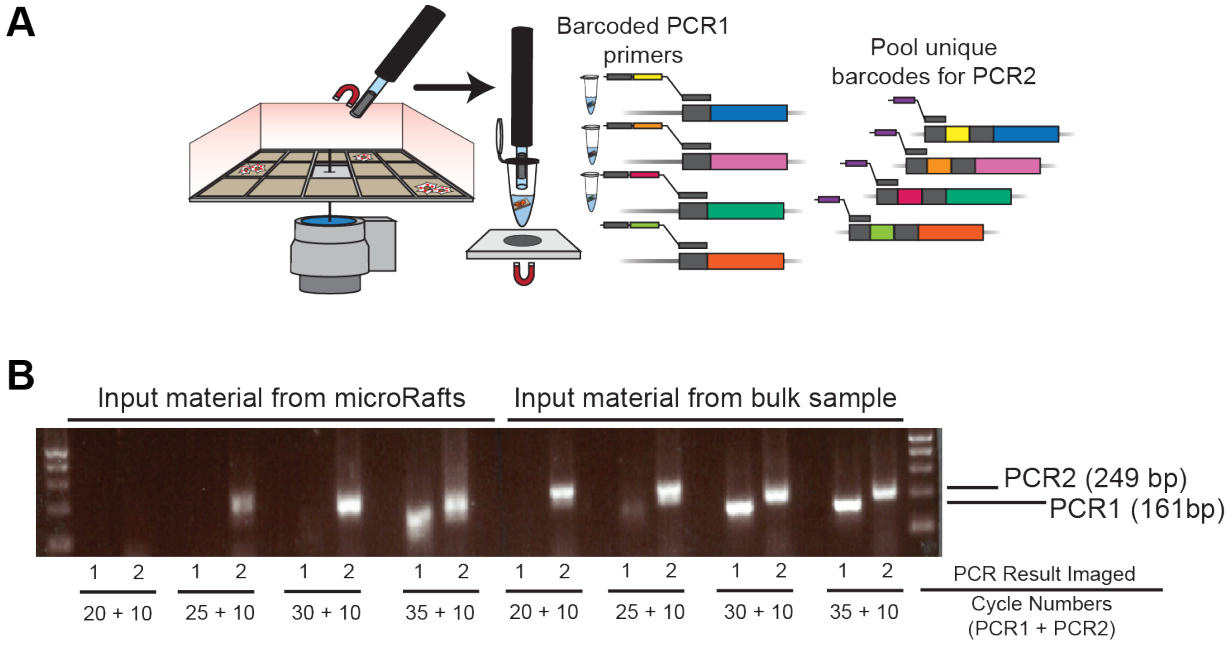


Figure 2.7: Isolation of target microrafts and library prep of gRNAs. (A) Diagram of microRaft isolation and sgRNA sequencing design. MicroRafts are removed from the array using a motorized microneedle fitted over a microscope objective. Dislodged rafts collected with a magnetized wand are placed into individual tubes for DNA extraction and a barcoded targeted-PCR of the sgRNA insert. (B) Agarose gel of PCR products with increasing cycle numbers to determine the minimum number of PCR cycles required to amplify a product for sequencing. Input material from bulk sample is used as a positive control. All rafts sequenced in this study were amplified with 22 cycles for PCR1 and 10 cycles for PCR2 (n=213 total, 173 successful).

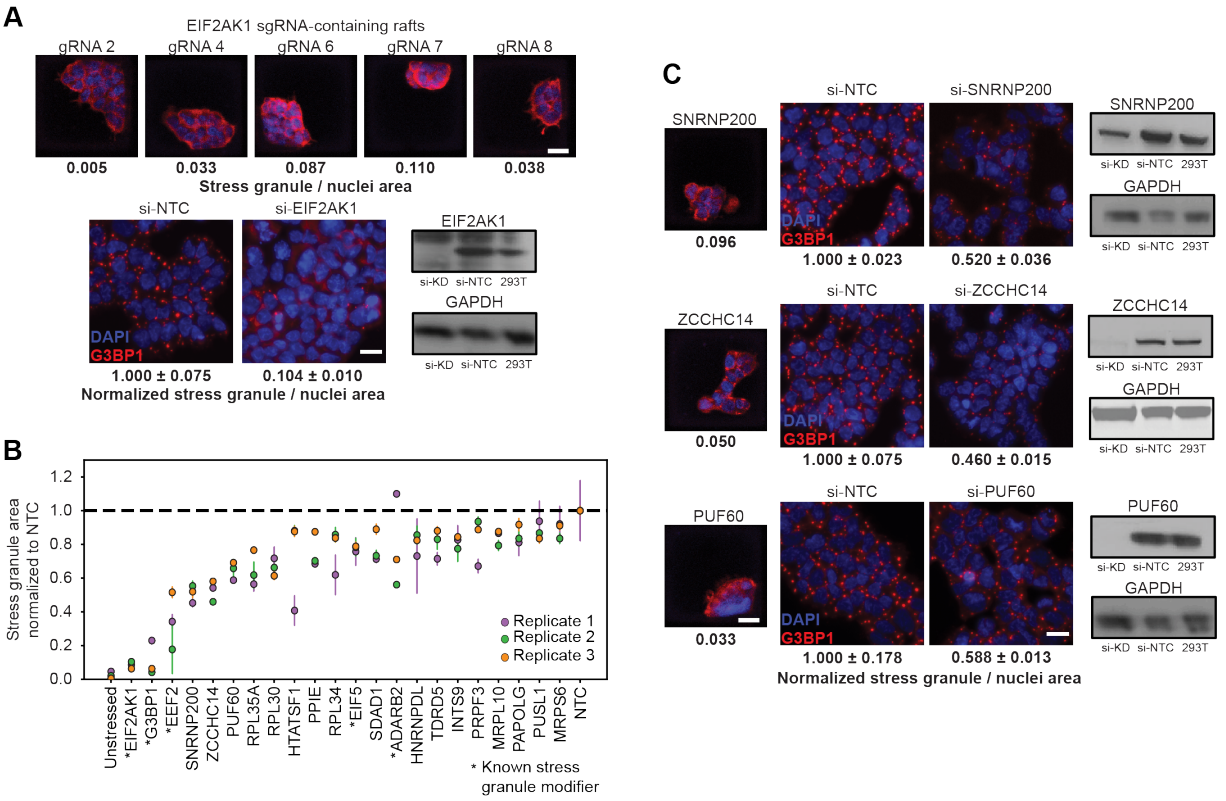


Figure 2.8: Validation of targets with siRNA-mediated depletion. (A) Top, images of individual, stress induced microRafts with 5 different sgRNAs targeting EIF2AK1. Scale bar = 20 μm. Bottom left, representative image of G3BP1 antibody staining in sodium arsenite stressed HEK293T after EIF2AK1 depletion by siRNA or non-targeting siRNA control (NTC). Data are mean $\pm$ s.d. across n = 3 wells / condition (4 images / well). Scale bar = 20 μm. Bottom right, Western blot validation of siRNA-mediated depletion of EIF2AK1 in HEK293T. (B) , siRNA depletion of target RBPs in bulk cells. Stress granule / nuclei area was normalized to the non-targeting control (NTC) for each experiment. RBPs are ordered on the x-axis by the lowest normalized stress granule area. RBPs shown here had significant reduction ($P < 0.05$, unpaired two-tailed t test, d.f. = 4, 95% confidence interval) of stress granule area relative to NTC in at least 2 of the 3 biological replicates. *Genes previously reported to modify stress granule abundance. Data are mean $\pm$ s.d. across n = 3 wells / condition (4 images / well). (C) Validation of SNRNP200, ZCCHC14, and PUF60 (top to bottom). Left, image of identified microRaft colony. Middle, representative image of stress induced G3BP1-positive (red) granule and DAPI (blue) after protein depletion by siRNA in HEK293T or HEK293T-G3BP1-GFP cells (false colored red for consistency). Scale bar = 20 μm. Quantification below image is mean stress granule area $\pm$ s.d. across n = 3 wells / condition (4 images / well) normalized to NTC. Right, Western blot validation of target protein depletion by siRNA.

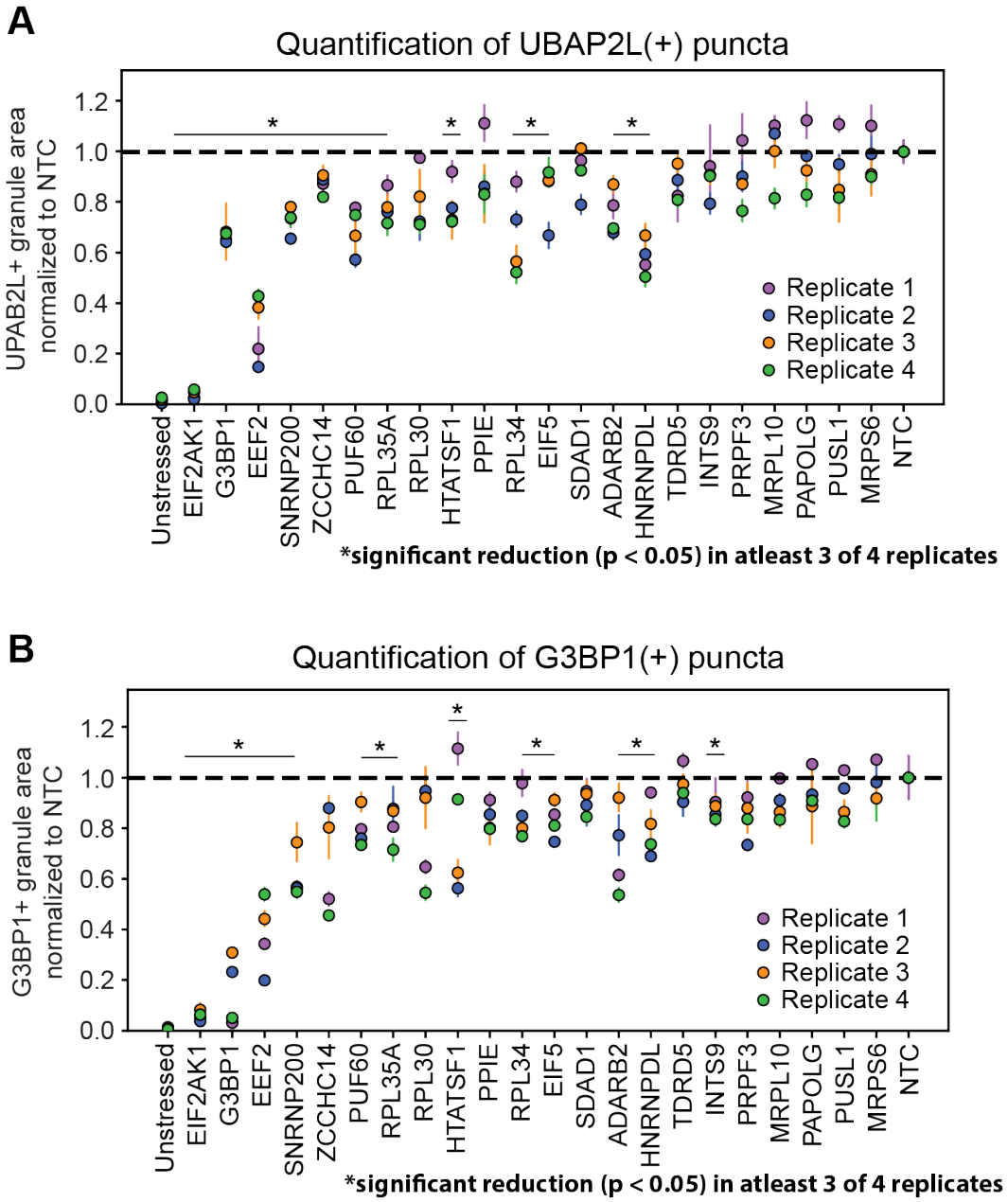


Figure 2.9: Quantification of UBAP2L-positive granules in siRNA depleted cells. (A) siRNA depletion of target RBPs. UBAP2L(+) granule / nuclei area was normalized to the non-targeting control (NTC) for each experiment. RBPs are ordered in order of appearance in Fig. 3c. *RBPs that had significant reduction ($P < 0.05$, unpaired two-tailed t test, d.f. = 4, 95% confidence interval) of UBAP2L (+) granule area relative to NTC in at least 3 of the 4 biological replicates. Data are mean $\pm$ s.d. across n = 3 wells / condition (4 images / well). (B) G3BP1(+) granule / nuclei area from respective wells measured in panel a. Values are normalized to non-targeting control (NTC) for each experiment. *RBPs that had significant reduction ($P < 0.05$, unpaired two-tailed t test, d.f. = 4, 95% confidence interval) of UBAP2L (+) granule area relative to NTC in at least 3 of the 4 biological replicates. Data are mean $\pm$ s.d. across n = 3 wells / condition (4 images / well).

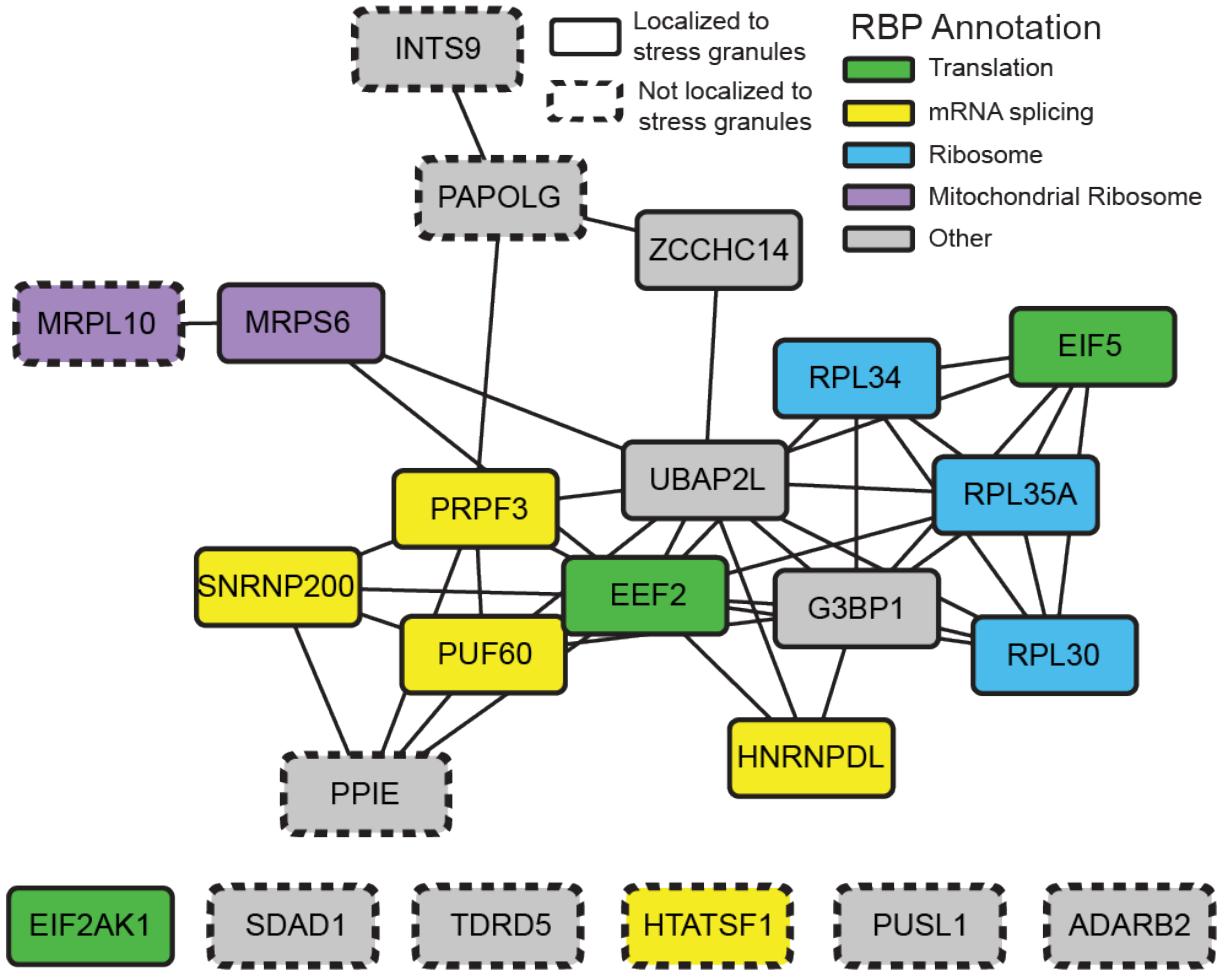


Figure 2.10: Protein-protein interaction network of the 23 RBP targets identified to modulate stress granule abundance. Network visualized in Cytoscape. Protein-protein interactions curated from BioPlex data and recent publications.

Chapter 3

Advances and challenges in the detection and interpretation of transcriptome-wide protein-RNA interactions with CLIP-seq methods.

RNA binding proteins (RBPs) regulate the processing of target RNAs and play key roles in determining cellular behavior. To understand RBP function, robust methods are required to detect their numerous binding sites across the transcriptome. RNA-immunoprecipitation followed by sequencing (RIP-seq) and crosslinking followed by immunoprecipitation and sequencing (CLIP-seq) are two state-of-the-art methods used to identify the RNA targets and specific binding sites of RBPs. Historically, CLIP methods have been confounded with challenges such as the requirement for tens of millions of cells per experiment, low RNA yields resulting in li-

braries that contain a high number of PCR duplicated reads, and technical inconveniences such as radioactive labeling of RNAs. However, recent improvements in the recovery of bound RNAs and the efficiency of converting isolated RNAs into a library for sequencing have enhanced our ability to perform the experiment at scale, from less starting material than has previously been possible, and resulting in high quality datasets for the confident identification of protein binding sites. These, along with additional improvements to protein capture, removal of non-specific signals, and methods to isolate non-canonical RBP targets have revolutionized the study of RNA processing regulation, and reveal a bright future for mapping the human protein-RNA regulatory network.

3.1 Introduction

Our appreciation of the importance of RNA processing in the maintenance of cellular homeostasis has increased significantly in recent years (Tartagila et al., 2016, Kudinov et al., 2017, Castello et al., 2013). RNA binding proteins (RBPs) interact with their target RNAs to affect the creation (Gertsberger et al., 2014), localization (Muller-McNicoll et al., 2013), and function of each RNA molecule in the cell (Hurt et al., 2013). Disruption of these protein-RNA interactions by mutations in RBPs has been implicated in many diseases including neurodegeneration (Nussbacjer et al., 2015) and cancer (Saez et al., 2017). Therefore, identifying the RNA targets of specific RBPs is important for deciphering the molecular mechanisms of RBP-mediated diseases. Here, we discuss current technologies that identify RBP targets and present technical challenges that need to be addressed in the future. In particular, we focus on three major areas of active research in the identification of high confidence, transcriptome-wide RNA binding sites: (1) advantages and disadvantages of UV and other crosslinking methods in

RBP:RNA complex capture (Figure 3-1), (2) recent efforts to optimize the efficiency of converting isolated RNA molecules into cDNA fragments for high-throughput sequencing (Figure 3-2), and (3) how incorporation of controls and normalization strategies enables proper ranking of binding sites and removal of artifacts (Figure 3-3). Although significant challenges still prevent the comprehensive identification and quantification of all RNA binding events in a cell, recent technical innovation has greatly enhanced our ability to identify RNA binding targets of hundreds of different RBPs from less than a million cells. With these improvements, RNA researchers are poised to dramatically expand the range of biological questions that can be addressed using transcriptome-wide profiling of protein-RNA interactions.

3.2 The development of transcriptome-wide discovery methods for protein-RNA interactions

As the need to identify transcriptome-wide protein-RNA interactions continues to grow, it is important for researchers to be aware of both the benefits and shortcomings of methods designed to capture and sequence protein-bound RNAs.

RNA immunoprecipitation (RIP)

RNA immunoprecipitation (RIP) followed by quantification on a microarray (RIP-chip) or with high-throughput sequencing (RIP-seq) was the earliest method, and has been used to provide insights into the biological function of RBPs (Tenenbaum et al., 2000, Keene et al., 2006). RIP methods involve cell/tissue lysis and immunoprecipitation of native RNA-RBP complexes with an antibody recognizing the RBP of interest. Since RNA-protein interactions

are not stabilized by a covalent crosslink, the stringency of the wash conditions needs to be carefully adjusted to ensure that direct protein-RNA interactions are maintained, while non-specifically bound RNAs are washed away. One drawback of not stabilizing direct interactions with a crosslink is that low-affinity or transient direct interactions are likely lost during the washing steps. There is also a debate as to what fraction of protein-RNA interactions can occur post-lysis depending on the lysis conditions, which may be difficult to distinguish from in vivo interactions(Penalva et al., 2004, Mili et al., 2004). Nevertheless, with the use of a total input RNA quantification from the lysate prior to immunoprecipitation, RIP provides a quantifiable binding score for enrichment of specific target RNAs (Tenenbaum et al., 2000).

'Binding-Site-resolution RIP'

Historically, RIP experiments could only identify transcripts bound by RBPs as the lack of an RNase digestion step prohibited the identification of specific binding sites within a transcript. Recently, a digestion-optimized RIP-seq method ('DO-RIP') was developed to investigate protein-RNA interactions with binding-site resolution (Nicholson et al., 2016). Using HuR as a proof of concept, the authors identified digestion conditions that resulted in the isolation of RBP-protected RNA fragments of 20 to 70 nucleotides, allowing binding-site mapping at high resolution. By using control samples that include total input RNA and a negative control IP with non-specific antibodies, DO-RIP quantitatively identifies transcriptome-wide protein binding sites by assigning relative enrichment scores that rank the protein occupancy of identified sites.

Crosslinking followed by immunoprecipitation (CLIP)

Crosslinking followed by immunoprecipitation (CLIP) was developed to enable more stringent purification of protein-RNA complexes (Ule et al., 2003). CLIP reduces the recovery of off-target RNAs indirectly associated with the RBP of interest by permitting more stringent washes of the protein-RNA complex, which disrupts protein-protein interactions. Immunoprecipitated RNAs are then treated with an optimized concentration of RNase to create short RBP-protected RNA fragments 20 to 70 nucleotides in length. CLIP was the first high-throughput method to identify transcriptome-wide protein binding sites on RNAs and has been used for over a decade to profile many RNA binding proteins (Ule et al., 2003, Licatalosi et al., 2008, Konig et al., 2010).

