CORRELATION OF TRANSCRIPTOME-WIDE PROTEIN BINDING TO INFORM RNA REGULATION

by

Veronica F. Busa

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

May 2022

© 2022 Veronica F. Busa

All Rights Reserved

Abstract

Dysregulated control of RNA can cause abnormal development and disease. RNA is targeted and regulated via the interaction of RNA structure, modifications, and RNA-binding proteins. Information about the interactions of transcriptomic features can therefore provide biologically relevant insight into RNA control. We develop a new algorithm, nearBynding, to calculate spatial correlations between transcriptomic data types and demonstrate its ability to recapitulate known biological relationships. We apply nearBynding to formulate new hypotheses about proteins that bind in regions of the transcriptome that are prone to G-quadruplex formation. We use nearBynding to observe the colocalization of G3BP1 and UPF1 within 3'UTRs for structure-mediated RNA decay (SRD) and to identify additional candidate factors involved in SRD. We provide evidence that there is a discrepancy in SRD activity between breast cancer samples derived from African American and European American patients. Finally, we investigate the role of the RNA-binding protein PARP13 in the innate immune response via integration of genomics data types. Evidence from genomic data, nearBynding analysis, and molecular biology experiments provides insight into the geometry of PARP13's interface with its cofactor TRIM25 and implicates PARP13 in creating and maintaining a cellular environment poised for an antiviral response.

Primary reader and advisor: Dr. Anthony K.L. Leung

Secondary readers: Dr. Sarah Woodson and Dr. Elana J. Fertig

Acknowledgements

I would like to thank Kurt Weir for his extensive suggestions and edits throughout all works presented herein; all members of the Leung Lab– especially Joe Fischer and Anthony– who provided experimental ideas, troubleshooting support, and wet lab training; members of the Fertig Lab– especially Genevieve Stein-O'Brien, Sasha Favorov, and Elana– who provided computational and statistical ideas, assistance, and resources; and the administrators of the Department of Human Genetics– especially Sandy Muscelli– who facilitated my ongoing education.

Chapter 2: Sasha Favorov provided computational training, troubleshooting, and discussion to create nearBynding. Elana Fertig and Anthony Leung provided mentoring to conceive of, test, and demonstrate the biological relevance of nearBynding.

Chapter 3: Rishi Goel and Ahmed El-Morsy assisted in developing the global RNA structure metric, identifying the race-based discrepancy in global RNA structure, and characterizing the drivers of the discrepancy. Sumit Siddarth assisted in the selection of cultured breast cancer cell lines and provided pellets of the selected lines. Elana Fertig, Dipali Sharma, and Anthony Leung provided mentoring to inform experimental design.

Chapter 4: Yoshi Ando generated the RNA-seq data and worked with Brian Yee and Gene Yeo to generate the CLIP-seq data. Many members of the Leung Lab–especially Junlin Zhuo, Shang-Jung Cheng and Aravinth Jayabalan– provided reagents and protocols for transfection, immunoprecipitation, and western blots. Anthony Leung provided mentoring to inform experimental design.

Table of Contents

ABSTRACT ..... **II**

ACKNOWLEDGEMENTS ..... **III**

TABLE OF CONTENTS ..... **IV**

LIST OF TABLES ..... **VIII**

LIST OF FIGURES ..... **IX**

LIST OF ABBREVIATIONS ..... **XI**

CHAPTER 1: INTRODUCTION **1**

CHAPTER 2: SPATIAL CORRELATION STATISTICS ENABLE TRANSCRIPTOME-WIDE CHARACTERIZATION OF RNA STRUCTURE BINDING **4**

HIGHLIGHTS 4

INTRODUCTION 4

RESULTS 7

nearBynding probes transcriptome-wide RBP binding to RNA structures *7*

Comparison of available spatial relationship algorithms *9*

Benchmarking on simulated data demonstrates robust signal detection *14*

Called peaks or aligned tracks for RBP binding produce similar binding profiles *16*

Cross-correlation tracks reproducibly cluster RBP data across biological replicates *18*

RNA structure cross-correlation signal is specific *19*

RBP binding profiles recapitulate known structural preferences *20*

Differentiating wild-type and mutant RBP structural preferences *22*

Probe protein binding profile symmetry within 3'UTRs *23*

Illustrate relative binding positions of RBPs *24*

Inform RBP binding preferences using experimentally-derived RNA annotations *24*

Analyzing RBP binding to G-quadruplexes 24

Analyzing RBP preferences for RNA modification 27

DISCUSSION 28
METHODS 31
Resource Availability *31*
Workflow details *33*
nearBynding inputs 33
Map data to pseudochromosomes 33
CapR RNA structure prediction 34
Relative binding position calculation 34
nearBynding output analyses 35
Quantification and Statistical Analysis *35*
FIGURES AND TABLES 36
CHAPTER 3: LINKING STRUCTURE-MEDIATED RNA DECAY AND RACIAL DISPARITIES IN BREAST CANCER 59
HIGHLIGHTS 59
INTRODUCTION 59
RESULTS 61
-ΔG/nt correlates with G3BP1-dependent decay for isoforms with highly-structured 3'UTRs *61*
G3BP1 expression is a predictor of novel global RNA structure metric *62*
HSU/PSU ratio separates African American and European American cancer samples *63*
SRD drives HSU/PSU difference between European American and African American breast cancer samples *64*
Verification of SRD-derived phenotype in cultured breast cancer cells *65*
Searching for additional factors associated with SRD in publicly-available datasets *66*
miRNAs are potential SRD regulators 66
Correlation with HSU/PSU ratio is a poor method to discover other SRD factor candidates 67
Spatial correlation via nearBynding software of protein binding from eCLIP data provides insight into SRD protein colocalization 68
DISCUSSION 70
METHODS 73

Code availability *73*

Real-time quantitative PCR *74*

FIGURES AND TABLES 74

CHAPTER 4: TRANSCRIPTOME REGULATION BY PARP13 IN BASAL AND ANTIVIRAL STATES . 89

HIGHLIGHTS 89

INTRODUCTION 89

RESULTS 91

PARP13 regulates cellular homeostasis and the innate antiviral response *91*

PARP13 suppresses PARP13.1 but not PARP13.2 translation in the absence of an antiviral response *92*

The antiviral response shifts PARP13 target localization but not its binding preferences *94*

PARP13 supports expression of transcripts related to the ISGylation and RIG-I pathway *95*

Binding to common transcripts reinforces TRIM25:PARP13 interactions *96*

DISCUSSION 98

METHODS 101

Cell culture *101*

ssRNA and 3p-RNA synthesis and transfection *101*

Dual luciferase assay *101*

siRNA transfection *102*

GFP construct transfection and immunoprecipitation *102*

Western blot *103*

RT-PCR *103*

RNA-seq analysis *103*

CLIP-seq analysis *104*

CHAPTER 5: DISCUSSIONS 114

LIMITATIONS OF NEARBYNDING 114

LIMITATIONS OF DATA COMPARISONS 116

NEW AREAS OF INTERROGATION 117

REFERENCES 119

APPENDIX A: RNA STRUCTURE IMPACTS CIRCULAR RNA REGULATION AND FUNCTION 134

List of Tables

Table 2.1 Comparison of functionality for correlation analysis programs 55

Table 2.2 Multivariate regression model predicting protein binding profile symmetry from all maximal RNA structure context preference signals 55

Table 2.3 Literature sources of RNA G4-binding proteins 56

Table 2.4 Influence of RBP domains on G4 binding signal 57

Table 2.5 Key Resources 58

Table 3.1 Coefficients of linear models assessing HSU/PSU ratio versus $\log_2$ G3BP1 expression for all TCGA cancer types 85