3.3 Stabilization of protein-RNA interactions by crosslinking

The implementation of UV-crosslinking was an important breakthrough enabling higher stringency immunoprecipitation and identification of binding sites at high resolution. UV-crosslinking at 254 nm wavelength only creates a covalent bond between an amino acid residue and the RNA base if they are in very close proximity, a constraint that is typically only met with specific, direct interactions (Meisenheimer et al., 1997). While this selectivity is a key strength of the CLIP approach, even favorable interactions are inefficiently captured, with crosslinking yields ranging from 5% to less than 1% in most cases using standard CLIP conditions of 400 mJoules (Darnell et al., 2010). Increased crosslinking energy can allow for higher yields, but causes additional problems with RNA-RNA crosslinks and RNA damage that may limit library generation and the ability to map sequenced fragments to the genome.

Inherent biases of 254nm UV-crosslinking

Although commonly used due to its simple procedure and applicability to unmodified cells or tissues, standard UV-crosslinking does have known biases that could affect interpretation of downstream results. First, in vitro biochemical studies suggest that there are biases in crosslinking efficiency for specific nucleotides and amino acid residues: notably, pyrimidines are more photoactivatable than purines. Similarly, while all amino acids are viable substrates for crosslinking, their reactivity is highly variable (with Cys, Lys, Phe, Trp and Tyr residues crosslinking with the highest efficiencies and His, Glu, and Asp crosslinking with moderate efficiency (Meisenheimer et al., 1997)). In addition, it is theorized that RBPs that bind double-stranded RNAs crosslink particularly poorly because the deep and narrow groove of A-form helical structures generally preclude access of amino acids to the nucleotide. Thus, depending on the nature of the RBP-RNA interaction, crosslinking will capture some more efficiently than others, while still others are missed entirely. These limitations have led to the development of alternative methods to stabilize protein-RNA interactions, increasing efficiency and widening the scope of protein-RNA interactions amenable to capture (summarized in Figure 3-1 and detailed below).

Increasing crosslinking efficiency

Photoactivatable ribonucleoside CLIP (PAR-CLIP) uses metabolic RNA labeling in cell culture to incorporate UV-reactive nucleoside analogs, such as 4-thiouridine or 6-thioguanosine, which crosslink at a wavelength of 365 nm (Hafner et al., 2010) (Figure 3-1). RNA yield after crosslinking is increased by the UV-reactive nucleosides compared to traditional UV-crosslinking (Hafner et al., 2010, which has driven high adoption of the method (Kishore et al., 2013, Maatz et al., 2017, Hamilton et al., 2016). PAR-CLIP has the added benefit that as many as 70% of

reads contain a T to C mutation after reverse transcription at the crosslinking site of 4-thiouridine (Hafner et al., 2010), which provides additional confidence in identified binding sites. However, the use of photoactivatable ribonucleoside analogs is not without drawbacks. As only the supplied nucleotide will be available for crosslinking, motifs and crosslinking sites are limited to sequence regions that contain this nucleotide. RNAs with high turnover rates will be overrepresented due to greater incorporation of the nucleoside analog on a short time scale, and many photoactivatable ribonucleoside analogs have shown toxicity in various cellular contexts (Burger et al., 2013).

Non-UV techniques to crosslink protein-bound RNA

Given that UV-crosslinking requires close association between an amino acid and a nucleotide residue, it is likely inefficient for dsRNA binders that interact with structural features of the backbone and have low accessibility of the nucleotide base (Singh et al., 2014). Although UV-crosslinking has successfully been used to assay double-stranded RNA-protein interactions with hiCLIP (Sugimoto et al., 2015) and CLASH (Helwak et al., 2016), on average less than 2% of reads from the resulting sequencing libraries contain the desired ligation events between the two interacting strands of RNA. Thus, there remains a need for the development of improved capture methods to profile RBPs that do not interact directly with a nitrogenous base.

Intercalating agents

One potential approach to specifically capture double stranded RNA (dsRNA)-RBP interactions is with treatment of methylene blue. Methylene blue intercalates between the bases of dsRNA to open up the RNA structure and allow crosslinking of dsRBPs in the presence of visible

light (Liu et al., 1996) (Figure 3-1). This method is highly specific for RBP-dsRNA interactions and is estimated to have 10-15% efficiency in vitro (Liu et al., 1996). Methylene blue treatment could be used in combination with 254nm UV crosslinking in vivo to capture both single and double-stranded RNA interactions, or performed in independent experiments to distinguish single and double stranded RNA interactions of the same protein.

Chemical crosslinking

Another alternative to UV is to perform crosslinking with low concentrations (0.1%) of formaldehyde, a reversible crosslinking agent (Figure 3-1)(D GH et al., 2016). Formaldehyde has long been used to map protein-DNA interactions, and was recently used to profile both RNA binding of proteins that have classically been characterized as chromatin modifiers (D GH et al., 2016) as well as direct RNA-chromatin interactions (Engreitz et al., 2014). However, as formaldehyde also crosslinks protein-protein interactions, the concern of indirect crosslinking to large ribonucleoprotein complexes (e.g. ribosomes, RNA polymerase II, P-bodies, or other RNP granules) has generally limited its use for studying RBPs. Additionally, at low formaldehyde concentration (10-fold lower than field standard for ChIP), the question of overall crosslinking efficiency has yet to be addressed.

Marking binding sites through direct RNA modification

The limitations of the above crosslinking methods have led researchers to explore non-crosslinking based approaches to profile RBP-RNA interactions by direct modification of target RNAs. One such method, TRIBE, fuses the deaminase domain of ADAR to the RBP of interest to cause ectopic RNA editing near the sites of RBP binding (McMahon et al., 2016)(Figure 3-1).

These editing sites can be detected by direct sequencing of RNAs compared to an endogenous, untagged control to identify ADAR-RBP fusion-dependent RNA editing events. A major advantage of this technique is that the editing readout is sequencing based (as A-to-I changes modify the cDNA sequence), rather than IP based, eliminating the requirement of crosslinking and immunoprecipitation (McMahon et al., 2016). However, substrate specificity inherent to the ADAR deaminase domain could impose a bias in the selection and efficiency of deamination at specific binding sites. Additionally, this method requires expression of the RBP-ADAR fusion protein, very deep sequencing, and a robust computational method for identifying editing sites. Another approach utilizes in vivo RNA modification by fusing the RBP of interest to a “bio” tag that recruits BirA ligase to biotinylate nearby RNAs (He et al., 2010) (Figure 3-1). Biotinylated RNAs can then be purified in stringent conditions with a streptavidin pull-down to isolate and sequence RNA sequences located near protein binding sites. Further work will be necessary to explore which RBPs would be suitable for this approach, and whether there are biases in biotinylation frequency.

3.4 Methods to select for a protein of interest

Antibody-based immunoprecipitation

The RBP of interest is enriched in a CLIP experiment by immunoprecipitation with a monoclonal or polyclonal antibody. While not all RBPs have an IP-grade antibody, many groups are working to generate databases of validated antibodies to expand the pool of RBPs that can be immunoprecipitated (Bjorling et al., 2008, Sundararaman et al., 2016). Proteins for which there is no validated IP-grade antibody requires fusion to a peptide tag for IP at either the N or

C terminus. This approach has been widely used in cell culture and in model organisms by both overexpression of a transgene or modification of the endogenous genomic locus. Regardless of the antibody used, a control IP with an IgG isotype-only control should be performed in parallel to provide a specificity metric for the IP. Additionally, there is variation in the specificity of antibodies raised against different epitope tags, and therefore it is important when using tags to perform a negative control IP on a sample that does not contain the tagged protein (Van Nostrand et al., 2016).

Non-Antibody Based Pulldown

CLIP protocols include a protein gel purification step in part because washing of the immunoprecipitated protein-RNA complex does not remove all indirect protein interactions; the strength of the protein-antibody interaction limits the stringency of washing. Other strategies have been developed to increase the specificity of the IP and remove the need for additional gel purification of the complexes. Singh et al. used a double IP strategy with two different antibodies in succession to increase the specificity and purity of the final sample (Singh et al., 2014). The double antibody selection without gel purification produced a library with the same purity as a single IP followed by gel purification. This method has the added benefit that it can either be applied to a single protein with two antibodies, or a protein complex in which two different components of the complex are targeted for IP. More recent iterations of tagging strategies are designed to increase the strength of protein capture to that of a covalent interaction, to allow for extremely harsh washing and complete removal of non-covalently linked protein and RNA species. Protein purification tags (such as the HIS, Bio/BirA (de Boer et al., 2003), or TAP-TAG (Gerace et al., 2015)) can be inserted next to a protein at the genomic locus

with CRISPR and allow for stringent protein purification conditions. The recently developed Halo Tag has extremely strong and highly specific interactions to the HaloLink resin for protein capture, although its large size (297 amino acids), can potentially disrupt nearby protein-protein interactions in an endogenous setting (Daniels et al., 2012). These and other tagging strategies hold promise for both increasing the signal to noise ratios in CLIP-seq data and eliminating the need for gel purification of the protein-RNA complexes.

3.5 Eliminating Radioactivity from CLIP methods

Replacing radiolabeling with fluorescence

Preparation of a CLIP library requires the optimization of RNA fragmentation after IP to ensure that the resulting RNA fragments are long enough to be uniquely mappable to the reference genome, but short enough to identify binding sites at high resolution. Historically, this has been performed by titrating RNase, resolving the RBP-RNA complexes by SDS PAGE gels, and autoradiographic visualization of radiolabeled RNA. RNase treatment results in a characteristic smear above the size of the protein of interest. As an additional control for IP specificity, at the highest RNase conditions all digested fragments would resolve to a single band at the size of the protein of interest. However, radiolabeling posed an inconvenience for widespread adoption of CLIP, and recent advances have eliminated it. For example, irCLIP uses an infrared-dye-conjugated and biotinylated RNA adapter that can be imaged with a digital fluorescence imager (Zargnegar et al., 2016). This adapter enables visualization of RNAs under different digestion conditions with a digital scanning readout that does not require radioactive materials.

Standardized RNA fragmentation

RBPs bind RNAs of varying size, and different cell types contain varying levels of endogenous RNases. As a result, RNA trimming should be optimized for each experiment. For large-scale enhanced CLIP (eCLIP) experiments performed as part of the ENCODE (https://www.encodeproject.org) efforts, we explored the requirement of customized fragmentation conditions for every CLIP experiment. In a single cell type, we tested a wide range of RNase concentrations for two RBPs representative of the extremes of RNA target lengths: RBFOX2, which binds intronic regions within pre-mRNAs that can be hundreds of kilobases in length, and SLBP, which exclusively binds the 3' untranslated regions of the 150nt intronless histone mRNAs. The total number of binding clusters and their distribution across and within genic regions (i.e., intronic regions, coding sequence, 5' and 3' UTRs) were surprisingly robust to the extent of RNA digestion (Van Nostrand et al., 2016). Therefore, once endogenous levels of RNase have been accounted for in the cell/tissue type of interest, a single, optimized concentration of RNase is often appropriate to yield informative binding profiles for most RBPs.

3.6 Improving the recovery rate of RNAs prepared for sequencing

Measuring PCR duplication

In early CLIP protocols, the low amount of RNA recovered after crosslinking and immunoprecipitation, coupled with inefficiencies in enzymatic reactions during library preparation, led to the need for many cycles of PCR amplification to generate sufficient material for sequencing. This often resulted in libraries of low sequence complexity, i.e. containing a large fraction of duplicated reads. We estimated the duplication rate from publicly available CLIP datasets and

found that on average, a staggering 83.8% of sequenced reads were flagged as PCR duplicates (Van Nostrand et al., 2016). To illustrate, a standard sequencing library requires 100 fmoles (6 x 10^{10}) DNA molecules. If 25 PCR cycles are required to produce this amount, and one assumes 80% PCR efficiency, then the initial unamplified library only contained 25,000 molecules ($6x10^{10}$ / (2^{25} x 0.8)). Therefore, only up to 25,000 reads in the final sequenced library will have originated from unique molecules and all additional reads are attributable to PCR duplicated sequences. To precisely quantify and remove PCR duplicated reads, adaptors containing short random sequences are ligated to the fragments during library preparation to uniquely tag each RNA fragment prior to PCR amplification. These unique molecular identifiers (UMIs) enable accurate classification of unique and duplicated reads by comparing the mapped genomic coordinates of reads that contain the same UMI (Konig et al., 2010, Darnell et al., 2011, Smith et al., 2017).

Addressing reverse transcription termination at crosslink sites

Increasing the total yield of recovered cDNA has led to some of the greatest improvements in library complexity. In the first generation CLIP protocols, such as HITS-CLIP (Licatalosi et al., 2008), both 5' and 3' RNA adapters are ligated prior to reverse transcription (RT) (Figure 3-2). After first-strand cDNA synthesis, the sample is PCR amplified with primers complementary to a portion of the adapters ligated at the ends of the RNA. Since UV-crosslinking chemically modifies the nucleotide bridging the RBP-RNA crosslink, RT enzymes are prone to termination at the crosslink site, and as many as 80% of the resulting cDNA products lacked the 5' adapter and therefore lacked the 5' primer binding site (Sugimoto et al., 2012). These sequences fail to be PCR amplified and are thus lost from the sequencing library. Individual nucleotide resolution

CLIP (iCLIP) and subsequent methods (eCLIP, irCLIP) addressed this issue by performing the second adapter ligation after RT, such that the second adapter is ligated to all cDNA fragments regardless of the RT termination site (Figure 3-2). Because a subset of reads terminate at the RT stops, this approach has the added advantage that the end of the reads, after genome mapping, mark crosslink sites and therefore enable binding sites to be identified at single nucleotide resolution (Konig et a;., 2010).

Increasing enzymatic reaction efficiency

In addition to recovering more RNAs by adding the second adapter after the RT step, improvements in enzymatic reaction efficiencies have greatly increased the yield of unique fragments prior to PCR amplification. Zarnegar et al. optimized the reaction condition for each step of library preparation by using an infrared-dye-conjugated adapter to quantify the amount of RNA before and after each reaction with a dot blot (Zarnegar et al., 2016). The resulting irCLIP libraries have a much higher yield of unique library fragments and therefore require a small number of PCR cycles to maintain a high complexity library. Similarly, the eCLIP method employed highly optimized reaction conditions to improve the overall yield of sequencing libraries by 1000-fold over libraries generated with the iCLIP method (Van Nostrand et al., 2016). These improvements have now made it possible to reliably generate informative CLIP libraries from limited amounts of starting material (less than one million cells (Zarnegar et al., 2016)), and at scale (hundreds of RBPs (Van Nostrand et al., 2016)).

3.7 Controlling for Background RNA Signal

As RNAs are dynamic in their expression, subcellular localization, and structural features, accurate quantification of protein bound RNAs has been quite challenging in CLIP-seq experiments. In contrast, the fixed number of copies of DNA elements in the nucleus allows for a much simpler background for robust quantifications of DNA binding from ChIP-seq experiments. Therefore, it is crucial to use control experiments in CLIP that account for background noise and variation in RNA abundances.

Input Control for Background

Washing and gel purification of the protein-RNA complex removes the majority of proteins bound nonspecifically. However, washing still does not result in a perfectly pure sample. To address this, the eCLIP protocol has incorporated a size-matched input to capture nonspecific, background RNA binding events in a CLIP experiment (Figure 3-3). The size-matched input contains 2% of the cell lysate that has been crosslinked, run on a gel, transferred to a membrane, and cut at the same size range as the IP (up to 75 kDa above the protein of interest). This input fraction has the same crosslinking, fragmentation, ligation, and amplification biases as the IP sample. Therefore it is an important control for highly abundant, sticky RNAs that are bound by many proteins and get called as binding sites in many CLIP experiments (Van Nostrand et al., 2016). An enrichment score of reads in the IP relative to the size-matched input for a given peak location provides a metric for the specificity of a binding event relative to background (Figure 3-3). This metric enabled quantification of enriched binding even at abundant transcripts commonly considered as artifacts across CLIP experiments, such as MALAT1 (Van Nostrand et al., 2016). It is important to note that the enrichment score calculated here is not

the same as what has been historically used for RIP experiments. The main difference is that the input in a RIP sample is total RNA, whereas the input used for eCLIP contains only the pool of crosslinked RNAs that are bound by proteins that run within the same size range on the gel. Therefore the eCLIP input captures the experimental background in the individual CLIP experiment, rather than providing an enrichment score relative to total RNA. A technical limitation that remains in CLIP-seq experiments is the ability to address in vivo protein binding occupancy on transcripts. Quantification of binding occupancy with CLIP methods is confounded by IP efficiency, crosslinking efficiency, and the inability to simultaneously measure RBP binding and RNA abundance in a sample. Therefore, if two peaks are identified with the same enrichment score on two independent transcripts, it is not yet possible to determine which one correlates to a higher stoichiometric binding occupancy at its respective site. However, this question can be critical to addressing the functional relevance of binding sites. For example, if one transcript has 20 copies in a cell and all 20 are bound by a protein, that is likely a different mode of regulation than a binding site on a transcript that has 200 copies in a cell, 20 of which are bound by protein. It remains an open challenge to identify modifications to the CLIP procedure that would enable such quantitative estimates of binding occupancy.

The next step: identification of functional binding sites

With the plethora of binding sites identified by RIP and CLIP-related technologies, there is an urgent need to evaluate and prioritize the function of these binding sites. Key questions that need to be addressed include: (1) Are all binding sites functional? (2) What are the properties of functional binding sites? (3) How do we define, measure, or experimentally validate functional binding sites? Orthogonal assays to measure RBP function can address these questions

(Van Nostrand et al., 2016). For example, a splicing map can be generated from an analysis of alternative splicing to compare binding location relative to exons that are alternatively spliced upon knockdown of a protein of interest (Ule et al., 2006). Emerging high-throughput screening techniques using CRISPR genome editing, tethering (Bos et al., 2016), and other methods to assay RNA processing allows for the manipulation of specific binding sites to evaluate function. To test the function of a given binding site, antisense oligonucleotides (ASOs) can be introduced to inhibit protein binding (Lovci et al., 2013), or CRISPR genome-editing can be used to modify the binding site and test the resulting behavior of the endogenous RNA target. Tethering strategies using MS2 hairpins or RNA-guided Cas9 (Nelles et al., 2016) can be used to probe RBP functions, such as splicing or mRNA degradation, by tethering the protein to a previously unregulated site. Comprehensive RBP-RNA interaction maps annotated with regions where binding is associated with a validated cellular function should serve as a valuable tool for engineering synthetic RNAs to direct proper packaging, expression, and behavior of particular RNAs for therapeutic use.

3.8 Conclusion

The ability to identify transcriptome-wide RBP binding sites with CLIP and RIP technologies has played a critical role in our ability to understand molecular mechanism of RBP function. In the past few years, dramatic improvements in library preparation efficiencies and removal of radioactivity from CLIP have led to more widespread adoption of CLIP. The incorporation of an input sample has greatly increased the ability to distinguish true binding sites from background signal in an experiment. Future improvements are needed to: 1) increase the efficiency of protein-RNA capture with methods other than UV-crosslinking, 2) increase the strength of

protein capture to allow for higher stringency washes without the need for gel purification, and 3) develop orthogonal assays to determine functional binding sites. These improvements will enhance our ability to identify and interpret functionally relevant RNA elements across the transcriptome.

3.9 Acknowledgements

We would like to thank all members of the Yeo Lab for review and critique of this manuscript with particular acknowledgment to Julia Nussbacher, Stefan Aigner, Mark Perelis and Ryan Marina. This work was supported by grants from the NIH (HG004659 and NS075449) to GWY. E.C.W. is supported by grants from the University of California, San Diego, Genetics Training Program (T23, GM008666) and the NSF Graduate Research Fellowship Program. E.L.V.N. is a Merck Fellow of the Damon Runyon Cancer Research Foundation (DRG-2172-13).

Chapter 3, in full, is a reprint of material as it appears in: **Wheeler, EC.***, Van Nostrand EL, Yeo GW. Advances and challenges in the detection of transcriptome-wide protein-RNA interactions. Wiley Interdiscip Rev RNA 9, doi:10.1002/wrna.1436 (2018). The dissertation author was first author of this paper.