Table 3.2 Top 10 hits HSU/PSU ~ $\log_2$(gene expression) 86

Table 3.3 Top 10 hits HSU/PSU ~ $\log_2$(gene expression) + $\log_2$(G3BP1) 86

Table 3.4 Breast cancer cell lines tested 87

Table 3.5 Proteins that co-localize with G3BP1 in HepG2 3'UTRs 87

Table 3.6 Key Resources 88

Table 4.1 Key resources 112

List of Figures

Figure 2.1 Overview of nearBynding 36

Figure 2.2 Removing cell line-specific background from binding profiles 37

Figure 2.3 Graphical representations of different data types 38

Figure 2.4 deepTools output 39

Figure 2.5 Graphed BEDtools *reldist* outputs 40

Figure 2.6 Cross-correlation distribution tracks of simulated RNA binding data to benchmark the performance of nearBynding 41

Figure 2.7 Comparison of aligned reads and called peaks as nearBynding inputs 43

Figure 2.8 Binding profiles of all RBPs with eCLIP data in ENCODE clustered using Wasserstein distance 45

Figure 2.9 Benchmarking reproducibility and specificity based on ENCODE data 47

Figure 2.10 Application of nearBynding to analyze RBP binding profiles for proteins with known structure preference 49

Figure 2.11 The symmetry of protein binding profiles within 3'UTRs correlates with the strength of a protein's preference for stem context 50

Figure 2.12 Recapitulating known RBP binding preferences 51

Figure 2.13 RBP binding to experimentally-derived exonic G4s 52

Figure 2.14 G4 binding signal strength relative to transcriptomic regions and protein domains .53

Figure 2.15 RBP binding profiles for experimentally-derived m6A 54

Figure 3.1 G3BP1 regulates much of the transcriptome 75

Figure 3.2 G3BP1 expression is a predictor of global RNA structure 76

Figure 3.3 Global RNA structure is often greater in African American cancer patients 77

Figure 3.4 Highly structured isoforms contribute disproportionately to race-based global RNA structure disparity 79

Figure 3.5 Cultured breast cancer samples demonstrate poor reproducibility 80

Figure 3.6 Race-based differences in miRNA expression may drive global RNA structure and cancer prognosis differences 82

Figure 3.7 nearBynding identifies proteins that colocalize with G3BP1 84

Figure 4.1 PARP13 expression affects expression of many constitutive transcripts and is required for transcriptomic upregulation of innate immune response.....105

Figure 4.2 PARP13 selectively binds PARP13.1 and regulates its translation via the 3' UTR..107

Figure 4.3 Antiviral response shifts PARP13 target localization but not binding preferences...109

Figure 4.4 Genes associated with ISGylation are dysregulated in PARP13 KO cells 110

Figure 4.5 TRIM25 and PARP13 interact via both RNA-dependent and RNA-independent contacts 111

List of Abbreviations

AAV	adeno-associated virus
actD	Actinomycin-D
APEX-RIP	(engineered ascorbate peroxidase)-catalyzed proximity biotinylation of endogenous proteins and RNA immunoprecipitation
ASOs	Antisense oligonucleotides
BRCA	Breast invasive carcinoma
BSJ	Back-splice junction
ChIP	Chromatin immunoprecipitation
circRNA	Circular RNA
CLIP-seq	Cross-link immunoprecipitation with high-throughput sequencing
CSD	Cold-shock domain
dsRBD	Double-stranded RNA binding domain
dsRNA	Double-stranded RNA
eCLIP	Enhanced cross-link immunoprecipitation
EJC	Exon junction complex
G4	G-quadruplex
GO	Gene ontology
hiCLIP	Hybrid individual-nucleotide resolution CLIP
HR	Hazard ratio
iCLIP	Individual-nucleotide resolution CLIP
icSHAPE	*in vivo* click SHAPE
IRES	Internal ribosome entry site
KO	Knock-out
KD	Knock-down
LDN	linear DNA nanostructure
m6A	N^6-methyadenosine
miCLIP-seq	m6A-iCLIP sequencing
OR	Odds ratio
RBP	RNA-binding protein
RIP	RNA immunoprecipitation
RPAD	RNase R Treatment, polyadenylation, and poly(A)+ RNA depletion
SHAPE	selective 2′-hydroxyl acylation analyzed by primer extension