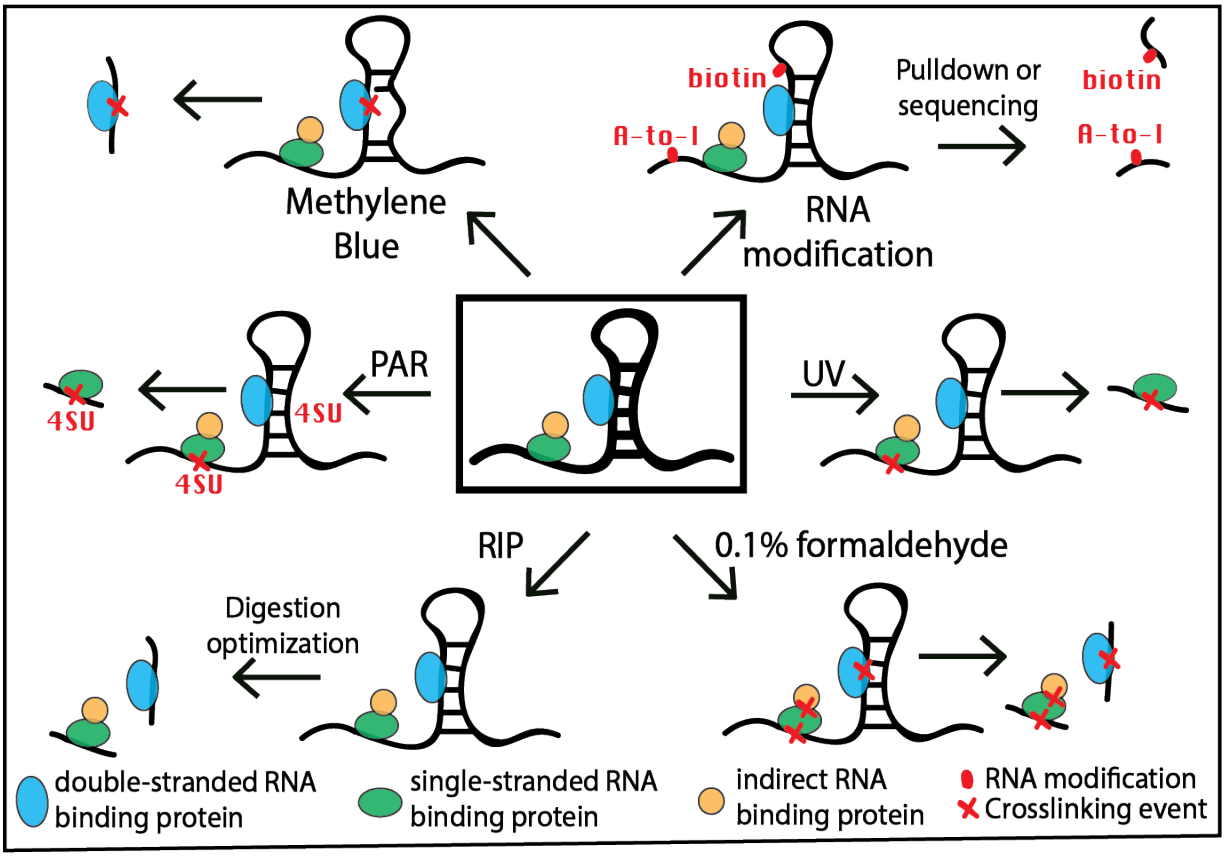
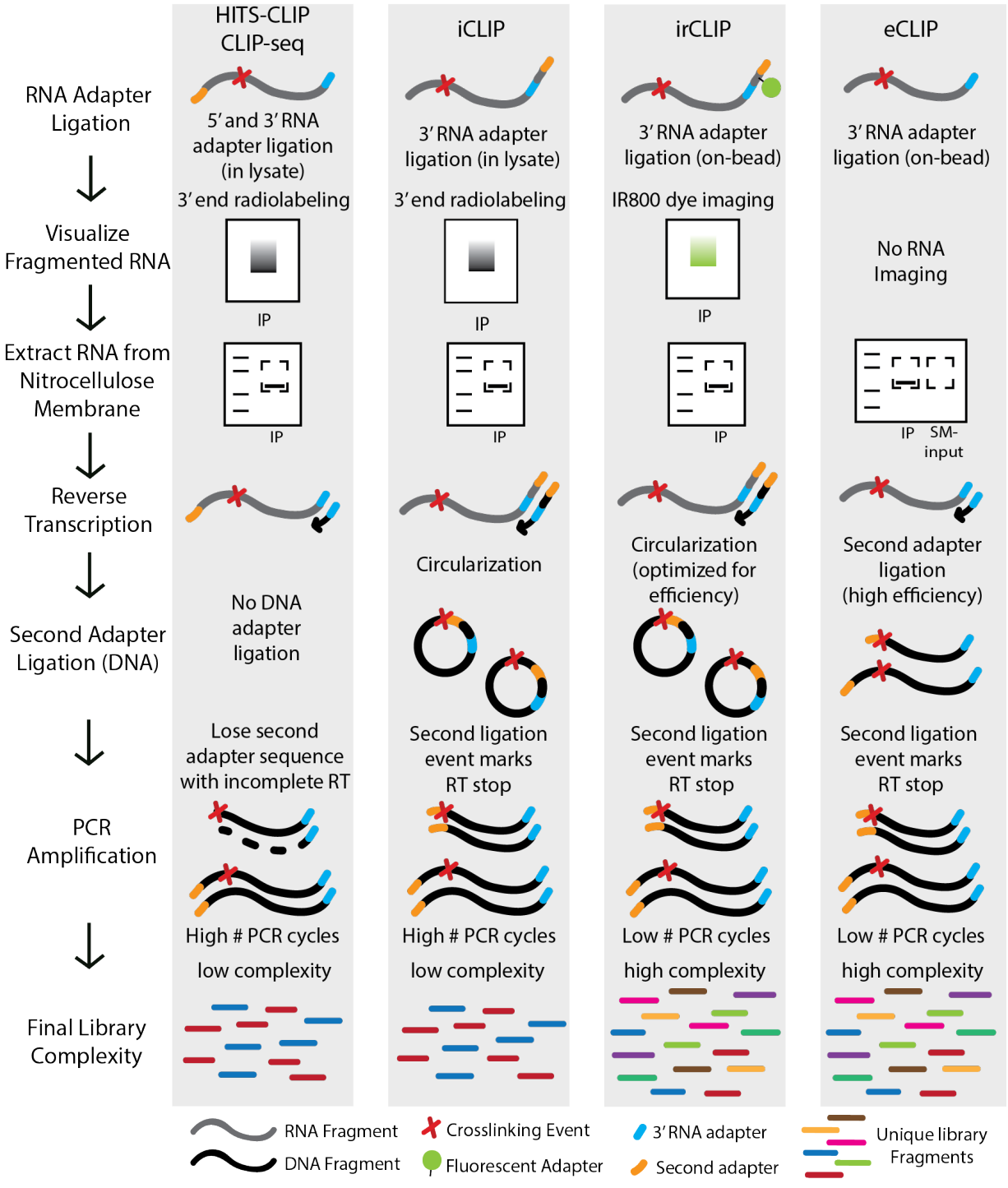


Figure 3.1: Methods to capture protein-RNA interactions. Different techniques are required to capture single-stranded (green), double-stranded (blue), and indirect (yellow) RNA interactions. Crosses (X) in red mark RNA sites that are crosslinked to the RNA binding protein. (right) UV treatment at 254 nm preferentially captures binding in single-stranded regions. (bottom right) 0.1% formaldehyde treatment captures all protein-protein and protein-RNA interactions. (bottom left) RNA immunoprecipitation (RIP) uses a native pulldown (no crosslinking) to capture binding events with antibody selection. Optimized RNA digestion conditions can reveal specific binding sites with RIP. (left) Photoactivatable ribonucleoside analog treatment (PAR) increases UV crosslinking efficiency at 365 nm. (top left) Methylene blue intercalates between the bases of double-stranded RNA to allow crosslinking in double-stranded regions in the presence of visible light. (top right) Protein-RNA interaction sites are marked by exogenous RNA modifications. This requires creating a fusion protein to modify RNA near binding sites with biotinylation (BioTag-BirA) or A-to-I RNA editing (ADAR).

Figure 3.2: Comparison of CLIP-seq library preparation protocols. HITS-CLIP/CLIP-seq perform adapter ligation on both ends of the RNA while other protocols only ligate an RNA adapter on the 3' end. RNA digested fragments are visualized with radiolabeling (HITS-CLIP/CLIP-seq and iCLIP), infrared dye imaging (irCLIP), or not at all (eCLIP). RNA is then transferred and isolated from a nitrocellulose membrane. In the case of eCLIP, a size-matched input sample is excised to control for background. For all, reverse transcription (RT) generates cDNA using the 3'RNA adapter as the priming site. As the RT enzyme commonly terminates at crosslinking sites, there will be a mixture of full length and truncated fragments. iCLIP and irCLIP have a circularization step to put the second adapter on the DNA fragment. Ligation in irCLIP has been highly optimized to improve efficiency. eCLIP uses a second ligation step that has also been optimized for efficiency. HITS-CLIP/CLIP-seq fragments that have incomplete RT cannot PCR amplify due to the loss of the second adapter sequence. For all, cDNA fragments are then PCR amplified to generate enough material for sequencing. Recent methods (irCLIP, eCLIP) have routinely high complexity libraries generated from a low number of PCR cycles.


HITS-CLIP
CLIP-seq
iCLIP
irCLIP
eCLIP
RNA Adapter Ligation
5′ and 3′ RNA adapter ligation (in lysate)
3′ RNA adapter ligation (in lysate)
3′ RNA adapter ligation (on-bead)
3′ RNA adapter ligation (on-bead)
3′ end radiolabeling
3′ end radiolabeling
IR800 dye imaging
Visualize Fragmented RNA
No RNA Imaging
IP
IP
IP
Extract RNA from Nitrocellulose Membrane
IP
IP
IP
IP SM-input
Reverse Transcription
Circularization
Circularization (optimized for efficiency)
Second adapter ligation (high efficiency)
Second Adapter Ligation (DNA)
No DNA adapter ligation
Second ligation event marks RT stop
Second ligation event marks RT stop
Second ligation event marks RT stop
Lose second adapter sequence with incomplete RT
PCR Amplification
High # PCR cycles
low complexity
High # PCR cycles
low complexity
Low # PCR cycles
high complexity
Low # PCR cycles
high complexity
Final Library Complexity
RNA Fragment
DNA Fragment
Crosslinking Event
Fluorescent Adapter
3′ RNA adapter
Second adapter
Unique library Fragments

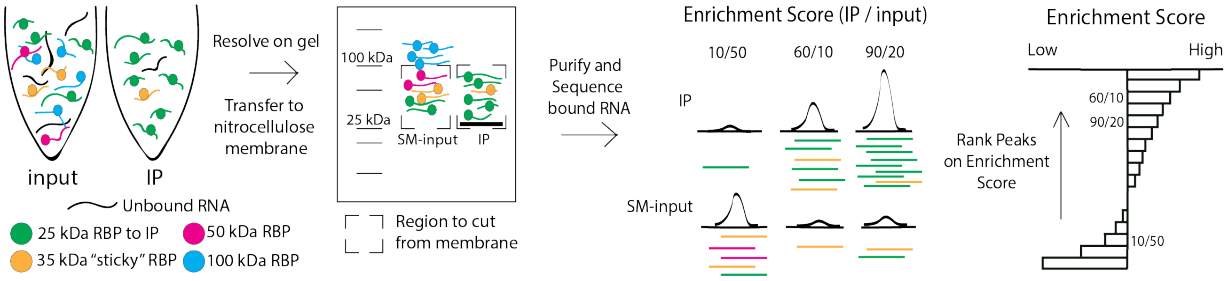


Figure 3.3: 2% of the lysate is taken prior to IP as the input sample. "Sticky" RBPs (yellow) are not completely purified away and contaminate the IP sample. The input and IP are run in parallel on the protein gel and extracted from the nitrocellulose membrane at the same size range. Called peaks are then normalized by dividing the number of reads in the IP by the number of reads in the input to remove signal coming from background RNAs (yellow). The enrichment score is a rank-based metric for specificity of binding.

Chapter 4

SRSF2 and *U2AF1* mutations drive MDS through alternate *GNAS* isoform usage and cooperate with *GNAS* hotspot mutations

Splicing factors (SF) are the most frequently mutated genes in myelodysplastic syndrome (MDS), but the critical downstream disease drivers remain unknown. Here, we developed human isogenic induced pluripotent stem cell (iPSC) models of the canonical *U2AF1 S34F* and *SRSF2 P95L* mutations. Integrated RNA-sequencing and eCLIP analyses in iPSC-derived hematopoietic stem/progenitor cells (HSPCs) revealed transcriptome-wide alterations in pre-mRNA splicing caused by altered RNA binding preferences of the mutant SFs. Intersectional analysis of altered splicing and differential RNA binding in SF-mutant iPSC-HSPCs and MDS pa-

tient cells identified the GNAS long isoform (*GNAS-L*) as a prominent alternatively spliced transcript common to both SF mutations. Overexpression of *GNAS-L* in normal hematopoietic cells reproduced the MDS phenotype. Conversely, suppressing the *GNAS-L* isoform rescued the differentiation defects in SF-mutant cells. MDS patient population studies, as well as biochemical and functional analyses reveal cooperation between the *GNAS-L* isoform and the GNAS R201 activating mutation. Our findings highlight G protein signaling as a convergent consequence of SF mutations, identifying a potential therapeutic avenue for SF-mutant MDS.

4.1 Introduction

Splicing factor (SF) gene mutations are the most common class of mutations in MDS, present in nearly 60% of MDS patients. SF mutations are early, potentially initiating events in the course of the disease and can thus provide attractive putative therapeutic targets (Papaemmanuil et al., 2013). Transcriptome-wide splicing alterations, consistent with altered binding of the mutant SFs, have been found in several cell and mouse models and patient cells (Dvinge et al., 2016). While splicing modulation may offer a therapeutic window, the effectiveness of such approaches in the clinic is highly uncertain (Lee et al., 2016, Chang et al., 2018, Seiler et al., 2018). Thus, the identification of key effector pathways of these mutations remains a top priority towards developing targeted therapies.

Serine/arginine-rich splicing factor 2 (SRSF2) is a member of the serine/arginine-rich (SR) protein family that regulates splicing by promoting inclusion of exons through binding to exonic splicing enhancer (ESE) sequences. Mutant SRSF2 preferentially recognizes 'CCNG'- versus 'GGNG'- containing ESEs, whereas wild-type (WT) SRSF2 binds to both with similar affinity (Kim et al., 2015, Zhang et al., 2015). U2 small nuclear RNA auxiliary factor 1 (U2AF1)

is a U2 auxiliary factor protein that recognizes the AG splice acceptor dinucleotide at the 3' end of introns (Dvinge et al., 2016). Mutant U2AF1 shows preference for 3' splice sites containing 'CAG'- versus 'UAG'- sequences, promoting preferential inclusion of the respective exons (Okeyo-Owuor et al., 2015, Ilagan et al., 2015).

Canonical SF mutations in MDS are always heterozygous and mutually exclusive to one another. While synthetic lethality between SF mutations has also been demonstrated (Lee et al., 2018), convergence in common downstream targets at the isoform, gene or cellular pathway level, has long been hypothesized to underlie their mutual exclusivity. In particular, convergence at the isoform level appears more likely in the case of *SRSF2* and *U2AF1* mutations, as both affect the 3' end of splicing. Efforts to identify common effectors of SF mutations in large-scale RNA-sequencing datasets from MDS patient cells and from murine and cellular models with SF mutations have revealed largely non-overlapping changes (Pellagatti et al., 2018, Shiozawa et al., 2018). However, these datasets have important limitations: primary patient cells have substantial heterogeneity due to co-occurring mutations and clonal composition; alternative splicing (AS) events are largely non-conserved between mouse and human (Yeo et al., 2005); and cellular models engineered in aneuploid leukemia cell lines express mutant SFs at non-physiological levels and stoichiometry.

To identify downstream effects common to *SRSF2* and *U2AF1* mutations, we used integrated transcriptome and eCLIP analyses in HSPCs from isogenic gene-edited iPSC models (iPSC-HSPCs) of the two mutations. We identify the *GNAS-L* isoform as an AS event common to both *SRSF2* and *U2AF1* mutations in iPSC-HSPCs and MDS patient cells and provide functional confirmation of its implication in driving the MDS phenotype. We further show that *GNAS-L* encodes a more active Gαs form and that it cooperates with the *GNAS* R201 hotspot

mutation in activating Gαs, a finding that may provide a mechanistic basis for the statistically significant co-occurence of *SRSF2* mutations with *GNAS* mutations in MDS patients that we also uncover. While *GNAS* R201 mutations activate the canonical cAMP/PKA signaling pathway, we find increased expression of the *GNAS-L* isoform associated with ERK pathway activation. These results pinpoint Gαs activation as a key effector of SF mutations in MDS.

4.2 Results and discussion

Generation of a stem cell model with splicing factor mutations

To interrogate the effects of *U2AF1* and *SRSF2* mutations and identify common downstream effectors driving MDS phenotypes, we used CRISPR/Cas9 to introduce a heterozygous U2AF1 S34F mutation in a normal iPSC line (N-2.12)(Kotini et al., 2017) (Figure 4-1A). Multiple mutant iPSC lines were generated with two different gRNAs to exclude potential confounding effects of off-target events on cellular and molecular phenotypes. Heterozygous *SRSF2* P95L mutation was previously introduced in the same parental iPSC line (Chang et al., 2018). Homozygous mutant clones were not recovered, in agreement with the previously proposed dependency of SF-mutant cells on one WT allele (Lee et al., 2016). To enable allele-specific immunoprecipitation, an epitope tag (3xFLAG) was introduced at the C-terminus of the endogenous locus of either the mutant allele (in SF-mutant lines) or one WT allele (in the isogenic SF-WT iPSC lines) (Figure 4-1B-E, Figure 4-2A).

Directed differentiation of independent lines of each genotype revealed no defects in hematopoietic specification, as indicated by the emergence of CD34+ and CD45+ HSPCs (Figure 4-2B-D). SF-mutant iPSC-HSPCs displayed decreased proliferation, as previously reported

in various cell line and mouse models of SF mutations (Chang et al., 2018, Saez et al., 2017) (Figure 4-2E). The number of colonies generated from SF-mutant iPSC-HSPCs in methylcellulose assays was reduced to approximately half of the number generated from WT lines (Figure 4-2F). These results show that SF mutations cause a differentiation defect, recapitulating phenotypes observed in MDS patient-derived iPSCs and ex vivo cultured primary MDS patient cells (Kotini et al., 2015, Hsu et al., 2019, Flores-Figueroa et al., 1999).

Alternative Splicing landscape of SF-mutant HSPCs

To evaluate the impact of these engineered SF mutations on RNA splicing, we performed RNA-seq analyses in sorted CD34+/CD45+ iPSC-HSPCs from at least 3 independent lines per genotype and quantified alternative splicing (AS) events (Figure 4-3A-B). Skipped exon events were the predominant AS event type in both genotypes, while alternative 3' and 5' splicing events were detected at a lower frequency (Figure 4-3C-D). *'CAG'* and *'AAG'* 3' splice site sequences were enriched in exons preferentially included in U2AF1-mutant cells ("S34F exons"), while exons preferentially skipped in U2AF1-mutant cells ("WT exons") were enriched for *'UAG'* 3' splice site sequences (Figure 4-4A-C). The same motif preferences were also identified at the 3' splice site of alternative 3' splicing events preferentially included ('CAG/AAG') and skipped ('UAG') in U2AF1-mutant compared to WT cells (Figure 4-5XX). Exons preferentially skipped in SRSF2-mutant cells ("WT exons") showed enrichment of 5-mers containing *'GGAG'* and *'GGUG'* sequences, while 5-mers containing *'GCAG'* and *'CCAG'* were enriched in exons preferentially included in SRSF2-mutant cells ("P95L exons") (Figure 4-4A,D-E). These sequence-specific splicing alterations are consistent with previous findings in other models and patient cells and establish that our iPSC models recapitulate the characteristic splicing alterations of SF-mutant

hematopoietic cells (Okeyo-Owuor et al., 2015, Zhang et al., 2015, Kim et al., 2015).

Differential RNA binding of wild-type and mutant SFs in HSPCs

To test the extent to which U2AF1 and SRSF2 mutations alter binding across the transcriptome, we performed eCLIP (Van Nostrand et al., 2016) in sorted CD34+/CD45+ iPSC-HSPCs from two independent iPSC lines per genotype with allele-specific immunoprecipitation of the epitope-tagged mutant (in *SRSF2*-mutant and *U2AF1*-mutant lines) or WT (in isogenic *SRSF2*-WT and *U2AF1*-WT lines) allele (Figure 4-6A-D). We identified 35,000 U2AF1 binding sites and 12,000 SRSF2 binding sites across the protein-coding transcriptome, the majority of which occurred within the expected RNA regions: 3' splice sites for U2AF1 (both WT and mutant) and protein-coding exonic regions for SRSF2 (both WT and mutant) (Figure 4-7A). Approximately half of the peaks of the mutant factors (54% of U2AF1 S34F peaks and 53% of SRSF2 P95L peaks) were shared with the WT factors (Figure 4-7B). To identify sequences preferentially bound by each mutant or WT factor, we calculated enrichment of all 6-mers within all peak regions. U2AF1-WT binding sites were enriched for 6-mers containing 'UAG', while U2AF1-S34F binding sites were enriched for 6-mers containing 'CAG' (Figure 4-7C). SRSF2-WT binding sites were enriched for 'GC' and 'GA'-rich 6-mers, while peaks bound by SRSF2-P95L were enriched for 6-mers containing 'CCUG' sequences (Figure 4-7D). These differential sequence preferences in binding between the WT and mutant forms of U2AF1 and SRSF2 mimic the sequence alterations found in skipped exon events by RNA-Seq, providing support that altered splicing is mediated through differential RNA recognition by mutant splicing factors.

To specifically interrogate differential binding in regulatory regions of skipped exons, we generated 'splicing maps' to visualize the position-specific effects of SF binding (Yee et al.,

2019). The binding density of U2AF1-WT was higher than that of U2AF1-S34F specifically at the 3' splice site of WT exons (skipped in mutant cells) (Figure 4-8A). Conversely, U2AF1-S34F had higher binding density than U2AF1-WT at the 3' splice site of S34F exons (included in mutant cells) (Figure 4-8A). We also observed increased binding density of SRSF2-WT within WT exons, but not within upstream or downstream control exons (Figure 4-8B). These results show differences in binding density between the WT and mutant factors specifically at the regulatory regions of skipped exons. Next, we grouped the binding events localized in the regulatory regions of skipped exons into 3 categories: (1) sites bound by the WT SF only (WT peaks); (2) sites bound by the mutant SF only (Mutant peaks); or (3) sites bound by both WT and mutant SF (Both peaks) (Figure 4-9A). We found that WT exons (positive delta PSI WT–S34F) more often contained WT peaks, and S34F exons (negative delta PSI WT–S34F) more often contained Mutant peaks (chi-square test, p=0.003). This trend was similar, albeit less strong, in the case of SRSF2, in which the overall number of peaks detected was much lower (Figure 4-9A). Importantly, these trends were not observed on downstream, unaltered exons (Figure 4-9B). This analysis, coupling alternative splicing to differential binding at the level of specific AS events, allowed us to identify events that also contain differential binding by mutant vs WT U2AF1 or SRSF2.

Convergence on Downstream Targets

While convergence in common downstream genes has been postulated to underlie the mutual exclusivity of SF mutations in MDS patients, no common alternatively spliced genes have been identified from existing datasets so far as candidate targets. We thus harnessed our isogenic model to find AS events common to both SF mutations. Of the 504 and 517 differential

AS events that we identified in *U2AF1* S34F and *SRSF2* P95L, compared to isogenic WT cells, respectively, 41 events, affecting 40 genes, were common to both genotypes (Figure 4-10A). We then quantified exon inclusion (delta PSI) for each of these events in previously published MDS patient data (Pellagatti et al., 2018). 20 of the 41 events had sufficient read coverage to allow evaluation of splicing and 15 of the 20 were alternatively spliced in the same direction in both *SRSF2* and *U2AF1* mutant cells compared to cells from MDS patients without SF mutations (Figure 4-10A). Of these, 3 AS events in genes *GNAS*, *PSMA4* and *ITGB3BP* also contained a differential eCLIP peak between U2AF1 S34F and U2AF1 WT at the 3' splice site of the alternatively spliced exon (Figure 4-10C).