SHAPE-MaP	SHAPE and mutational profiling
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
SRD	Structure-mediated RNA decay
ssRNA	Single-stranded RNA
Tornado	Twister-optimized RNA for durable overexpression
tpm	Transcripts per million sequencing reads
tricRNA	tRNA intron-derived circular RNA
TSEN	tRNA splicing endonuclease
UTR	Untranslated region
WT	Wild-type
ZF	Zinc-finger domain

Chapter 1: Introduction

RNA is at the center of gene regulation—either as the intermediate for protein-coding gene expression or as the terminal functional product in noncoding RNAs. Dysregulated control of RNA, either at biogenesis, processing, or degradation, can result in abnormal development [reviewed (1,2)] and can drive disease such as cancer [reviewed (3,4)] or disorders of the immune system [reviewed (5–7)]. RNA structure, protein occupancy, and modification are all linked to RNA function and regulation [Appendix A and (8–10)]. An expanding number of RNA profiling methods have been developed to measure transcriptional attributes associated with RNA structure, modification, and processing. Sequencing-based tools have been developed to elucidate RNA-binding protein (RBP) targets (11); to provide a snapshot of the precise locations of ribosomes on RNA (12); to identify loci with adenosine-to-inosine editing (13); and to provide RNA structure information (14), to name only a small sample of diverse transcriptomic data types and features. To complicate matters, RNA structure, protein occupancy, and modification can each affect one another. Although low-throughput experiments can provide insight about the interactions and causal relationships between two features on a single transcript, each RNA can have dozens or even thousands of modifications and interactors, and each RBP can similarly affect thousands of transcripts. In studying the etiology of disease, it is therefore essential to understand how different RNA features interrelate on a transcriptome-wide level.

Analysis approaches that integrate diverse transcriptomic data modalities and identify inter-related features can lead to novel hypotheses about biological regulation (15). Because these regulatory processes generally occur at co-localized or adjacent transcriptional coordinates, assessing spatial relationships at a transcriptome-wide scale represents a powerful means of evaluating RNA structure, modification and protein occupation. Efficient methods to correlate genome-wide features are available (15,16). Still, robust transcriptome-wide spatial correlation requires new tailored extensions. Transcriptomic data constitutes only a fraction of

the genome and its analysis often requires nucleotide-level spatial relationship resolution, unlike the 100 bp- to megabase-resolution information that usually suffices for genomic data (15,17,18). Therefore, applying genomic tools directly to calculate the spatial correlation of transcriptomic data analysis is computationally inefficient and imprecise. To overcome this limitation, the main approach currently used to assess colocalization of transcriptomic features adapts genome-based tools to compare features at identical transcriptomic locations or within binned regions (overlapping coordinates). However, biologically important relationships often occur at proximal but non-overlapping transcriptomic sites (adjacent coordinates) (19–23), necessitating analyses that evaluate the relative positions of transcriptomic features. Moreover, because of the expanding number of RNA profiling methods, there is a need for tools that are flexible enough to accommodate associations of binary features from processed RNA profiling data and continuous track data from assays that resolve quantitative features along the transcriptome.

Molecular interactions at identical transcriptomic locations or at proximal but non-overlapping sites can mediate RNA modification and regulation, necessitating tools to uncover these spatial relationships. One notable example in which transcriptome-wide correlation is particularly applicable is in the example of RBP preference for RNA sequence, structure, and/or modification. Most RBP binding research has focused on sequence preference. Some RBPs recognize RNA structure more than sequence (24,25), but binding preferences to structured RNA have thoroughly been described for only a few proteins, and RNA structure surrounding protein binding events is rarely characterized. Sequence motifs ascribed to RBPs are often insufficient for explaining a large proportion of binding occurrences (10,26–30). Describing the unexplained binding of RBPs—especially for RBPs that bind structured RNAs—will increase our potential to elucidate the etiology of diseases driven by dysregulated protein-RNA interactions.