We focused on *GNAS* (Guanine Nucleotide binding protein, Alpha Stimulating), the gene encoding the α (alpha) subunit of the stimulatory G-protein (Gαs), as it plays a major regulatory role in cell signaling with previous links to oncogenesis and has been found to be recurrently mutated in MDS (Bejar et al., 2011). Both SF mutations promoted the inclusion of exon 3, resulting in the preferential usage of a long isoform (*GNAS-L*), over a short isoform (*GNAS-S*), in both our iPSC-HSPC models and in MDS patient cells. (Figure 4-10D-F). Knockdown of *GNAS-L* with an shRNA specifically targeting exon 3 rescued the differentiation defect of SF-mutant iPSC-HSPCs to a level comparable to that of WT iPSC-HSPCs (Figure 4-11A-B). Conversely, overexpression of *GNAS-L* in WT iPSC-HSPCs decreased their differentiation potential by approximately half, reproducing the SF-mutant phenotype (Figure 4-11C). These results show that *GNAS-L* plays a critical role in driving the MDS phenotype.

Activity of *GNAS-L* in combination with point mutation

GNAS mutations in MDS and other tumors most commonly involve the R201 residue and result in constitutive activation by maintaining Gαs GTP-bound (Sprang 2016). We thus hypothesized that *GNAS-L*, promoted by SF mutations, may also encode a more active form of the protein to confer a selective advantage in cancer. Signaling activity through Gαs is controlled by many factors including: GTP affinity, GTP hydrolysis, and GDP release. Exon3 is located in a region of the protein that forms a hinge between the open and closed conformation, a structural change that is essential for GDP release and re-association with GTP (Dror et al. 2015). Therefore, we measured stability of the open conformation and affinity of GTP binding in purified Gαs-L and Gαs-S proteins with and without R201 mutation. These data showed that Gαs-L is more stable in the open conformation, and has a higher affinity for GTP than Gαs-S (Figure 4-12A,D-E). Surprisingly, the R201 mutation further increased both the stability of the open conformation (Figure 4-12B-C) and GTP affinity of Gαs-L, but not Gαs-S (Figure 4-12D-E). These results are consistent with Gαs-L being more active than Gαs-S and suggest cooperation between the long isoform and point mutation in activating Gαs.

In support of cooperativity, interrogation of published datasets from a total of 3,298 MDS and AML patients revealed significant co-occurrence between SRSF2 and GNAS mutations (odds ratio = 4.1; p = 0.01)(Papaemmanuil et al., 2016, Papaemmanuil et al., 2013, Haferlach et al., 2014) (Figure 4-12F). To test this further, we compared the effects of ectopic expression of GNAS-L or GNAS-S with and without the R201H mutation in colony formation in WT iPSC-HSPCs. Both GNAS-L and the R201H mutation decreased colony formation and their combination had an even more pronounced effect, consistent with cooperation (Figure 4-12G).

Activated Gas stimulates adenylyl cyclase and conversion of ATP to cAMP, which in turn

activates protein kinase A (PKA) (Mealiea et al., 1990). Although surprisingly, we did not detect an increase in cAMP or PKA activity in cells expressing the *GNAS-L* isoform (data not shown). In turn, we speculate that expression of *GNAS-L* may activate ERK signaling, although the point mutation does not amplify this effect (data not shown). Future work will address the signaling pathways that are regulated through *GNAS-L* expression, to identify druggable pathways that may be effective to treat SF-mutant MDS patients.

Discussion

While the GNAS AS event that we report here was present in previous datasets of mainly U2AF1-mutant cells and one study using a conditional Srsf2 P95H mutant mouse (Fei et al., 2018, Przychodzen et al., 2013, Shirai et al., 2015, Kon et al., 2018), it was not prioritized for follow up in any previous studies. Notably, we found the GNAS-L isoform to be more prominent in U2AF1 than SRSF2- mutant cells. However, two key features of the present study focused our attention on GNAS: our isogenic conditions that empowered the identification of convergent targets of both U2AF1 and SRSF2 mutations; and the high-quality eCLIP-Seq data in relatively homogeneous iPSC-derived HSPC populations and faithful genomic context (diploid human genome with one SF-WT and one mutant allele), which provided a critical filter for selection of direct targets of the mutant SFs.

It has long been known that Gαs exists in two isoforms, and previous studies with in vitro purified proteins has shown that *GNAS-L* had suggested that *GNAS-L* may encode a more active protein relative to *GNAS-S* (Graziano et al. 1989). However, the biological significance of this isoform, if any, has remained elusive. In this work we show, for the first time, a cancer-related phenotype is directly mediated through expression of the *GNAS-L* isoform. Strikingly, we

also uncover a novel biochemical feature of this protein that point mutation in cooperation with usage of the long isoform results in hyper-active function. This result is corroborated through biochemical data and analysis of mutational landscape from large patient cohorts. Interestingly, activation of Gαs through GNAS mutations is typically found in low-grade solid tumors and benign metaplasias. SF mutations – which we show here cause Gαs activation through preferential *GNAS-L* usage – are typical of MDS, a myeloid malignancy with a low proliferation index and of clonal hematopoiesis, a premalignant condition.

Thus, the findings we report here provide a new understanding of the role of SF mutations in MDS and open avenues for targeted therapeutic interventions for MDS patients. Future work will is needed to better characterize the downstream signaling components driven by *GNAS-L* to reveal drugs that may already be clinically approved for other purposes that may show efficacy in MDS patients.

4.3 Concluding remarks

This work has revealed, for the first time, a disease-relevant consequence of differential usage of GNAS isoforms. While this work has laid the foundation to suggest G protein signaling is dysregulated through this alternate isoform in MDS, there is much work to be done to convert this finding into a therapeutic option for patients. There are two major pieces of missing information that would make this target a druggable option. First, we need to identify the G-protein coupled receptor (GPCR) that specifically couples to the long isoform (if there is one). And secondly, we need to determine the downstream signaling pathways that are driven by the long isoform. Preliminary data suggests that it is not the canonical PKA signaling axis, but rather ERK signaling. If we are able to confirm a signaling pathway that is dependent on this long

isoform, it can be targeted from many different angles in an attempt to kill these cancer cells with splicing factor mutations.

4.4 Materials and Methods

CRISPR/Cas9 gene editing of iPSCs

We used the previously described normal iPSC line N-2.12-D-1-1 as the parental line (Kotini et al., 2015). We used CRISPR/Cas9-mediated homology directed repair (HDR) for all editing, including introduction of point mutations and the 3xFLAG epitope. Editing of the SRSF2 P95L mutation was performed as previously described (Chang et al., 2018). Multiple independent clones with the desired mutation were isolated after each gene editing step and, following genetic and preliminary phenotypic characterization to exclude potential outliers, one clone was selected for the subsequent step. To introduce the *U2AF1* S34F, we used a plasmid expressing gRNAs under the U6 promoter and Cas9 linked to mCitrine with a P2A driven by the CMV promoter (Chang et al., 2018) and donor DNA plasmids. For *U2AF1* S34F, two different gRNAs targeting the U2AF1 locus within the within exon 3 (cutting site between 1 bp and 10 bp from the C¿T mutation site were designed, assembled by a two-step overlapping PCR reaction downstream of the U6 promoter sequence and cloned in the gRNA/Cas9 plasmid. Two sets (one for each gRNA) of two donor DNA plasmids, one containing the S34F mutation and one the corresponding wild-type (WT) sequence, containing 5' and 3' homology arms were constructed. The donor plasmids also contained silent mutations to introduce a new restriction site sequence (SphI and to prevent further cleavage by Cas9. The entire 5'+3' homology sequence was amplified from N-2.12 genomic DNA and the c.101C¿T and/or silent mutations to introduce new

restriction enzyme recognition sites and to prevent cleavage by Cas9 were introduced by two-step overlapping PCR before subsequent cloning into the donor plasmid. To generate *U2AF1* S34F iPSCs, the N-2.12 iPSC line was cultured in hESC media containing 10 mM Y-27632 for at least one hour before nucleofection. The cells were dissociated into single cells with accutase and 1 million cells were used for nucleofection with 5 μg of gRNA/Cas9 plasmid and 5 μg of each donor plasmid (WT and G12D) using Nucleofector II (Lonza) and program B-16. Immediately after nucleofection the cells were replated on MEFs. mCitrine+ cells were FACS-sorted 48 hr after transfection and plated as single cells at clonal density (1000 FACS-sorted cells per 60-mm dish). After 10-12 days, single colonies were picked in separate wells of a 6-well plate, allowed to grow for approximately 3-6 days and screened by PCR. 1-3 medium-sized colonies from each individual clone were picked directly into a 0.2 ml tube, pelleted and lysed. Restriction Fragment Length Polymorphism (RFLP) analysis was performed after PCR with primers and digestion of the product with SphI. Bi-allelically targeted clones were selected and the PCR products were cloned into the PCR-4 TOPO TA vector (Invitrogen) and sequenced to select clones heterozygous the *U2AF1* S34F mutation. As genetically identical lines exhibit very little phenotypic variation, a representative line for *U2AF1* S34F and *SRSF2* P95L, respectively, was selected in which for a second round of CRISPR-Cas9 gene editing to introduce an epitope tag. Specifically, a 3xFLAG tag was inserted into the C-terminus of the endogenous locus of the mutant allele and in one allele of the wild-type isogenic normal line. Editing for the epitope tag was done as described above, with the exception that only one donor encoding the 3xFLAG immediately before the stop codon was used for HDR with no concurrent WT donor. Western blot was done to ensure all selected clones expressed FLAG-tagged protein of the appropriate size. To ensure clonality, an additional step of single-cell cloning was performed after each step

of gene editing.

Human iPSC culture, hematopoietic differentiation and in vitro phenotypic characterization

Culture of human iPSCs on mitotically inactivated MEFs or feeder-free conditions, was performed as previously described (Chang et al., 2018). Hematopoietic differentiation was performed using a spin-EB protocol previously described (Chang et al., 2018). At the end of the differentiation culture, the cells were collected and dissociated with accutase into single cells and used for flow cytometry, cytological analyses or clonogenic assays, as described (Chang et al., 2018). Competitive growth assays were performed using an isogenic GFP-marked iPSC line (N-2.12-GFP), as previously described (Chang et al., 2018).

Flow cytometry and FACS sorting

Following hematopoietic differentiation, the cells were dissociated with accutase into single cells. The following flow cytometry antibodies were utilized: CD34-PE (clone 563, BD Pharmingen) and CD45-APC (clone HI30, BD Pharmingen). Cell viability was assessed with DAPI (Life Technologies). Cells were assayed on a BD Fortessa and data were analyzed with FlowJo software (Tree Star). Cell sorting was performed on a BD FACS Aria II.

MACS sorting, mRNA isolation, and RNA-Seq library generation

iPSC-HSPCs were generated via hematopoietic differentiation from at least three clones per genotype. Magnetic cell sorting of CD45+ cells was performed using the MACS cell separation microbeads and reagents (Miltenyi Biotec). RNA was isolated from pellets ranging in size

from 500K to 1million CD34+/CD45+ cells using Direct-zol RNA purification kit (Zymo R2061). Sequencing libraries were prepared using the TruSeq Stranded mRNA library prep kit (Illumina 20020594) starting with 500ng of input RNA. Libraries were sequenced to a depth of 40 million reads in PE100 mode on an Illumina HiSeq4000.

eCLIP library preparation

eCLIP was performed as previously described (Van Nostrand et al., 2017). Briefly, CD34+/CD45+ cells were UV-crosslinked (400 mJ/cm^2, 254 nm) and snap-frozen. Crosslinked cell pellets were combined among multiple differentiations to yield a total of 10 million cells, per replicate, per genotype. Lysed pellets were sonicated and treated with RNaseI for RNA fragmentation. 2% of lysate was retained for preparation of a size-matched input library, and the remaining 98% was subject to immunoprecipitation (IP) using and anti-FLAG antibody (Sigma F1804), coupled to magnetic dynabeads (Invitrogen 11203D). Bound RNA fragments were dephosphorylated and 3'-end ligated with an RNA adapter. Protein-RNA complexes from both input and IP samples were run on SDS polyacrylamide gel and transferred to nitrocellulose membrane for extraction of bound RNA fragments. Membrane regions from the size of the protein to 75 kDa above the protein size were cut and RNA was released with proteinase K. Input samples were then dephosphorylated and 3'-end ligated with an RNA adapter. Reverse transcription was performed with AffinityScript (Agilent) and cDNAs were 5'-end ligated with a DNA adaptor. cDNA products were amplified with Q5 PCR mix (NEB) to yield a sequencing library. Libraries were sequenced on the Illumina HiSeq4000 in SE75 mode to a depth of 20 million reads per library.

Biotin-based visualization of RBP-coupled RNA

Visualization of bound RNA fragments was performed as described previously (Van Nostrand et al., 2020) Briefly, 10 million iPSCs were collected in accutase and crosslinked with standard eCLIP conditions. Cell pellets were prepared as in eCLIP up until the first RNA adapter ligation. A modified RNA linker ligation was performed with standard eCLIP ligation conditions with 500 pmol pCp-Botin (Kena Bioscience) in place of the RNA adapter. Samples were incubated at 16 DegC overnight. For IP samples, samples were washed once with high-salt wash buffer and three times with wash buffer. Input samples were loaded directly on SDS-PAGE gel for electrophoresis and transfer to nitrocellulose membrane. Visualization was performed using the Chemiluminescent Nucleic Acid Detection Module Kit (Thermo Fisher) according to manufacturer's instructions.

RNA-seq data processing

RNA-sequencing reads were trimmed using cutadapt (v1.4.0) of adaptor sequences and mapped to repetitive elements (RepBase v18.04) using the STAR (v2.4.0i). Reads that did not map to repetitive elements were then mapped to the human genome (hg19). GENCODE (v19) gene annotations and featureCounts (v.1.5.0) were used to create read count matrices.

Quantification of Alternative Splicing

rMATS v4.0.2 was used to perform alternative splicing analysis among replicate RNA-seq datasets. Significant events were calculated as those with ¿5% change in isoform ratio between genotypes with a FDR ¡ 0.05. Each event was required to contain an average of at least 10 reads supporting the inclusion and exclusion isoform across replicates in one genotype,

while also containing at least an average of 10 reads total between both isoforms in the other genotype. We previously reported that rMATS often calls multiple splicing events with different flanking regions but overlapping alternatively splicing regions (PMID 30413564). To remove these artifacts and avoid double-counting of splicing events, we used custom scripts to only keep non-overlapping alternative splicing events for analysis (subsetrmatsjunctioncountonly.py found in https://github.com/YeoLab/rbp-maps).

Selection of unchanged cassette exons

To generate a background list of cassette exons that are unchanged between genotypes, we randomly selected exons that were matched for inclusion levels. The splicing events of interest were grouped in ranges of inclusion levels from: 0-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1. We then calculated the total number of events that fell within each range and randomly selected that same number of exons from a list that were not differentially regulated between genotypes and fell within the same range of inclusion level.

eCLIP-seq data processing

Reads were processed as described previously (Van Nostrand et al., 2016). Briefly, reads were adapter-trimmed and mapped to human-specific repetitive elements from RepBase (version 18.05) by STAR (Dobin et al., 2013). Repeat-mapping reads were removed, and remaining reads mapped to the human genome assembly hg19 with STAR. PCR duplicate reads were removed using the unique molecular identifier (UMI) sequences in the 5' adapter to generate 'usable reads' used in peak calling. Peaks were called on the usable reads by CLIPper (Lovci et al., 2013) and assigned to gene regions annotated in GENCODE (v19) with the follow-

ing descending priority order: 3' splice site, 5' splice site, CDS, 3'UTR, 5'UTR, proximal intron, and distal intron, noncoding sequences. Proximal intron regions are defined as extending up to 500 bp from an exon-intron junction. Each peak was normalized to the size-matched input (SMInput) by calculating the fraction of the number of usable reads from immunoprecipitation to that of the usable reads from the SMInput. Peaks were deemed significant at $>$4-fold enrichment and p$<10^{-5}$ (χ^2 test, or Fisher's exact test if the observed or expected read number in eCLIP or SMInput was below 5). Peaks passing significance thresholds in either replicate were kept for downstream analyses.

Kmer enrichment analysis of eCLIP peaks

Kmer sequences were counted in regions of interest using kvector (https://github.com/olgabot/kvector). The resulting counts were summed by genotype and enrichment was calculated using a chi-squared test (Scipy v1.2.0, chi2contingency) comparing the kmer-frequencies observed in each genotype. FDR was calculated to correct for multiple hypothesis testing and 6-mer sequences with FDR ¡ 0.05 were reported as significantly enriched.

Publicly available RNA-seq data

Unprocessed fastq files from CD34+ cells reported in Pellagatti et al. Blood 2018 were downloaded from GEO (accession GSE114922). Files were processed as described above.

Knockdown and overexpression

For knockdown, GNAS-L shRNA and scramble shRNA sequences were inserted into the 3' UTR of the G-U6 lentiviral vector (Kotini et al., 2015). Vectors were packaged as described (Kotini et al., 2015). 500,000 cells iPSC-HSPCs separated on day 11 of hematopoietic differentiation were transduced with either GNAS-L shRNA or scramble shRNA. Two days after transduction, the cells were plated on methylcellulose media for colony forming assay and RT-qPCR. For overexpression, GNAS Short wild-type (S-WT), Long wild-type (L-WT), Short R201H (S-R201H), Long R201H (L-R201H), SRSF2 P95L and empty GFP vector sequences were inserted into the mP2A lentiviral vector (Chang et al., 2018). 500,000 cells of the appropriate cell type (iPSC-HSPCs, K562, or CB CD34+) were transduced. iPSC-HSPCs were transduced on d11 of hematopoietic differentiation and taken for subsequent colony formation assay and/or western blots. Two days after transduction the cells were harvested and assessed by flow cytometry. K562 and CB CD34+ cells were subsequently collected for western blot and/or cAMP ELISA assay.

Western Blot

All cells transduced with a lentiviral vector encoding the previously specified GNAS OE constructs, SRSF2 P95L or empty vector were lysed two days after transduction with high salt buffer (0.3 M KCl) supplemented with protease inhibitor and phosphatase inhibitor. Additionally, cells taken upon hematopoietic differentiation were collected at d14 of differentiation. Cytokine starved cells were collected at d13, washed with 1X PBS and resuspended in the previously specified formula for Stem Pro34 SFM medium (cite above or Ng et al., 2008) without cytokines and cultured for 24hrs. 500,000 to 1 million cells were collected per sample for all western blots.

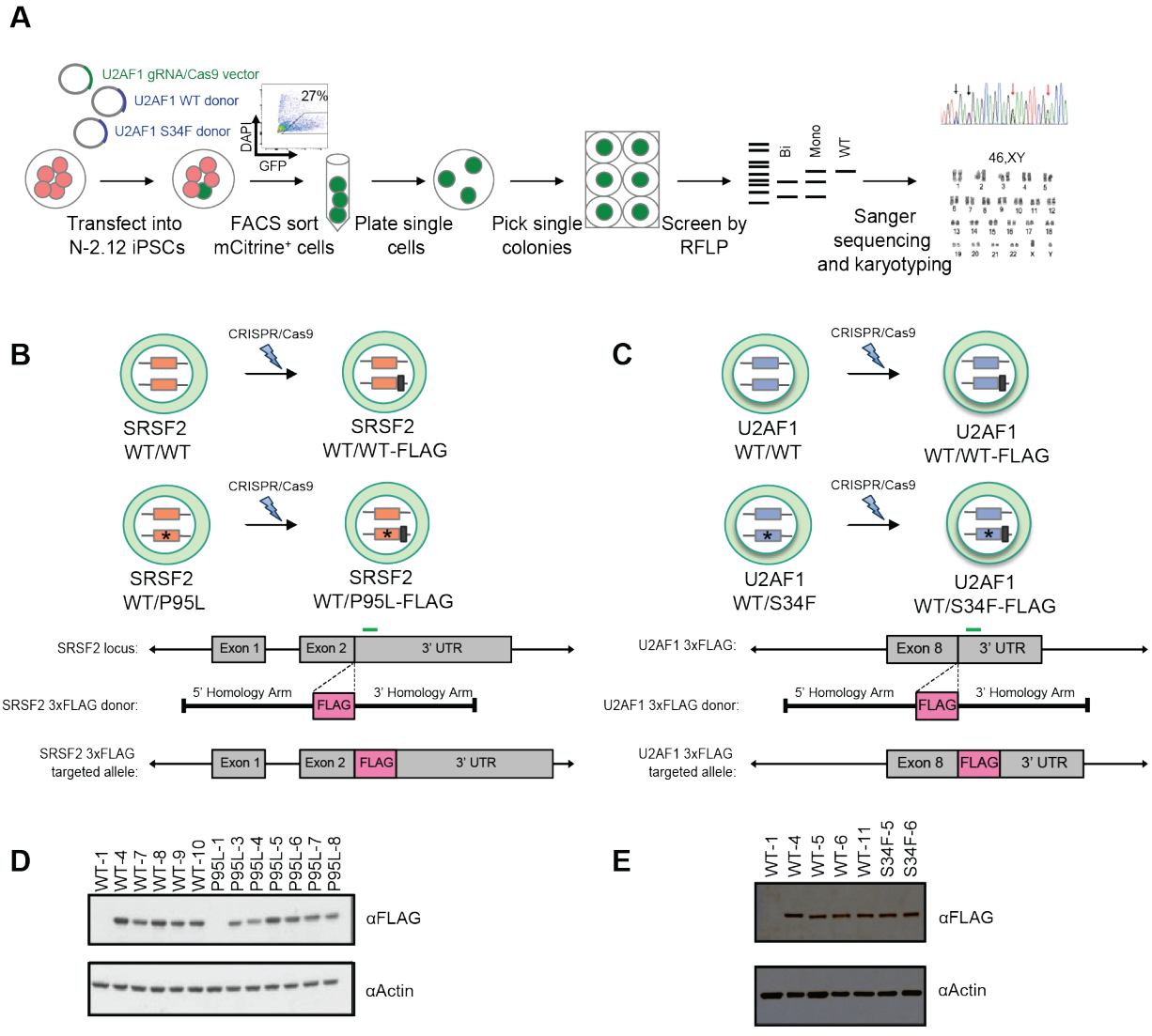
Protein concentrations were determined by bicinchoninic acid assay (Pierce Biotechnology Inc.) and 20 μg of protein from each extract were diluted in Laemmli SDS sample buffer and resolved by electrophoresis on Bolt 4% to 12% Bis-Tris precast gels (Invitrogen) and blotted on nitrocellulose membrane. The membranes were blocked with 5% BSA in Tris-buffered saline and incubated with primary antibody p-(Ser/Thr) PKA substrate or β-Actin (Cell Signaling). After washing, blots were incubated with HRP-conjugated secondary antibody and developed using ECL Western Blotting Detection Reagents. Band intensity was quantified by ImageJ.