G3BP1 is one example of a structured-RNA-binding protein that has been linked to many diseases [reviewed (31)]. It was recently shown to drive structure-mediated RNA decay (SRD), though current methods to predict G3BP1 binding and therefore SRD targets is poor (32). There are no currently-known disease-related effects of dysregulated SRD, but G3BP1's broad functions in RNA regulation suggest that a better understanding of G3BP1 binding and cofactors may provide insight into the mechanism of SRD and its role in disease. Spatially correlating G3BP1 transcriptomic occupancy with that of other RBPs could provide that understanding.

In this work, we first develop a new algorithm, nearBynding, to calculate spatial correlations between transcriptomic data types. We demonstrate nearBynding's ability to recapitulate known RBP preferences for binding structured and modified RNA, relative positions of RBPs along pre-mRNA, and discrepancies between wild-type (WT) and mutant protein binding. We next apply nearBynding to help generate hypotheses about G3BP1's role in SRD and breast cancer. We interrogate the relationship between the SRD pathway and race in breast cancer outcomes by creating a novel metric to predict global RNA structure—and therefore SRD—within a sample. We leverage this metric and correlation profiles generated using nearBynding to produce a candidate list of additional factors that drive SRD. Lastly, we apply nearBynding to explore the role of the RNA-binding protein PARP13 in the innate immune response. We partner nearBynding and molecular biology to understand the geometry and manner of PARP13's interface with its cofactor TRIM25 and implicate PARP13 in maintaining a poised antiviral state.

Chapter 2: Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding

Published:

Busa, Veronica F., Alexander V. Favorov, Elana J. Fertig, and Anthony K. L. Leung. "Spatial Correlation Statistics Enable Transcriptome-Wide Characterization of RNA Structure Binding." *Cell Reports Methods* 1, no. 6 (October 25, 2021). https://doi.org/10.1016/j.crmeth.2021.100088.

Highlights

- nearBynding is an R/Bioconductor algorithm to identify spatial relationships via cross-correlation of diverse transcriptomic features
- nearBynding can accommodate interval or continuous data formats to calculate and visualize features at or adjacent to annotation sites transcriptome-wide
- nearBynding recapitulates known RNA-binding protein binding to structural motifs and provides novel biological insights into RBP binding preference of G-quadruplexes

Introduction

Several machine learning algorithms have been developed to resolve structure-based RBP motifs using cross-link immunoprecipitation (CLIP) data and RNA structure prediction (33–35). Algorithms such as GraphProt and iDeepS incorporate a post-processing step to easily visualize RBP sequence and structure preferences (33,34), but these algorithms only provide visualization of structure information for a short binding motif (7-12 nucleotides). The predictive power of many state-of-the-art algorithms may be limited by their reliance exclusively on sequence-based RNA structure prediction and their lack of accommodation for experimentally-derived RNA structure information [reviewed (36,37)]. The recent algorithm PrismNet has begun addressing these problems by allowing the incorporation of in vivo click selective 2'-hydroxyl

acetylation and profiling experiment (icSHAPE) data (38), but all algorithms to resolve structure-based RBP motifs still rely exclusively on analysis of overlapping coordinates and do not offer insight about the RNA structure surrounding those motifs, despite evidence that such local context can be important in RBP binding (19,39). Current methods interrogating RNA structure binding contexts at and adjacent to the binding site can only be performed in a low-throughput manner (19,39); therefore, RBPs known to have preferred secondary structures are sparse. Efficient methods to perform transcriptome-wide correlation are needed to overcome these limitations and resolve global RBP binding to RNA structure.