4.5 Acknowledgements

The authors would like to thank Eric VanNostrand for critical discussion of analysis methods used in this manuscript. The authors would like to thank Dana Steffen and Silvio Gutkind for helpful discussion of cell signaling through PKA. This work was supported by NIH grant HL137219 to E.P and G.Y. E.C.W is supported by the National Science Foundation Graduate Research Fellowship. E.C.W is supported in part by a Ruth L. Kirschstein Institutional National Research Award from the National Institute for General Medical Sciences, T32 GM008666.

Chapter 4 in part is a reprint of material submitted to be published in: **Wheeler, EC.***, Vora S.*, Kotini A, Mayer D, Olszewaska M, Park SS, Sunahara RK, Yeo GW*, Papapetrou EP*. *SRSF2* and *U2AF1* mutations drive MDS through alternate *GNAS* isoform usage and cooperate with *GNAS* hotspot mutations. *in preparation*. The dissertation author was the primary author.

Figure 4.1: Generation of isogenic CRISPR edited stem cell lines. (A) Scheme of the experimental strategy for generating CRISPR/Cas9-mediated gene edited clones. iPSCs were co-transfected with a plasmid expressing Cas9 together with mCitrine and the gRNA and two donor DNA plasmids. Two donor plasmids (one carrying the S34F mutation and one the WT sequence) were co-delivered to facilitate selection of heterozygous mutant clones following selection of biallelically edited clones. mCitrine+ cells were sorted 48 hours after nucleofection to enrich for transfected and gene-edited cells and plated at clonal density. Individual clones were replated and initially screened by restriction fragment length polymorphism (RFLP) analysis and selected clones were further characterized by Sanger sequencing and karyotyping. (B) Upper: Schematic of strategy for introducing a 3xFLAG tag at the C-terminus of the mutant allele in the SRSF2 P95L iPSC line P95L-1 and of one allele of the WT line WT-1. The asterisk indicates the P95L mutation. Lower: Schematic of the SRSF2 locus with the position of the gRNA target sequence (green line) shown and of the donor template, harboring the 3xFLAG tag sequence, with the positions of the 5' and 3' homology arms shown. (C) Upper: Schematic of strategy for introducing a 3xFLAG tag at the C-terminus of the mutant allele in the U2AF1 S43F iPSC line S34F-2 and of one allele of the WT line WT-1. The asterisk indicates the S34F mutation. Lower: Schematic of the U2AF1 locus with the position of the gRNA target sequence (green line) shown and of the donor template, harboring the 3xFLAG tag sequence, with the positions of the 5' and 3' homology arms shown. (D, E) Detection of the FLAG tag by Western blot in the indicated edited iPSC lines.


A
U2AF1 gRNA/Cas9 vector
U2AF1 WT donor
U2AF1 S34F donor
27%
DAPI
GFP
Transfect into N-2.12 iPSCs
FACS sort mCitrine+ cells
Plate single cells
Pick single colonies
Bi
Mono
WT
Screen by RFLP
Sanger sequencing and karyotyping
46,XY
B
CRISPR/Cas9
SRSF2 WT/WT
SRSF2 WT/WT-FLAG
SRSF2 WT/P95L
SRSF2 WT/P95L-FLAG
SRSF2 locus:
Exon 1
Exon 2
3' UTR
SRSF2 3xFLAG donor:
5' Homology Arm
FLAG
3' Homology Arm
SRSF2 3xFLAG targeted allele:
C
U2AF1 WT/WT
U2AF1 WT/WT-FLAG
U2AF1 WT/S34F
U2AF1 WT/S34F-FLAG
U2AF1 3xFLAG:
Exon 8
U2AF1 3xFLAG donor:
U2AF1 3xFLAG targeted allele:
D
WT-1
WT-4
WT-7
WT-8
WT-9
WT-10
P95L-1
P95L-3
P95L-4
P95L-5
P95L-6
P95L-7
P95L-8
αFLAG
αActin
E
WT-1
WT-4
WT-5
WT-6
WT-11
S34F-5
S34F-6
αFLAG
αActin

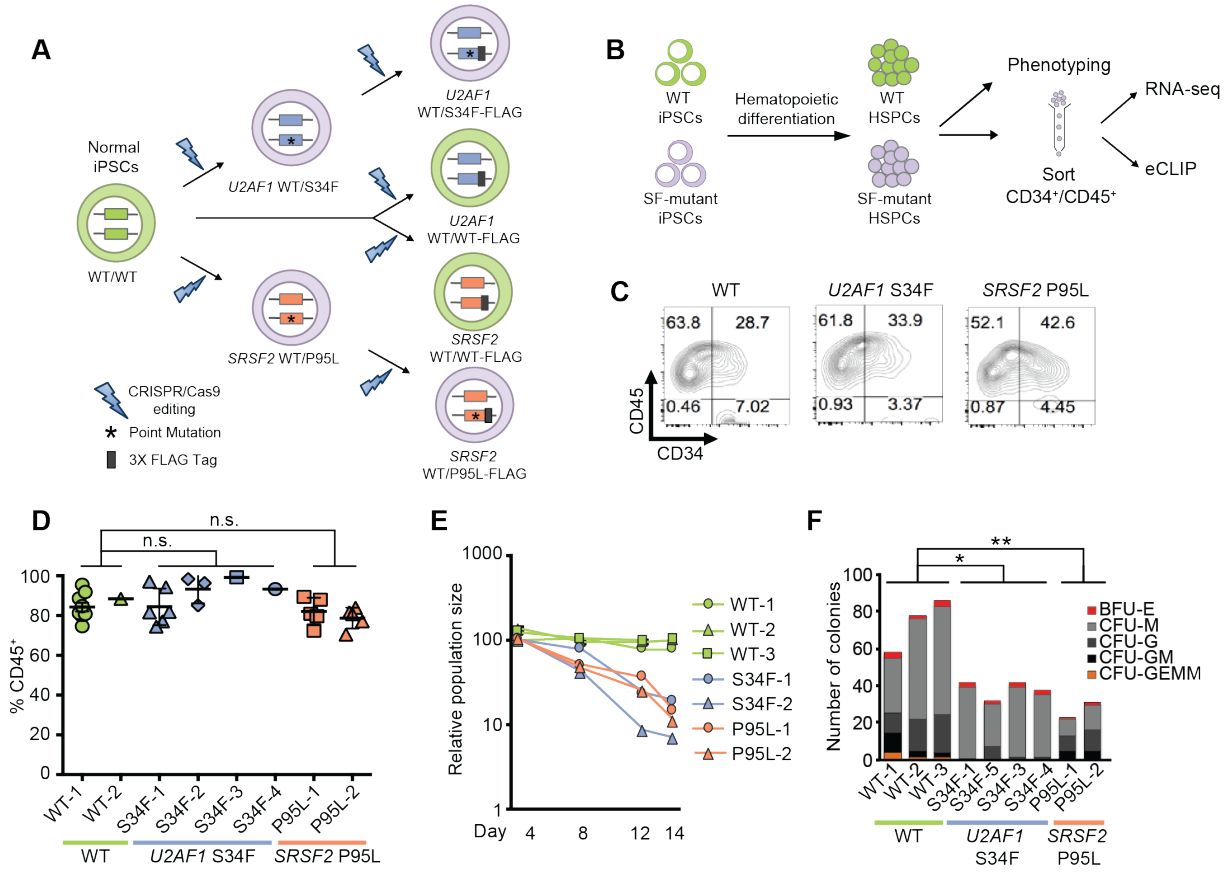


Figure 4.2: Hematopoietic specification and phenotyping of SF-mutant stem cell lines. (A) . Schematic overview of the generation of isogenic clonal iPSC lines with the canonical U2AF1 S34F and SRSF2 P95L mutations and allele-specific epitope tags through CRISPR Cas9 mediated gene editing. (B) Schematic overview of the experimental design of this study. Upon directed hematopoietic differentiation, iPSC derived cells were subjected to phenotypic characterization and CD34+ CD45+ HSPCs were magnetically sorted for RNA-seq and eCLIP analyses. (C) Representative flow cytometry plots on day 14 of hematopoietic differentiation. (D) Fraction of CD45+ cells on day 14 of hematopoietic differentiation of the indicated WT and SF mutant iPSC lines. Mean and SEM of values from 1 to 7 independent differentiation experiments for each line are shown. (E) Competitive growth assay. The cells were mixed 1:1 at the onset of hematopoietic differentiation with an isogenic WT iPSC line stably expressing GFP (derived from the parental line WT 1). The relative population size was estimated as the percentage of GFP negative cells (measured by flow cytometry) at each time point (days 2 to 12 of differentiation) relative to the population size on day 2. Results from 1 or 2 independent experiments per line are shown. (F) Number of colonies obtained from 5,000 cells of the indicated iPSC lines seeded in methylcellulose assays on day 14 of hematopoietic differentiation. Mean of 1-6 independent differentiation experiments for each line is shown.

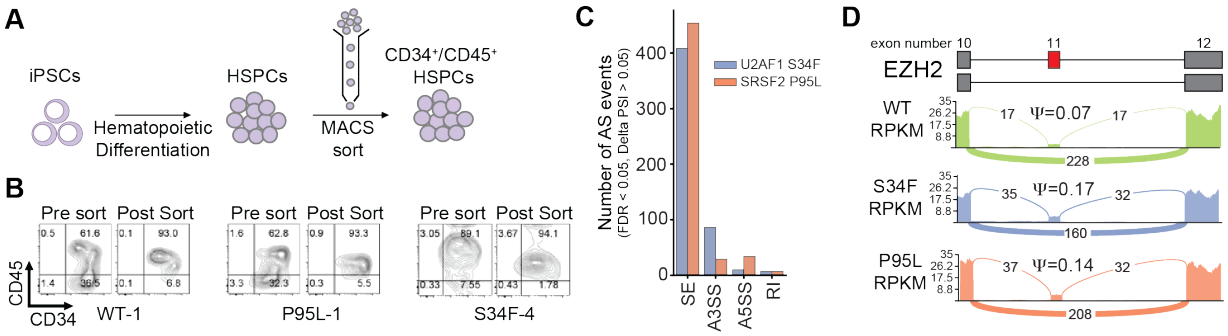


Figure 4.3: Quantification of alternative splicing in HSPCs. (A) Schematic of isolation of iPSC-derived CD34+/CD45+ HSPCs for RNA-Seq and eCLIP analyses using magnetic-activated cell sorting (MACS) with anti-CD45 beads on an empirically determined day of differentiation culture when nearly 100% of CD45+ cells are also still CD34+ (ranging from day 10 to 13, depending on the individual line and differentiation experiment). (B) Flow cytometric assessment of cell purity of one representative MACS-sorted iPSC-HSPC sample per genotype used for the RNA-seq and eCLIP analyses. (C) Number of AS events detected for each genotype in comparison to WT cells (FDR ¡ 0.05, delta PSI ¿ 5%). SE: skipped exon; A3SS: alternative 3' splice site; A5SS: alternative 5' splice site; RI: retained intron. (D) Sashimi plot showing percent spliced in (Ψ) of the *EZH2* poison exon from one representative sample of each genotype.

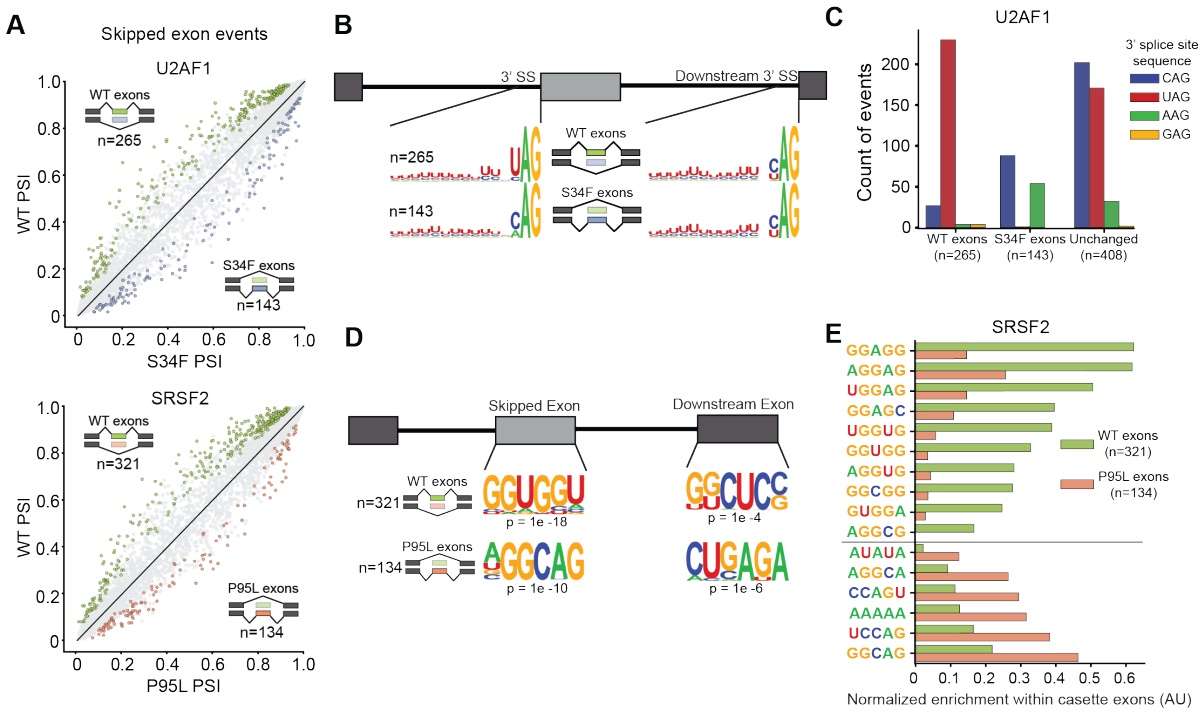


Figure 4.4: Sequence-specific alterations of splicing factor mutant cells. (A) Scatterplots of percent spliced in (PSI) values of all skipped exon events for each genotype. Significant events (FDR $<$ 0.05, Delta PSI $>$ 5%) are colored. (B) Sequence logos of 3' splice sites flanking skipped exon events and downstream exons. WT exons and S34F exons denote exons skipped or included, respectively, in U2AF1-mutant cells. (C) Count of skipped exon events containing the indicated 3' splice site sequence. WT exons and S34F exons denote exons skipped or included, respectively, in U2AF1-mutant cells. Unchanged exons are a randomly selected control set of exons that are not differentially spliced in U2AF1-mutant cells. (D) Motif enrichment of sequence preferences in skipped exons and downstream control exons. WT exons and P95L exons denote exons skipped or included, respectively, in SRSF2-mutant cells. (E) 5-mer sequences with most significant differential enrichment in WT exons and P95L exons. $p < 0.005$ by chi-squared test with Bonferroni-correction for all 5-mers plotted.

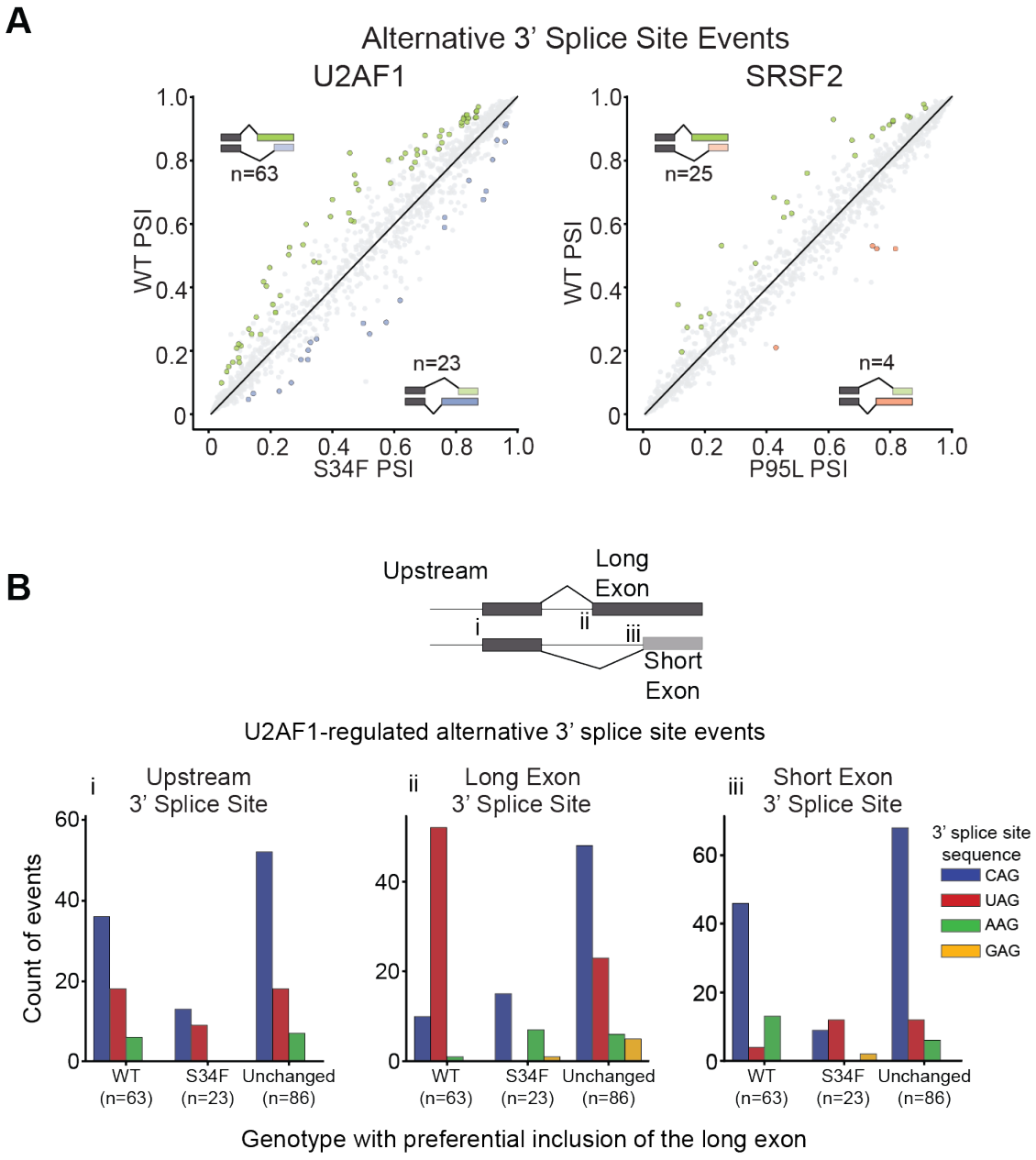


Figure 4.5: Alternative 3' splice site alterations in U2AF1-mutant cells. (A) Scatterplots of percent spliced in (PSI) values of all alternative 3' splicing events for each genotype. Significant events (FDR < 0.05, Delta PSI $> 5\%$) are colored. (B) Count of alternative 3' splicing events containing the indicated 3' splice site sequence. WT exons and S34F exons denote exons skipped or included, respectively, in U2AF1-mutant cells. Unchanged exons are a randomly selected control set of exons that are not differentially spliced in U2AF1-mutant cells.

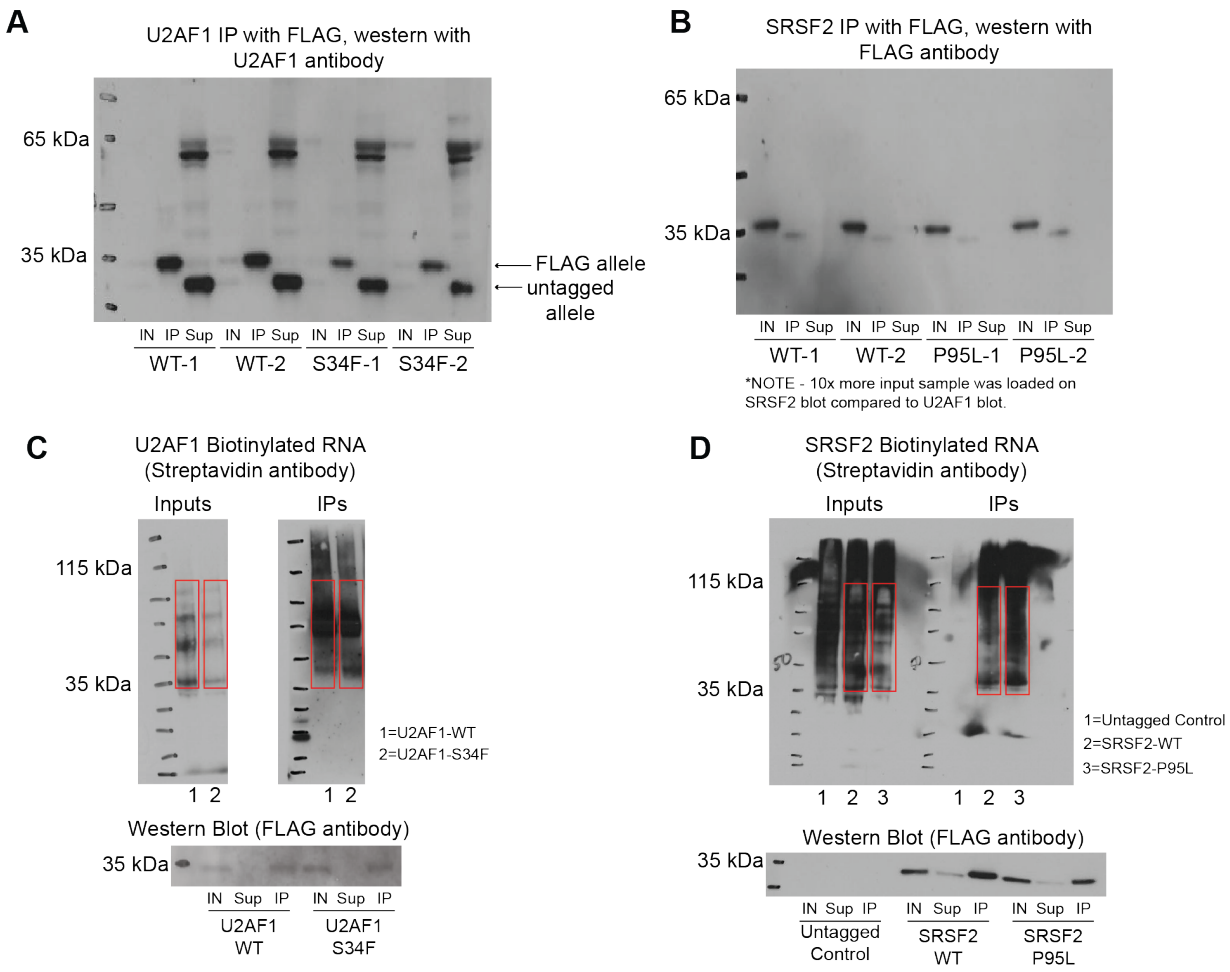


Figure 4.6: eCLIP western blot and RNA visualization. (A,B) Western blot of input (IN), immunoprecipitated (IP), and supernatant (Sup) samples in biological duplicate of each genotype from U2AF1 and SRSF2 eCLIP experiments. IP was performed with M2 FLAG antibody, and western was probed with endogenous U2AF1 antibody (A), or M2 FLAG antibody (B). (C,D) Visualization of RNA in input and IP samples from biotin ligation of RNA fragments and detection with streptavidin antibody following membrane transfer. Red boxes indicate regions from the membrane in which RNA was isolated and prepared for eCLIP sequencing libraries. Western blots below serve as a control for total protein amount and quality of immunoprecipitation from each experiment. Images shown are a representative from one of two (U2AF1) or three (SRSF2) replicate experiments.