Here, we present a new algorithm, nearBynding, to calculate spatial correlations between transcriptomic data types. The nearBynding algorithm is unique in three ways: first, it calculates correlations to indicate features colocalized at, or adjacent to, annotation sites in a transcriptome-wide manner; second, it is a flexible scaffold to cross-correlate diverse transcriptomic data types; and third, it can analyze transcriptomic data in either interval or continuous data formats. The algorithm for nearBynding is implemented in an R/Bioconductor package by the same name to promote accessibility and ease of use. This software also incorporates visualization functions for its statistics to identify colocalizations and adjacent features. To illustrate the application of this transcriptome-wide spatial correlation statistics algorithm, we used nearBynding to analyze RBP preference for RNA structure and RNA modification.

nearBynding provides the software to model spatial relationships between transcriptomic data types, such as between protein binding and RNA structure information, in a targeted manner by allowing users to isolate regions of the genome that correspond to unspliced or mature transcripts as well as subsections of the transcriptome (e.g., 3’UTRs or CDS only). The algorithm achieves this transcriptome-wide correlation by expanding our current tool for genome-wide correlation StereoGene (15) to allow users to select specific transcripts and

specific regions of those transcripts to interrogate (see Appendix 1 and Table S1 for capability comparison). Users can then use nearBynding to input two tracks, calculate the pairwise correlation between those tracks, and visualize the correlation along the transcriptome. If users wish to input only one track from an RNA profiling assay, such as RBP binding, and utilize features from *in silico* RNA structure predictions for the other correlation track, nearBynding includes functions to identify the sequences of selected regions of the transcriptome and derive RNA folding prediction information for those regions. The correlation output for the input track (e.g., RNA binding) relative to all predicted RNA folding conformations can then be visualized simultaneously.

We benchmark nearBynding using simulated data and replicate enhanced CLIP (eCLIP) experiments. We demonstrate nearBynding's utility by comparing our results to known RBP binding preferences, such as RBP binding to RNA structure, RBP binding relative to a second RBP, and RBP binding to RNA modifications. We also demonstrate how nearBynding could be applied to discern binding preference differences between a wild-type and mutant RBP. Besides *in silico* predictions, we employ diverse experimentally-derived data types (e.g., miCLIP-seq, rG4-seq) that are not utilized by currently available RBP motif-finding software in our correlation-based predictions of RBP binding preferences. We then use these discovered RBP binding preferences from our transcriptome-wide correlation analyses to hypothesize RBP characteristics that may predispose binding preferences for or against specific RNA structures. We show that aligned reads (continuous data format) provide qualitatively similar outputs and comparable reproducibility between technical replicates compared to peak-called data (interval data). The ability of nearBynding to correlate any interval or continuous feature annotated across selected regions of the transcriptome makes it a diverse, flexible tool to study RBP binding, RNA structure, RNA modification, and potentially other RNA features.

Results

nearBynding probes transcriptome-wide RBP binding to RNA structures

We use the nearBynding algorithm to incorporate transcriptome-wide information of RBP binding sites and RNA structure to discern local RNA structure for regions bound by an RBP as an example of its potential (Fig. 2.1A). To visualize the local binding context of an RBP, nearBynding requires only a list of transcripts, an annotated genome, and aligned CLIP-seq data as inputs. RNA structure data is an optional input and can be predicted within the pipeline. nearBynding produces a concatenated transcriptome made of only the transcriptome regions being probed and maps the data tracks (e.g. CLIP-seq and RNA structure) to it, which drastically reduces the magnitude of the datasets to only the intervals of interest. With this extension enabling transcriptome-scale analysis, the algorithm for efficient calculation of spatial correlations in the software StereoGene (15) is then applied to calculate the pairwise correlation between the two data tracks.

To enable analysis of RBP binding, the nearBynding software further includes visualization tools of the output statistics that are tailored to illustrate the relative positions of RBP binding and RNA structure. By default, nearBynding uses RNA structure probabilities predicted from sequence by CapR (40) for the selected transcriptomic intervals (see Methods). While CapR provides the default structural data input, the algorithm can also accept alternative inputs of custom RNA structure tracks or intervals, such as RNA modifications that affect RNA structure (e.g. N^6-methyladenosine [m6A]). StereoGene generates cross-correlation densities for RNA folding contexts relative to RBP binding. Since cross-correlation shows the relative position of one track (e.g., RBP binding) to another track (e.g., RNA structure), we can use it as a tool to visually represent the location of the RNA structure relative to the RBP binding site (i.e. upstream, at, or downstream of RBP binding). To account for the case in which a transcriptional track is correlated to *in silico* predictions of structure, the visualizations are performed

separately for RNA structures based on their categorization as double-stranded (stem) or one of five single-stranded types: hairpin, multibranch, internal, exterior, and bulge (Fig. 2.1B).