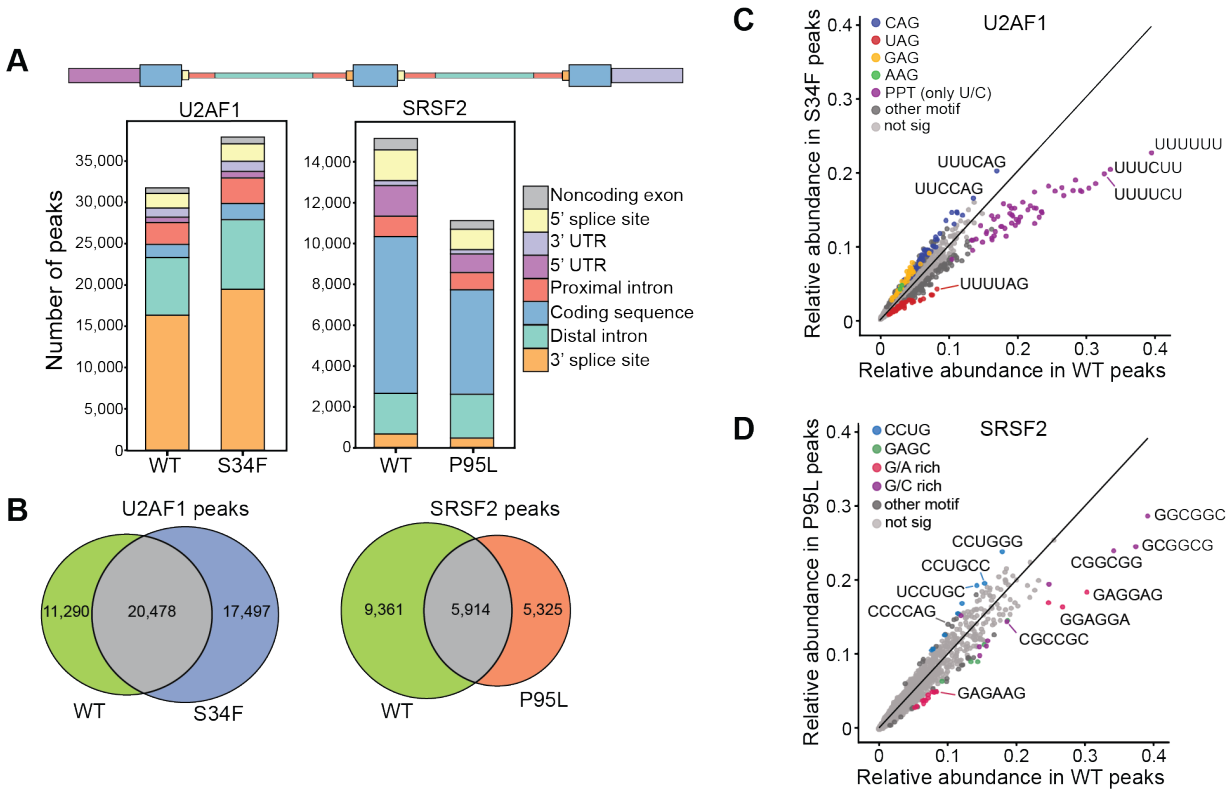


Figure 4.7: Summary of eCLIP binding and sequence-specific binding. (A) Count of significantly enriched binding sites (peaks) identified in each genic region indicated. Significantly enriched peaks are peaks with fold change $>$ 4 relative to input and p-value $<$ 0.001 (chi-squared test) in at least one of two replicate experiments. (B) Venn diagram of all peaks for each genotype showing overlap between WT and mutant factor binding sites. (C,D) Relative abundances of 6-mer sequences counted in all peak regions. Differentially used 6-mers were identified with a chi-squared test and FDR $<$ 0.05. 6-mers that are not differentially used (not sig) are plotted in light grey.

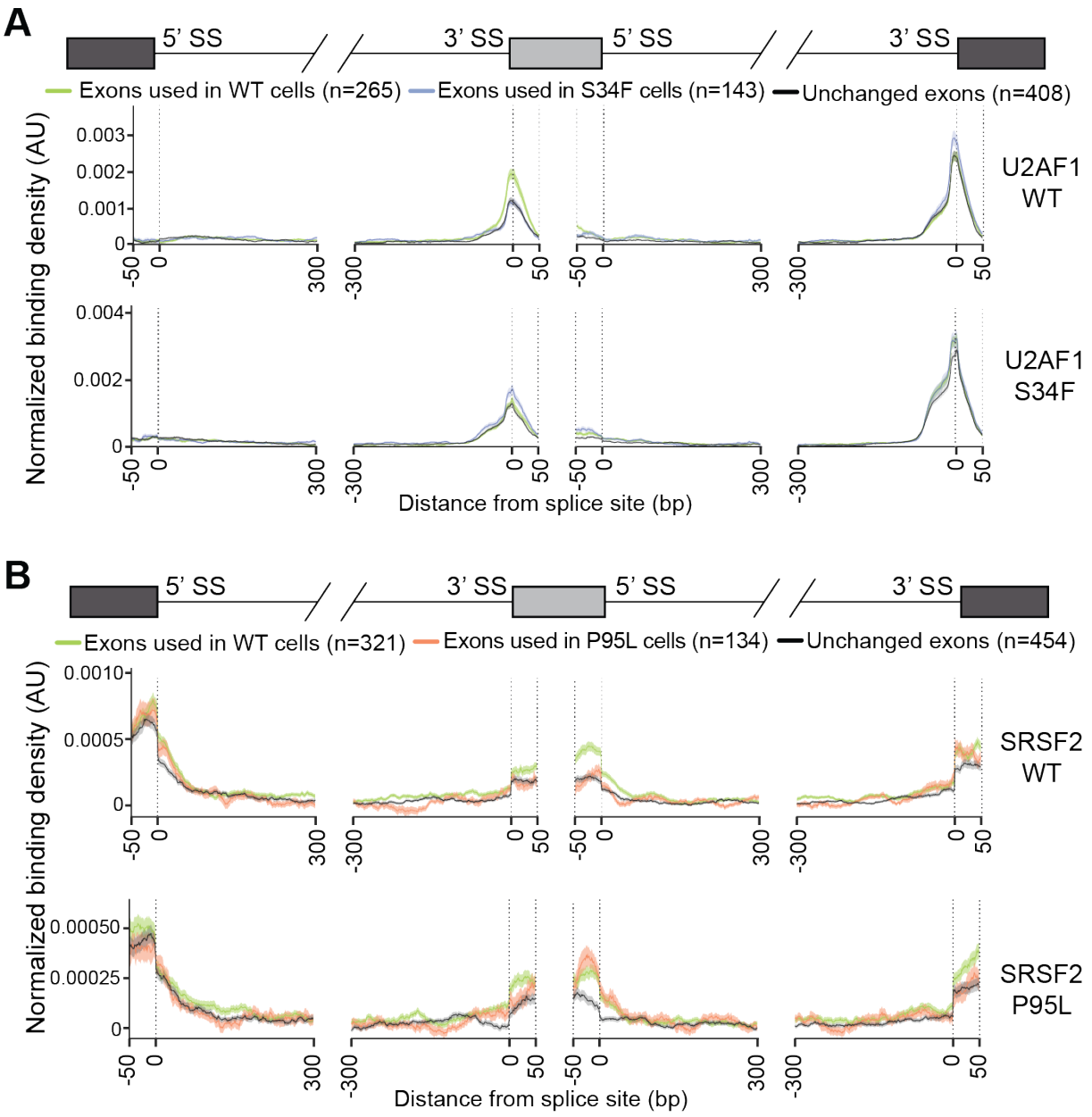


Figure 4.8: Splicing maps of binding in regulatory regions around alternative splicing events. (A,B) Splicing maps of eCLIP binding density in regions that are alternatively spliced as identified with RNA-Seq. A random background set of exons that are unchanged between the two conditions is used as control. Binding density is averaged across regions (50bp on each end of exon sequences and 300bp into surrounding introns) and normalized to the input sample. (AU = arbitrary units).

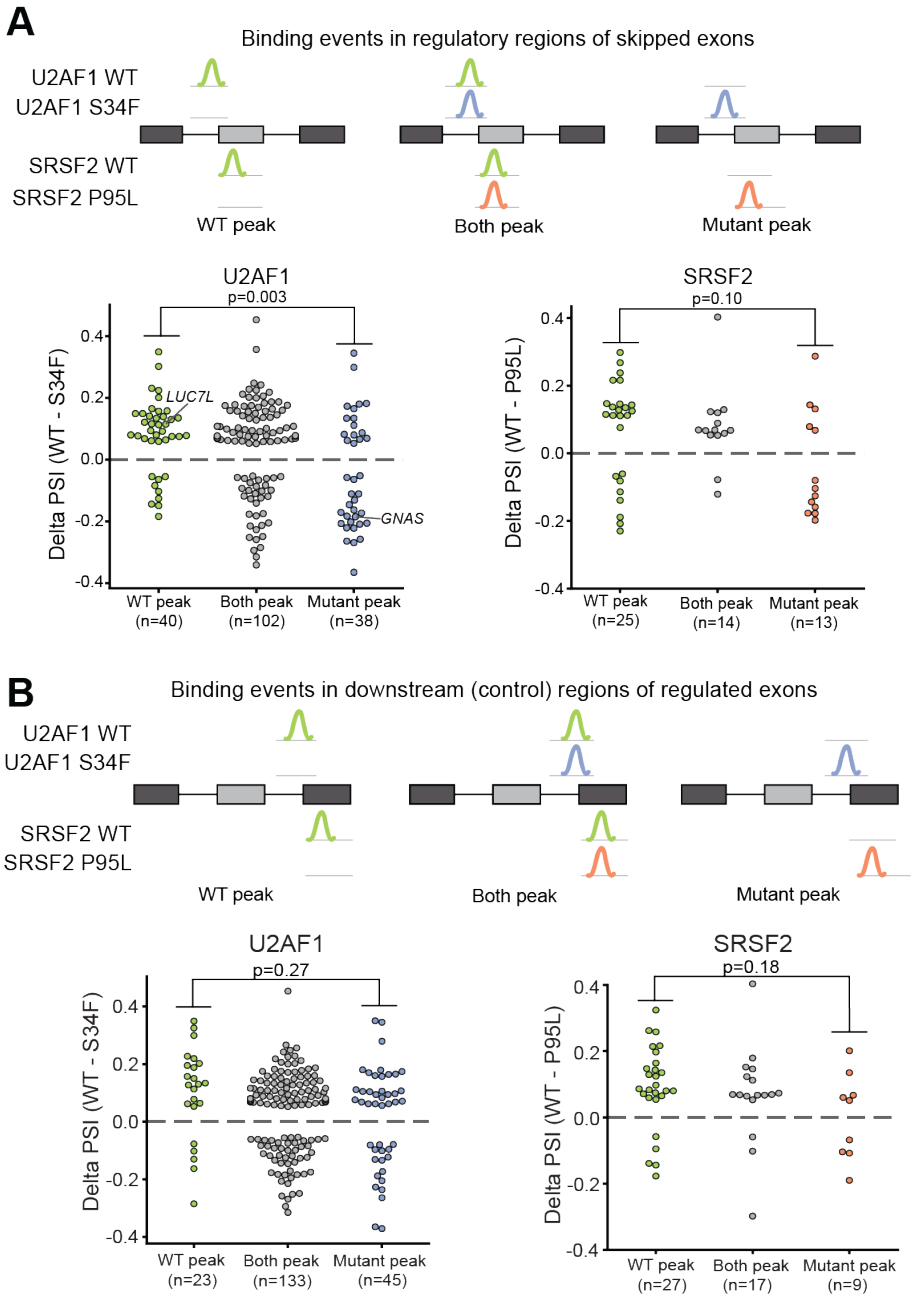
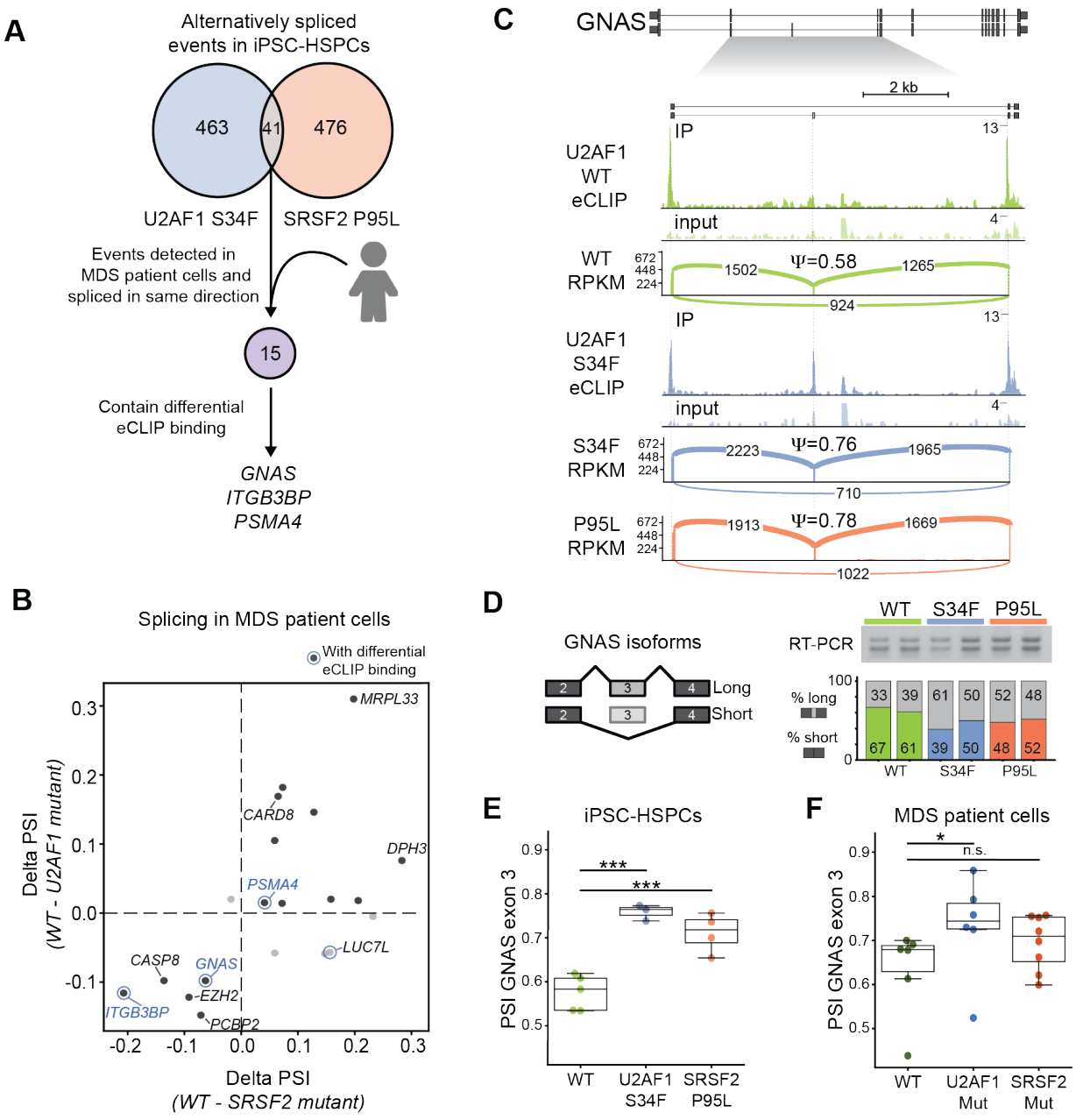


Figure 4.9: Differential binding in regulatory regions of alternative splicing events. (A,B) Upper: schematic of strategy for identification of skipped exon events with differential peaks in regulatory regions (upstream intron for U2AF1 and cassette exon for SRSF2). Skipped exons are categorized in 3 groups according to presence of WT only SF peak, mutant only SF peak, or both. Lower: Delta PSI (WT–mutant) of skipped exon events with binding events plotted for each of the 3 groups. A chi-square test was used to determine dependence of WT only or Mutant only peak status on the direction.

Figure 4.10: Integrative analyses nominate alternate usage of the *GNAS-L* as a common effector of SF mutations. (A) Workflow showing the integration of AS analyses in both genotypes with datasets from U2AF1- and SRSF2- mutant MDS patient cells and with eCLIP analyses of differentially bound exons. (B) Delta PSI of splicing events common to both genotypes in iPSC-HSPCs that were detected in MDS patient cells. Black dots represent events that trend in the same direction in both genotypes, blue circle shows events that also contain a differential eCLIP peak in the regulatory region. (C) eCLIP binding track plotting normalized read density of IP and input sample at exons 2 - 5 of the GNAS transcript for one replicate of U2AF1 WT and U2AF1 S34F. Sashimi plots display inclusion levels of exon3 in WT iPSC-HSPCs (green), S34F (blue), or P95L (orange) - mutant iPSC-HSPCs for 1 representative sample (n=3-6). Counts shown in sashimi plots curves represent the number of reads supporting exon inclusion (top), or exon skipping (bottom). Ψ = Percent spliced in of exon 3. (D) Quantitation of GNAS isoform expression by RT-PCR in WT and SF-mutant iPSC-HSPCs, as indicated. (E,F) Quantification of PSI (percent spliced in) of exon 3 in all replicates of iPSC-HSPCs (E) or MDS-patient cells (F). Boxes represent the IQR (25th, 50th and 75th percentiles), while whiskers represent 1.5 times the IQR from the 25th and 75th percentiles. Significance is determined with a wilcoxon rank-sum test. ***$p<0.005$, *$p<0.05$, ns = not significant.


A
Alternatively spliced events in iPSC-HSPCs
463
41
476
U2AF1 S34F
SRSF2 P95L
Events detected in MDS patient cells and spliced in same direction
15
Contain differential eCLIP binding
GNAS
ITGB3BP
PSMA4
B
Splicing in MDS patient cells
With differential eCLIP binding
Delta PSI (WT - U2AF1 mutant)
Delta PSI (WT - SRSF2 mutant)
MRPL33
CARD8
DPH3
PSMA4
LUC7L
CASP8
GNAS
EZH2
ITGB3BP
PCBP2
C
GNAS
2 kb
U2AF1 WT eCLIP
IP
input
WT RPKM
Ψ=0.58
1502
1265
924
U2AF1 S34F eCLIP
S34F RPKM
Ψ=0.76
2223
1965
710
P95L RPKM
Ψ=0.78
1913
1669
1022
D
GNAS isoforms
Long
Short
RT-PCR
WT
S34F
P95L
% long
% short
E
iPSC-HSPCs
PSI GNAS exon 3
WT
U2AF1 S34F
SRSF2 P95L
F
MDS patient cells
n.s.
WT
U2AF1 Mut
SRSF2 Mut

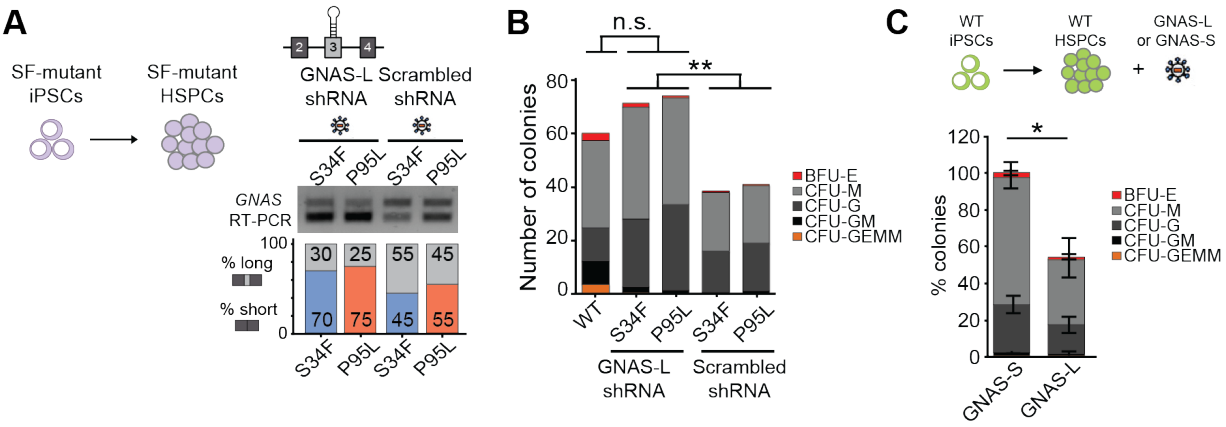


Figure 4.11: Expression of *GNAS-L* modulates disease phenotypes in HSPCs. (A) SF-mutant iPSC-HSPCs on d12 of differentiation were transduced with a lentiviral vector encoding an shRNA targeting GNAS-L or scrambled shRNA and GNAS isoform expression was quantitated by RT-PCR 48 hours later. One representative experiment out of 3 is shown. (B) Number of methylcellulose colonies from 5,000 SF-mutant iPSC-HSPCs transduced with GNAS-L shRNA or scrambled shRNA. Values of 1-2 representative experiments are shown. (C) Methylcellulose colonies from WT iPSC-HSPCs transduced with GNAS-L or GNAS-S. Mean and SEM of 3 experiments with 3 different WT lines (WT-1, WT-2, WT-3) is shown.

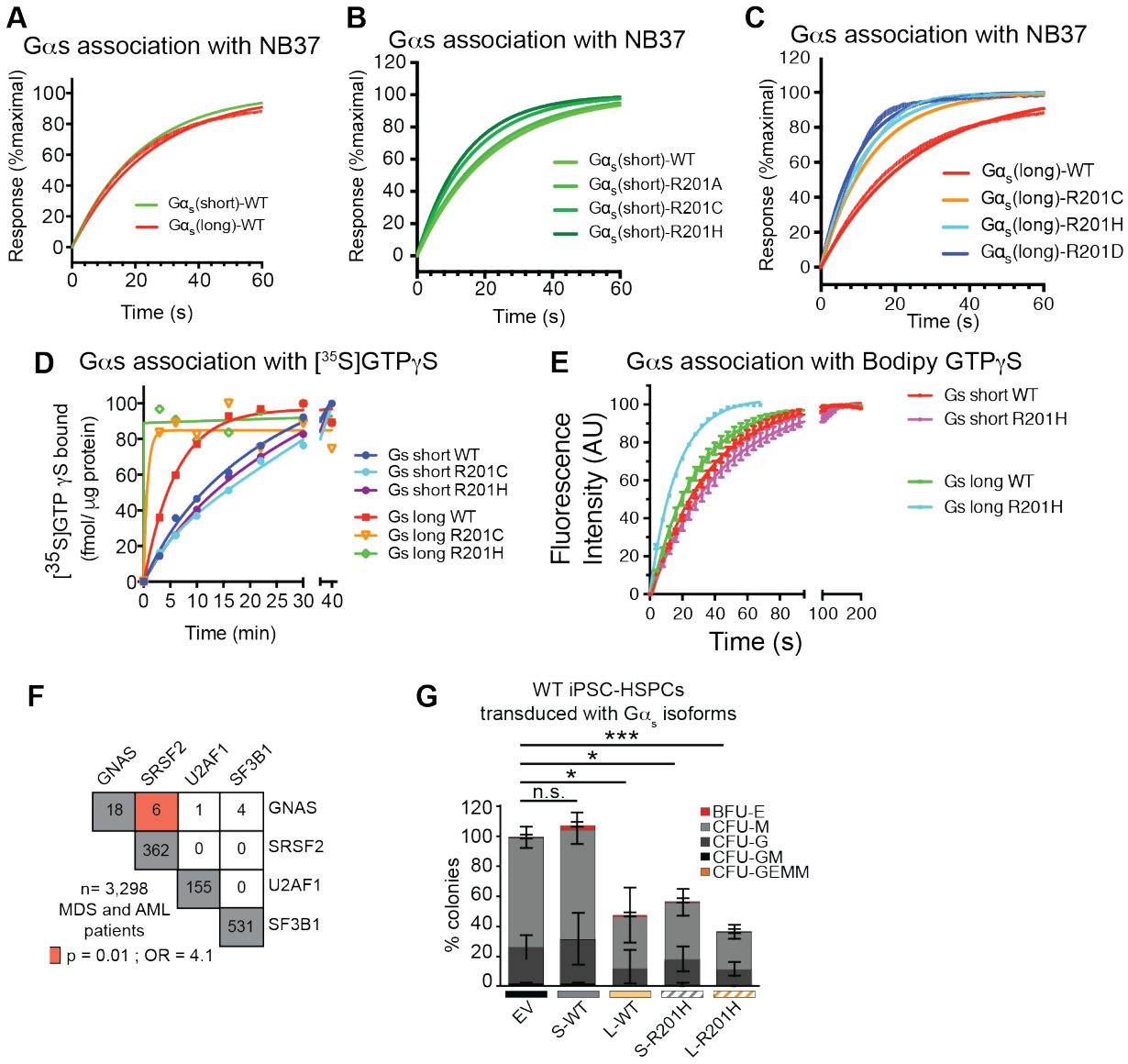


Figure 4.12: *GNAS-L* in cooperation with R201 point mutation confers a hyperactive phenotype. (A-C) Interferometry data showing NB37 association with purified forms of Gαs in the short or long isoform, with and without point mutations at the R201 residue. (D) Affinity of Gαs for GTP and Bodiby GTP (E). (F) Co-occurrence of mutations in GNAS, SRSF2, U2AF1, and SF3B1 in MDS and AML patient cohorts (p-value calculated with a Fisher's exact test) (G) Methylcellulose colonies from WT iPSC-HSPCs transduced with the indicated vectors. Mean and SEM of 3 experiments with 3 WT iPSC lines (WT-1, WT-4, and WT-5) are shown.

Bibliography

1 Adamson, B., Norman, T. M., Jost, M., Cho, M. Y., Nunez, J. K., Chen, Y., Villalta, J. E., Gilbert, L. A., Horlbeck, M. A., Hein, M. Y., Pak, R. A., Gray, A. N., Gross, C. A., Dixit, A., Parnas, O., Regev, A. & Weissman, J. S. A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell* **167**, 1867-1882 e1821, doi:10.1016/j.cell.2016.11.048 (2016).

2 Anderson, P., Kedersha, N. & Ivanov, P. Stress granules, P-bodies and cancer. *Biochim Biophys Acta* **1849**, 861-870, doi:10.1016/j.bbagrm.2014.11.009 (2015).

3 Bauer, D. C., McMorran, B. J., Foote, S. J. & Burgio, G. Genome-wide analysis of chemically induced mutations in mouse in phenotype-driven screens. *BMC Genomics* **16**, 866, doi:10.1186/s12864-015-2073-4 (2015).

4 Bejar, R. & Steensma, D. P. Recent developments in myelodysplastic syndromes. *Blood* **124**, 2793-2803, doi:10.1182/blood-2014-04-522136 (2014).

5 Bejar, R., Stevenson, K., Abdel-Wahab, O., Galili, N., Nilsson, B., Garcia-Manero, G., Kantarjian, H., Raza, A., Levine, R. L., Neuberg, D. & Ebert, B. L. Clinical effect of point mutations in myelodysplastic syndromes. *N Engl J Med* **364**, 2496-2506, doi:10.1056/NEJMoa1013343 (2011).

6 Bjorling, E. & Uhlen, M. Antibodypedia, a portal for sharing antibody and antigen validation data. *Mol Cell Proteomics* **7**, 2028-2037, doi:10.1074/mcp.M800264-MCP200 (2008).

7 Blomen, V. A., Majek, P., Jae, L. T., Bigenzahn, J. W., Nieuwenhuis, J., Staring, J., Sacco, R., van Diemen, F. R., Olk, N., Stukalov, A., Marceau, C., Janssen, H., Carette, J. E., Bennett, K. L., Colinge, J., Superti-Furga, G. & Brummelkamp, T. R. Gene essentiality and synthetic lethality in haploid human cells. *Science* **350**, 1092-1096, doi:10.1126/science.aac7557 (2015).

8 Bos, T. J., Nussbacher, J. K., Aigner, S. & Yeo, G. W. Tethered Function Assays as Tools to Elucidate the Molecular Roles of RNA-Binding Proteins. *Adv Exp Med Biol* **907**, 61-88, doi:10.1007/978-3-319-29073-7_3 (2016).

9 Burger, K., Muhl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dolken, L. & Eick, D. 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biol* **10**, 1623-1630, doi:10.4161/rna.26214 (2013).

10 Caicedo, J. C., Singh, S. & Carpenter, A. E. Applications in image-based profiling of perturbations. *Curr Opin Biotechnol* **39**, 134-142, doi:10.1016/j.copbio.2016.04.003 (2016).

11 Castello, A., Fischer, B., Hentze, M. W. & Preiss, T. RNA-binding proteins in Mendelian disease. *Trends Genet* **29**, 318-327, doi:10.1016/j.tig.2013.01.004 (2013).

12 Chang, C. J., Kotini, A. G., Olszewska, M., Georgomanoli, M., Teruya-Feldstein, J., Sperber, H., Sanchez, R., DeVita, R., Martins, T. J., Abdel-Wahab, O., Bradley, R. K. & Papapetrou, E. P. Dissecting the Contributions of Cooperating Gene Mutations to Cancer Phenotypes and Drug Responses with Patient-Derived iPSCs. *Stem Cell Reports* **10**, 1610-1624, doi:10.1016/j.stemcr.2018.03.020 (2018).

13 Chung, J., Marini, S., Pera, J., Norrving, B., Jimenez-Conde, J., Roquer, J., Fernandez-Cadenas, I., Tirschwell, D. L., Selim, M., Brown, D. L., Silliman, S. L., Worrall, B. B., Meschia, J. F., Demel, S., Greenberg, S. M., Slowik, A., Lindgren, A., Schmidt, R., Traylor, M., Sargurupremraj, M., Tiedt, S., Malik, R., Debette, S., Dichgans, M., Langefeld, C. D., Woo, D., Rosand, J. & Anderson, C. D. Genome-wide association study of cerebral small vessel disease reveals established and novel loci. *Brain* **142**, 3176-3189, doi:10.1093/brain/awz233 (2019).

14 Cirillo, L., Cieren, A., Barbieri, S., Khong, A., Schwager, F., Parker, R. & Gotta, M. UBAP2L Forms Distinct Cores that Act in Nucleating Stress Granules Upstream of G3BP1. *Curr Biol* **30**, 698-707 e696, doi:10.1016/j.cub.2019.12.020 (2020).

15 D, G. H., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol* **17**, 28, doi:10.1186/s13059-016-0878-3 (2016).

16 Daniels, D. L., Mendez, J., Mosley, A. L., Ramisetty, S. R., Murphy, N., Benink, H., Wood, K. V., Urh, M. & Washburn, M. P. Examining the complexity of human RNA polymerase complexes using HaloTag technology coupled to label free quantitative proteomics. *J Proteome Res* **11**, 564-575, doi:10.1021/pr200459c (2012).

17 Darnell, J. C., Van Driesche, S. J., Zhang, C., Hung, K. Y., Mele, A., Fraser, C. E., Stone, E. F., Chen, C., Fak, J. J., Chi, S. W., Licatalosi, D. D., Richter, J. D. & Darnell, R. B. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell* **146**, 247-261, doi:10.1016/j.cell.2011.06.013 (2011).

18 Darnell, R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA* **1**, 266-286, doi:10.1002/wrna.31 (2010).

19 Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D. & Bock, C. Pooled CRISPR screening with single-cell transcriptome readout. *Nat Methods* **14**, 297-301, doi:10.1038/nmeth.4177 (2017).

20 de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F. & Strouboulis, J. Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. *Proc Natl Acad Sci U S A* **100**, 7480-7485, doi:10.1073/pnas.1332608100 (2003).

21 de Groot, R., Luthi, J., Lindsay, H., Holtackers, R. & Pelkmans, L. Large-scale image-based profiling of single-cell phenotypes in arrayed CRISPR-Cas9 gene perturbation screens. *Mol Syst Biol* **14**, e8064, doi:10.15252/msb.20178064 (2018).

22 DeJesus, R., Moretti, F., McAllister, G., Wang, Z., Bergman, P., Liu, S., Frias, E., Alford, J., Reece-Hoyes, J. S., Lindeman, A., Kelliher, J., Russ, C., Knehr, J., Carbone, W., Beibel, M., Roma, G., Ng, A., Tallarico, J. A., Porter, J. A., Xavier, R. J., Mickanin, C., Murphy, L. O., Hoffman, G. R. & Nyfeler, B. Functional CRISPR screening identifies the ufmylation pathway as a regulator of SQSTM1/p62. *Elife* **5**, doi:10.7554/eLife.17290 (2016).

23 DiSalvo, M., Smiddy, N. M. & Allbritton, N. L. Automated sensing and splitting of stem cell colonies on microraft arrays. *APL Bioeng* **3**, 036106, doi:10.1063/1.5113719 (2019).

24 Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., Marjanovic, N. D., Dionne, D., Burks, T., Raychowdhury, R., Adamson, B., Norman, T. M., Lander, E. S., Weissman, J. S., Friedman, N. & Regev, A. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853-1866 e1817, doi:10.1016/j.cell.2016.11.038 (2016).

25 Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

26 Dror, R. O., Mildorf, T. J., Hilger, D., Manglik, A., Borhani, D. W., Arlow, D. H., Philippsen, A., Villanueva, N., Yang, Z., Lerch, M. T., Hubbell, W. L., Kobilka, B. K., Sunahara, R. K. & Shaw, D. E. SIGNAL TRANSDUCTION. Structural basis for nucleotide exchange in heterotrimeric G proteins. *Science* **348**, 1361-1365, doi:10.1126/science.aaa5264 (2015).

27 Dvinge, H., Kim, E., Abdel-Wahab, O. & Bradley, R. K. RNA splicing factors as oncoproteins and tumour suppressors. *Nat Rev Cancer* **16**, 413-430, doi:10.1038/nrc.2016.51 (2016).

28 Engreitz, J. M., Sirokman, K., McDonel, P., Shishkin, A. A., Surka, C., Russell, P., Grossman, S. R., Chow, A. Y., Guttman, M. & Lander, E. S. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* **159**, 188-199, doi:10.1016/j.cell.2014.08.018 (2014).

29 Fei, D. L., Zhen, T., Durham, B., Ferrarone, J., Zhang, T., Garrett, L., Yoshimi, A., Abdel-Wahab, O., Bradley, R. K., Liu, P. & Varmus, H. Impaired hematopoiesis and leukemia development in mice with a conditional knock-in allele of a mutant splicing factor gene U2af1. *Proc Natl Acad Sci U S A* **115**, E10437-E10446, doi:10.1073/pnas.1812669115 (2018).

30 Feldman, D., Singh, A., Schmid-Burgk, J. L., Carlson, R. J., Mezger, A., Garrity, A. J., Zhang, F. & Blainey, P. C. Optical Pooled Screens in Human Cells. *Cell* **179**, 787-799 e717, doi:10.1016/j.cell.2019.09.016 (2019).

31 Flores-Figueroa, E., Gutierrez-Espindola, G., Guerrero-Rivera, S., Pizzuto-Chavez, J. & Mayani, H. Hematopoietic progenitor cells from patients with myelodysplastic syndromes: in vitro colony growth and long-term proliferation. *Leuk Res* **23**, 385-394, doi:10.1016/s0145-2126(98)00176-3 (1999).

32 Gach, P. C., Wang, Y., Phillips, C., Sims, C. E. & Allbritton, N. L. Isolation and manipulation of living adherent cells by micromolded magnetic rafts. *Biomicrofluidics* **5**, 32002-3200212, doi:10.1063/1.3608133 (2011).

33 Gerace, E. & Moazed, D. Affinity Purification of Protein Complexes Using TAP Tags. *Methods Enzymol* **559**, 37-52, doi:10.1016/bs.mie.2014.11.007 (2015).

34 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat Rev Genet* **15**, 829-845, doi:10.1038/nrg3813 (2014).

35 Grabocka, E. & Bar-Sagi, D. Mutant KRAS Enhances Tumor Cell Fitness by Upregulating Stress Granules. *Cell* **167**, 1803-1813 e1812, doi:10.1016/j.cell.2016.11.035 (2016).

36 Haferlach, T., Nagata, Y., Grossmann, V., Okuno, Y., Bacher, U., Nagae, G., Schnittger, S., Sanada, M., Kon, A., Alpermann, T., Yoshida, K., Roller, A., Nadarajah, N., Shiraishi, Y., Shiozawa, Y., Chiba, K., Tanaka, H., Koeffler, H. P., Klein, H. U., Dugas, M., Aburatani, H., Kohlmann, A., Miyano, S., Haferlach, C., Kern, W. & Ogawa, S. Landscape of genetic lesions in 944 patients with myelodysplastic syndromes. *Leukemia* **28**, 241-247, doi:10.1038/leu.2013.336 (2014).

37 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. & Tuschl, T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **141**, 129-141, doi:10.1016/j.cell.2010.03.009 (2010).

38 Hamilton, M. P., Rajapakshe, K. I., Bader, D. A., Cerne, J. Z., Smith, E. A., Coarfa, C., Hartig, S. M. & McGuire, S. E. The Landscape of microRNA Targeting in Prostate Cancer Defined by AGO-PAR-CLIP. *Neoplasia* **18**, 356-370, doi:10.1016/j.neo.2016.04.008 (2016).

39 Handrigan, G. R., Chitayat, D., Lionel, A. C., Pinsk, M., Vaags, A. K., Marshall, C. R., Dyack, S., Escobar, L. F., Fernandez, B. A., Stegman, J. C., Rosenfeld, J. A., Shaffer, L. G., Goodenberger, M., Hodge, J. C., Cain, J. E., Babul-Hirji, R., Stavropoulos, D. J., Yiu, V., Scherer, S. W. & Rosenblum, N. D. Deletions in 16q24.2 are associated with autism spectrum disorder, intellectual disability and congenital renal malformation. *J Med Genet* **50**, 163-173, doi:10.1136/jmedgenet-2012-101288 (2013).

40 Hart, T., Tong, A. H. Y., Chan, K., Van Leeuwen, J., Seetharaman, A., Aregger, M., Chandrashekhar, M., Hustedt, N., Seth, S., Noonan, A., Habsid, A., Sizova, O., Nedyalkova, L., Climie, R., Tworzyanski, L., Lawson, K., Sartori, M. A., Alibeh, S., Tieu, D., Masud, S., Mero, P., Weiss, A., Brown, K. R., Usaj, M., Billmann, M., Rahman, M., Constanzo, M., Myers, C. L., Andrews, B. J., Boone, C., Durocher, D. & Moffat, J. Evaluation and Design of Genome-Wide CRISPR/SpCas9 Knockout Screens. *G3 (Bethesda)* **7**, 2719-2727, doi:10.1534/g3.117.041277 (2017).

41 He, A. & Pu, W. T. Genome-wide location analysis by pull down of in vivo biotinylated transcription factors. *Curr Protoc Mol Biol* **Chapter 21**, Unit 21 20, doi:10.1002/0471142727.mb2120s92 (2010).

42 Helwak, A. & Tollervey, D. Identification of miRNA-Target RNA Interactions Using CLASH. *Methods Mol Biol* **1358**, 229-251, doi:10.1007/978-1-4939-3067-8_14 (2016).

43 Hsu, J., Reilly, A., Hayes, B. J., Clough, C. A., Konnick, E. Q., Torok-Storb, B., Gulsuner, S., Wu, D., Becker, P. S., Keel, S. B., Abkowitz, J. L. & Doulatov, S. Reprogramming identifies functionally distinct stages of clonal evolution in myelodysplastic syndromes. *Blood* **134**, 186-198, doi:10.1182/blood.2018884338 (2019).

44 Hurt, J. A., Robertson, A. D. & Burge, C. B. Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res* **23**, 1636-1650, doi:10.1101/gr.157354.113 (2013).

45 Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wuhr, M., Chick, J., Zhai, B., Kolippakkam, D., Mintseris, J., Obar, R. A., Harris, T., Artavanis-Tsakonas, S., Sowa, M. E., De Camilli, P., Paulo, J. A., Harper, J. W. & Gygi, S. P. The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440, doi:10.1016/j.cell.2015.06.043 (2015).

46 Ilagan, J. O., Ramakrishnan, A., Hayes, B., Murphy, M. E., Zebari, A. S., Bradley, P. & Bradley, R. K. U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res* **25**, 14-26, doi:10.1101/gr.181016.114 (2015).

47 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

48 Jain, S., Wheeler, J. R., Walters, R. W., Agrawal, A., Barsic, A. & Parker, R. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **164**, 487-498, doi:10.1016/j.cell.2015.12.038 (2016).

49 Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., Salame, T. M., Tanay, A., van Oudenaarden, A. & Amit, I. Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell* **167**, 1883-1896 e1815, doi:10.1016/j.cell.2016.11.039 (2016).

50 Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A. & Charpentier, E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821, doi:10.1126/science.1225829 (2012).

51 Kedersha, N. & Anderson, P. Mammalian stress granules and processing bodies. *Methods Enzymol* **431**, 61-81, doi:10.1016/S0076-6879(07)31005-7 (2007).

52 Keene, J. D., Komisarow, J. M. & Friedersdorf, M. B. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc* **1**, 302-307, doi:10.1038/nprot.2006.47 (2006).

53 Kiger, A. A., Baum, B., Jones, S., Jones, M. R., Coulson, A., Echeverri, C. & Perrimon, N. A functional genomic analysis of cell morphology using RNA interference. *J Biol* **2**, 27, doi:10.1186/1475-4924-2-27 (2003).

54 Kim, E., Ilagan, J. O., Liang, Y., Daubner, G. M., Lee, S. C., Ramakrishnan, A., Li, Y., Chung, Y. R., Micol, J. B., Murphy, M. E., Cho, H., Kim, M. K., Zebari, A. S., Aumann, S., Park, C. Y., Buonamici, S., Smith, P. G., Deeg, H. J., Lobry, C., Aifantis, I., Modis, Y., Allain, F. H., Halene, S., Bradley, R. K. & Abdel-Wahab, O. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27**, 617-630, doi:10.1016/j.ccell.2015.04.006 (2015).

55 Kishore, S., Gruber, A. R., Jedlinski, D. J., Syed, A. P., Jorjani, H. & Zavolan, M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol* **14**, R45, doi:10.1186/gb-2013-14-5-r45 (2013).

56 Kon, A., Yamazaki, S., Nannya, Y., Kataoka, K., Ota, Y., Nakagawa, M. M., Yoshida, K., Shiozawa, Y., Morita, M., Yoshizato, T., Sanada, M., Nakayama, M., Koseki, H., Nakauchi, H. & Ogawa, S. Physiological Srsf2 P95H expression causes impaired hematopoietic stem cell functions and aberrant RNA splicing in mice. *Blood* **131**, 621-635, doi:10.1182/blood-2017-01-762393 (2018).

57 Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M. & Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909-915, doi:10.1038/nsmb.1838 (2010).

58 Kotini, A. G., Chang, C. J., Boussaad, I., Delrow, J. J., Dolezal, E. K., Nagulapally, A. B., Perna, F., Fishbein, G. A., Klimek, V. M., Hawkins, R. D., Huangfu, D., Murry, C. E., Graubert, T., Nimer, S. D. & Papapetrou, E. P. Functional analysis of a chromosomal deletion associated with myelodysplastic syndromes using isogenic human induced pluripotent stem cells. *Nat Biotechnol* **33**, 646-655, doi:10.1038/nbt.3178 (2015).

59 Kotini, A. G., Chang, C. J., Chow, A., Yuan, H., Ho, T. C., Wang, T., Vora, S., Solovyov, A., Husser, C., Olszewska, M., Teruya-Feldstein, J., Perumal, D., Klimek, V. M., Spyridonidis, A., Rampal, R. K., Silverman, L., Reddy, E. P., Papaemmanuil, E., Parekh, S., Greenbaum, B. D., Leslie, C. S., Kharas, M. G. & Papapetrou, E. P. Stage-Specific Human Induced Pluripotent Stem Cells Map the Progression of Myeloid Transformation to Transplantable Leukemia. *Cell Stem Cell* **20**, 315-328 e317, doi:10.1016/j.stem.2017.01.009 (2017).

60 Kudinov, A. E., Karanicolas, J., Golemis, E. A. & Boumber, Y. Musashi RNA-Binding Proteins as Cancer Drivers and Novel Therapeutic Targets. *Clin Cancer Res* **23**, 2143-2153, doi:10.1158/1078-0432.CCR-16-2728 (2017).

61 LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444, doi:10.1038/nature14539 (2015).

62 Lee, K. H., Zhang, P., Kim, H. J., Mitrea, D. M., Sarkar, M., Freibaum, B. D., Cika, J., Coughlin, M., Messing, J., Molliex, A., Maxwell, B. A., Kim, N. C., Temirov, J., Moore, J., Kolaitis, R. M., Shaw, T. I., Bai, B., Peng, J., Kriwacki, R. W. & Taylor, J. P. C9orf72 Dipeptide Repeats Impair the Assembly, Dynamics, and Function of Membrane-Less Organelles. *Cell* **167**, 774-788 e717, doi:10.1016/j.cell.2016.10.002 (2016).

63 Lee, S. C., Dvinge, H., Kim, E., Cho, H., Micol, J. B., Chung, Y. R., Durham, B. H., Yoshimi, A., Kim, Y. J., Thomas, M., Lobry, C., Chen, C. W., Pastore, A., Taylor, J., Wang, X., Krivtsov, A., Armstrong, S. A., Palacino, J., Buonamici, S., Smith, P. G., Bradley, R. K. & Abdel-Wahab, O. Modulation of splicing catalysis for therapeutic targeting of leukemia with mutations in genes encoding spliceosomal proteins. *Nat Med* **22**, 672-678, doi:10.1038/nm.4097 (2016).

64 Lee, S. C., North, K., Kim, E., Jang, E., Obeng, E., Lu, S. X., Liu, B., Inoue, D., Yoshimi, A., Ki, M., Yeo, M., Zhang, X. J., Kim, M. K., Cho, H., Chung, Y. R., Taylor, J., Durham, B. H., Kim, Y. J., Pastore, A., Monette, S., Palacino, J., Seiler, M., Buonamici, S., Smith, P. G., Ebert, B. L., Bradley, R. K. & Abdel-Wahab, O. Synthetic Lethal and Convergent Biological Effects of Cancer-Associated Spliceosomal Gene Mutations. *Cancer Cell* **34**, 225-241 e228, doi:10.1016/j.ccell.2018.07.003 (2018).

65 Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C. & Darnell, R. B. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469, doi:10.1038/nature07488 (2008).

66 Link, W., Oyarzabal, J., Serelde, B. G., Albarran, M. I., Rabal, O., Cebria, A., Alfonso, P., Fominaya, J., Renner, O., Peregrina, S., Soilan, D., Ceballos, P. A., Hernandez, A. I., Lorenzo, M., Pevarello, P., Granda, T. G., Kurz, G., Carnero, A. & Bischoff, J. R. Chemical interrogation of FOXO3a nuclear translocation identifies potent and selective inhibitors of phosphoinositide 3-kinases. *J Biol Chem* **284**, 28392-28400, doi:10.1074/jbc.M109.038984 (2009).

67 Liu, T., Sims, D. & Baum, B. Parallel RNAi screens across different cell lines identify generic and cell type-specific regulators of actin organization and cell morphology. *Genome Biol* **10**, R26, doi:10.1186/gb-2009-10-3-r26 (2009).

68 Liu, Z. R., Wilkie, A. M., Clemens, M. J. & Smith, C. W. Detection of double-stranded RNA-protein interactions by methylene blue-mediated photo-crosslinking. *RNA* **2**, 611-621 (1996).

69 Lovci, M. T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T. Y., Stark, T. J., Gehman, L. T., Hoon, S., Massirer, K. B., Pratt, G. A., Black, D. L., Gray, J. W.,

Conboy, J. G. & Yeo, G. W. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat Struct Mol Biol* **20**, 1434-1442, doi:10.1038/nsmb.2699 (2013).

70 Low, K. J., Ansari, M., Abou Jamra, R., Clarke, A., El Chehadeh, S., FitzPatrick, D. R., Greenslade, M., Henderson, A., Hurst, J., Keller, K., Kuentz, P., Prescott, T., Roessler, F., Selmer, K. K., Schneider, M. C., Stewart, F., Tatton-Brown, K., Thevenon, J., Vigeland, M. D., Vogt, J., Willems, M., Zonana, J., Study, D. D. & Smithson, S. F. PUF60 variants cause a syndrome of ID, short stature, microcephaly, coloboma, craniofacial, cardiac, renal and spinal features. *Eur J Hum Genet* **25**, 552-559, doi:10.1038/ejhg.2017.27 (2017).

71 Maatz, H., Kolinski, M., Hubner, N. & Landthaler, M. Transcriptome-wide Identification of RNA-binding Protein Binding Sites Using Photoactivatable-Ribonucleoside-Enhanced Crosslinking Immunoprecipitation (PAR-CLIP). *Curr Protoc Mol Biol* **118**, 27 26 21-27 26 19, doi:10.1002/cpmb.35 (2017).

72 Maharana, S., Wang, J., Papadopoulos, D. K., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillen-Boixet, J., Franzmann, T. M., Jahnel, M., Marrone, L., Chang, Y. T., Sterneckert, J., Tomancak, P., Hyman, A. A. & Alberti, S. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science* **360**, 918-921, doi:10.1126/science.aar7366 (2018).

73 Markmiller, S., Soltanieh, S., Server, K. L., Mak, R., Jin, W., Fang, M. Y., Luo, E. C., Krach, F., Yang, D., Sen, A., Fulzele, A., Wozniak, J. M., Gonzalez, D. J., Kankel, M. W., Gao, F. B., Bennett, E. J., Lecuyer, E. & Yeo, G. W. Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell* **172**, 590-604 e513, doi:10.1016/j.cell.2017.12.032 (2018).

74 Martinez, F. J., Pratt, G. A., Van Nostrand, E. L., Batra, R., Huelga, S. C., Kapeli, K., Freese, P., Chun, S. J., Ling, K., Gelboin-Burkhart, C., Fijany, L., Wang, H. C., Nussbacher, J. K., Broski, S. M., Kim, H. J., Lardelli, R., Sundararaman, B., Donohue, J. P., Javaherian, A., Lykke-Andersen, J., Finkbeiner, S., Bennett, C. F., Ares, M., Jr., Burge, C. B., Taylor, J. P., Rigo, F. & Yeo, G. W. Protein-RNA Networks Regulated by Normal and ALS-Associated Mutant HNRNPA2B1 in the Nervous System. *Neuron* **92**, 780-795, doi:10.1016/j.neuron.2016.09.050 (2016).

75 McEwen, E., Kedersha, N., Song, B., Scheuner, D., Gilks, N., Han, A., Chen, J. J., Anderson, P. & Kaufman, R. J. Heme-regulated inhibitor kinase-mediated phosphorylation of eukaryotic translation initiation factor 2 inhibits translation, induces stress granule formation, and mediates survival upon arsenite exposure. *J Biol Chem* **280**, 16925-16933, doi:10.1074/jbc.M412882200 (2005).

76 McMahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W. & Rosbash, M. TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell* **165**, 742-753, doi:10.1016/j.cell.2016.03.007 (2016).

77 Mealiea, W. L., Jr. & Courts, F. J. Computer technology and dentistry. *Todays FDA* **2**, 1C, 3C (1990).

78 Meisenheimer, K. M. & Koch, T. H. Photocross-linking of nucleic acids to associated proteins. *Crit Rev Biochem Mol Biol* **32**, 101-140, doi:10.3109/10409239709108550 (1997).

79 Mili, S. & Steitz, J. A. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA* **10**, 1692-1694, doi:10.1261/rna.7151404 (2004).

80 Muller-McNicoll, M. & Neugebauer, K. M. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet* **14**, 275-287, doi:10.1038/nrg3434 (2013).

81 Nelles, D. A., Fang, M. Y., O'Connell, M. R., Xu, J. L., Markmiller, S. J., Doudna, J. A. & Yeo, G. W. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell* **165**, 488-496, doi:10.1016/j.cell.2016.02.054 (2016).

82 Nicholson, C. O., Friedersdorf, M. & Keene, J. D. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. *RNA* **23**, 32-46, doi:10.1261/rna.058115.116 (2017).

83 Nussbacher, J. K., Batra, R., Lagier-Tourenne, C. & Yeo, G. W. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends Neurosci* **38**, 226-236, doi:10.1016/j.tins.2015.02.003 (2015).

84 Ohn, T., Kedersha, N., Hickman, T., Tisdale, S. & Anderson, P. A functional RNAi screen links O-GlcNAc modification of ribosomal proteins to stress granule and processing body assembly. *Nat Cell Biol* **10**, 1224-1231, doi:10.1038/ncb1783 (2008).

85 Okeyo-Owuor, T., White, B. S., Chatrikhi, R., Mohan, D. R., Kim, S., Griffith, M., Ding, L., Ketkar-Kulkarni, S., Hundal, J., Laird, K. M., Kielkopf, C. L., Ley, T. J., Walter, M. J. & Graubert, T. A. U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. *Leukemia* **29**, 909-917, doi:10.1038/leu.2014.303 (2015).

86 Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V. I., Paschka, P., Roberts, N. D., Potter, N. E., Heuser, M., Thol, F., Bolli, N., Gundem, G., Van Loo, P., Martincorena, I., Ganly, P., Mudie, L., McLaren, S., O'Meara, S., Raine, K., Jones, D. R., Teague, J. W., Butler, A. P., Greaves, M. F., Ganser, A., Dohner, K., Schlenk, R. F., Dohner, H. & Campbell, P. J. Genomic Classification and Prognosis in Acute Myeloid Leukemia. *N Engl J Med* **374**, 2209-2221, doi:10.1056/NEJMoa1516192 (2016).

87 Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C. J., Ellis, P., Wedge, D. C., Pellagatti, A., Shlien, A., Groves, M. J., Forbes, S. A., Raine, K., Hinton, J., Mudie, L. J., McLaren, S., Hardy, C., Latimer, C., Della Porta, M. G., O'Meara, S., Ambaglio, I., Galli, A., Butler, A. P., Walldin, G., Teague, J. W., Quek, L., Sternberg, A., Gambacorti-Passerini, C., Cross, N. C., Green, A. R., Boultwood, J., Vyas, P., Hellstrom-Lindberg, E., Bowen, D., Cazzola, M., Stratton, M. R., Campbell, P. J. & Chronic Myeloid Disorders Working Group of the International Cancer Genome, C. Clinical and biological implications of driver mutations in myelodysplastic

syndromes. *Blood* **122**, 3616-3627; quiz 3699, doi:10.1182/blood-2013-08-518886 (2013).

88 Parnas, O., Jovanovic, M., Eisenhaure, T. M., Herbst, R. H., Dixit, A., Ye, C. J., Przybylski, D., Platt, R. J., Tirosh, I., Sanjana, N. E., Shalem, O., Satija, R., Raychowdhury, R., Mertins, P., Carr, S. A., Zhang, F., Hacohen, N. & Regev, A. A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell* **162**, 675-686, doi:10.1016/j.cell.2015.06.059 (2015).

89 Pellagatti, A., Armstrong, R. N., Steeples, V., Sharma, E., Repapi, E., Singh, S., Sanchi, A., Radujkovic, A., Horn, P., Dolatshad, H., Roy, S., Broxholme, J., Lockstone, H., Taylor, S., Giagounidis, A., Vyas, P., Schuh, A., Hamblin, A., Papaemmanuil, E., Killick, S., Malcovati, L., Hennrich, M. L., Gavin, A. C., Ho, A. D., Luft, T., Hellstrom-Lindberg, E., Cazzola, M., Smith, C. W. J., Smith, S. & Boultwood, J. Impact of spliceosome mutations on RNA splicing in myelodysplasia: dysregulated genes/pathways and clinical associations. *Blood* **132**, 1225-1240, doi:10.1182/blood-2018-04-843771 (2018).

90 Penalva, L. O., Burdick, M. D., Lin, S. M., Sutterluety, H. & Keene, J. D. RNA-binding proteins to assess gene expression states of co-cultivated cells in response to tumor cells. *Mol Cancer* **3**, 24, doi:10.1186/1476-4598-3-24 (2004).

91 Przychodzen, B., Jerez, A., Guinta, K., Sekeres, M. A., Padgett, R., Maciejewski, J. P. & Makishima, H. Patterns of missplicing due to somatic U2AF1 mutations in myeloid neoplasms. *Blood* **122**, 999-1006, doi:10.1182/blood-2013-01-480970 (2013).

92 Saez, B., Walter, M. J. & Graubert, T. A. Splicing factor gene mutations in hematologic malignancies. *Blood* **129**, 1260-1269, doi:10.1182/blood-2016-10-692400 (2017).

93 Schneider-Poetsch, T., Ju, J., Eyler, D. E., Dang, Y., Bhat, S., Merrick, W. C., Green, R., Shen, B. & Liu, J. O. Inhibition of eukaryotic translation elongation by cycloheximide and lactimidomycin. *Nat Chem Biol* **6**, 209-217, doi:10.1038/nchembio.304 (2010).

94 Seiler, M., Yoshimi, A., Darman, R., Chan, B., Keaney, G., Thomas, M., Agrawal, A. A., Caleb, B., Csibi, A., Sean, E., Fekkes, P., Karr, C., Klimek, V., Lai, G., Lee, L., Kumar, P., Lee, S. C., Liu, X., Mackenzie, C., Meeske, C., Mizui, Y., Padron, E., Park, E., Pazolli, E., Peng, S., Prajapati, S., Taylor, J., Teng, T., Wang, J., Warmuth, M., Yao, H., Yu, L., Zhu, P., Abdel-Wahab, O., Smith, P. G. & Buonamici, S. H3B-8800, an orally available small-molecule splicing modulator, induces lethality in spliceosome-mutant cancers. *Nat Med* **24**, 497-504, doi:10.1038/nm.4493 (2018).

95 Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelson, T., Heckl, D., Ebert, B. L., Root, D. E., Doench, J. G. & Zhang, F. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).

96 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

97 Shiozawa, Y., Malcovati, L., Galli, A., Sato-Otsubo, A., Kataoka, K., Sato, Y., Watatani, Y., Suzuki, H., Yoshizato, T., Yoshida, K., Sanada, M., Makishima, H., Shiraishi, Y., Chiba, K., Hellstrom-Lindberg, E., Miyano, S., Ogawa, S. & Cazzola, M. Aberrant splicing and defective mRNA production induced by somatic spliceosome mutations in myelodysplasia. *Nat Commun* **9**, 3649, doi:10.1038/s41467-018-06063-x (2018).

98 Shirai, C. L., Ley, J. N., White, B. S., Kim, S., Tibbitts, J., Shao, J., Ndonwi, M., Wadugu, B., Duncavage, E. J., Okeyo-Owuor, T., Liu, T., Griffith, M., McGrath, S., Magrini, V., Fulton, R. S., Fronick, C., O'Laughlin, M., Graubert, T. A. & Walter, M. J. Mutant U2AF1 Expression Alters Hematopoiesis and Pre-mRNA Splicing In Vivo. *Cancer Cell* **27**, 631-643, doi:10.1016/j.ccell.2015.04.008 (2015).

99 Singh, G., Ricci, E. P. & Moore, M. J. RIPiT-Seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods* **65**, 320-332, doi:10.1016/j.ymeth.2013.09.013 (2014).

100 Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491-499, doi:10.1101/gr.209601.116 (2017).

101 Sprang, S. R. Invited review: Activation of G proteins by GTP and the mechanism of Galpha-catalyzed GTP hydrolysis. *Biopolymers* **105**, 449-462, doi:10.1002/bip.22836 (2016).

102 Sugimoto, Y., Konig, J., Hussain, S., Zupan, B., Curk, T., Frye, M. & Ule, J. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol* **13**, R67, doi:10.1186/gb-2012-13-8-r67 (2012).

103 Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N. M. & Ule, J. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature* **519**, 491-494, doi:10.1038/nature14280 (2015).

104 Sundararaman, B., Zhan, L., Blue, S. M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E. L., Pratt, G. A., Huelga, S. C., Smalec, B. M., Wang, X., Hong, E. L., Davidson, J. M., Lecuyer, E., Graveley, B. R. & Yeo, G. W. Resources for the Comprehensive Discovery of Functional RNA Elements. *Mol Cell* **61**, 903-913, doi:10.1016/j.molcel.2016.02.012 (2016).

105 Tartaglia, G. G. The Grand Challenge of Characterizing Ribonucleoprotein Networks. *Front Mol Biosci* **3**, 24, doi:10.3389/fmolb.2016.00024 (2016).

106 Tenenbaum, S. A., Carson, C. C., Lager, P. J. & Keene, J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A* **97**, 14085-14090, doi:10.1073/pnas.97.26.14085 (2000).

107 Tourriere, H., Chebli, K., Zekri, L., Courselaud, B., Blanchard, J. M., Bertrand, E. & Tazi, J. The RasGAP-associated endoribonuclease G3BP assembles stress granules. *J Cell Biol* **160**, 823-831, doi:10.1083/jcb.200212128 (2003).

108 Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A. & Darnell, R. B. CLIP identifies Nova-regulated RNA networks in the brain. *Science* **302**, 1212-1215, doi:10.1126/science.1090095 (2003).

109 Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., Blencowe, B. J. & Darnell, R. B. An RNA map predicting Nova-dependent splicing regulation. *Nature* **444**, 580-586, doi:10.1038/nature05304 (2006).

110 Van Nostrand, E. L., Gelboin-Burkhart, C., Wang, R., Pratt, G. A., Blue, S. M. & Yeo, G. W. CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods* **118-119**, 50-59, doi:10.1016/j.ymeth.2016.12.007 (2017).

111 Van Nostrand, E. L., Huelga, S. C. & Yeo, G. W. Experimental and Computational Considerations in the Study of RNA-Binding Protein-RNA Interactions. *Adv Exp Med Biol* **907**, 1-28, doi:10.1007/978-3-319-29073-7_1 (2016).

112 Van Nostrand, E. L., Nguyen, T. B., Gelboin-Burkhart, C., Wang, R., Blue, S. M., Pratt, G. A., Louie, A. L. & Yeo, G. W. Robust, Cost-Effective Profiling of RNA Binding Protein Targets with Single-end Enhanced Crosslinking and Immunoprecipitation (seCLIP). *Methods Mol Biol* **1648**, 177-200, doi:10.1007/978-1-4939-7204-3_14 (2017).

113 Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508-514, doi:10.1038/nmeth.3810 (2016).

114 Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., Rabano, I., Stanton, R., Sundararaman, B., Wang, R., Fu, X. D., Graveley, B. R. & Yeo, G. W. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol* **21**, 90, doi:10.1186/s13059-020-01982-9 (2020).

115 Wang, C., Lu, T., Emanuel, G., Babcock, H. P. & Zhuang, X. Imaging-based pooled CRISPR screening reveals regulators of lncRNA localization. *Proc Natl Acad Sci U S A* **116**, 10842-10851, doi:10.1073/pnas.1903808116 (2019).

116 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).

117 Wang, Y., Phillips, C., Xu, W., Pai, J. H., Dhopeshwarkar, R., Sims, C. E. & Allbritton, N. Micromolded arrays for separation of adherent cells. *Lab Chip* **10**, 2917-2924, doi:10.1039/c0lc00186d (2010).

118 Wolozin, B. & Ivanov, P. Stress granules and neurodegeneration. *Nat Rev Neurosci* **20**, 649-666, doi:10.1038/s41583-019-0222-5 (2019).

119 Yee, B. A., Pratt, G. A., Graveley, B. R., Van Nostrand, E. L. & Yeo, G. W. RBP-Maps enables robust generation of splicing regulatory maps. *RNA* **25**, 193-204, doi:10.1261/rna.069237.118 (2019).

120 Yeo, G. W., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. B. Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci U S A* **102**, 2850-2855, doi:10.1073/pnas.0409742102 (2005).

121 Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., Sato, Y., Sato-Otsubo, A., Kon, A., Nagasaki, M., Chalkidis, G., Suzuki, Y., Shiosaka, M., Kawahata, R., Yamaguchi, T., Otsu, M., Obara, N., Sakata-Yanagimoto, M., Ishiyama, K., Mori, H., Nolte, F., Hofmann, W. K., Miyawaki, S., Sugano, S., Haferlach, C., Koeffler, H. P., Shih, L. Y., Haferlach, T., Chiba, S., Nakauchi, H., Miyano, S. & Ogawa, S. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69, doi:10.1038/nature10496 (2011).

122 Youn, J. Y., Dunham, W. H., Hong, S. J., Knight, J. D. R., Bashkurov, M., Chen, G. I., Bagci, H., Rathod, B., MacLeod, G., Eng, S. W. M., Angers, S., Morris, Q., Fabian, M., Cote, J. F. & Gingras, A. C. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. *Mol Cell* **69**, 517-532 e511, doi:10.1016/j.molcel.2017.12.020 (2018).

123 Zarnegar, B. J., Flynn, R. A., Shen, Y., Do, B. T., Chang, H. Y. & Khavari, P. A. irCLIP platform for efficient characterization of protein-RNA interactions. *Nat Methods* **13**, 489-492, doi:10.1038/nmeth.3840 (2016).

124 Zhang, J., Lieu, Y. K., Ali, A. M., Penson, A., Reggio, K. S., Rabadan, R., Raza, A., Mukherjee, S. & Manley, J. L. Disease-associated mutation in SRSF2 misregulates splicing by altering RNA-binding affinities. *Proc Natl Acad Sci U S A* **112**, E4726-4734, doi:10.1073/pnas.1514105112 (2015).

125 Zhang, X., Lai, T. Y., Chiang, S. W., Tam, P. O., Liu, D. T., Chan, C. K., Pang, C. P., Zhao, C. & Chen, L. J. Contribution of SNRNP200 sequence variations to retinitis pigmentosa. *Eye (Lond)* **27**, 1204-1213, doi:10.1038/eye.2013.137 (2013).