Binding profiles illustrating RNA structures at and proximal to RBP binding can be visualized either as line plots with standard errors for cases with multiple replicates, or as heatmaps (Fig. 2.1C). nearBynding, via StereoGene, calculates a null signal derived from the distribution of the correlation metrics for randomly shuffled windows (black line; Fig. 2.1C, lower panel). nearBynding also calculates a +/- one standard error confidence interval for the foreground signal when more than one experimental replicate is input for analysis; statistical significance can be assessed by comparing the distribution of the foreground signal computed from replicates to the null distribution computed from randomly shuffled windows, and users can alter the number of standard errors shown by the error bars to increase or decrease confidence intervals (e.g., three standard errors were shown in all figures presented here). This combination of visualization and statistics can be used to predict RBP binding to and adjacent to RNA structures transcriptome-wide. In addition to allowing visual assessment, nearBynding includes functions to quantitatively compare RBP binding cross-correlation distributions between two different RBPs. Specifically, the software computes the Wasserstein, or earth-mover, distance (41) between pairs of RBP binding profiles. For example, a short Wasserstein distance suggests similarity between two RBP profiles and possible binding competition between RBPs.

A transcriptional track for RBP binding information from CLIP-seq data can be input to the nearBynding software as a BAM file containing aligned reads or as a BED file containing peak intervals or protein–RNA cross-linked sites. Our transcriptome-wide correlation algorithm is applicable to a variety of CLIP-seq data types. Some processing methods estimate binding peaks or cross-linking sites that correct the CLIP-seq data for a size-matched input (42,43) or modeled background signal (44,45). When inputting raw CLIP data, the nearBynding software also allows for input of a size-matched input background track, the output of which is subtracted

from the output signal of the foreground track prior to computing the correlation statistics. We hypothesized that including a background track would ensure that the observed signal is from the RBP of interest rather than from experimental artifacts, such as cell-specific transcript levels or size-matched input noise. To test this hypothesis, we applied nearBynding to calculate and visualize the RNA structure preferences of the poly(C)-binding protein HNRNPK (46) and found that the binding profiles for HNRNPK in HepG2 and K562 cells were much more similar after background signal was removed (Wasserstein distance of 1.80 between non-corrected profiles versus 0.23 for background-corrected profiles) (Fig. 2.2).

Comparison of available spatial relationship algorithms

Multiple tools are available to evaluate the spatial relationships across genomic data types to draw conclusions about biological interactions (e.g. histone modifications and transcription start sites; CpG islands and gene promoters; splice sites and *Alu* elements). A common approach to assessing colocalization of features is to compare features at identical transcriptomic locations or within windows (overlapping coordinates, Fig. 2.3A), such as with the *plotCorrelation* function in deepTools (47). Biologically important relationships can also occur at proximal but non-overlapping transcriptomic sites (adjacent coordinates), which can be identified and quantified by tools such as the *reldist* function in BEDtools (16). StereoGene extended spatial relationship analysis to allow for correlation of continuous values, in addition to interval datatypes (Fig. 2.3B) (15). StereoGene is optimized for genomic annotation information and cannot selectively and efficiently analyze transcriptome data directly, so nearBynding expands its context of usability to transcriptomic data. Here, we provide an in-depth comparison of the spatial analysis capabilities of deepTools, BEDtools, and StereoGene and detail the ways in which nearBynding extends StereoGene to expand usability to transcriptomic analysis (summarized in Table 2.1).

We consider three main categories when comparing algorithms to correlate genomic relationships: the data types that can be analyzed, the types of correlation employed, and whether correlation results are visualizable. Different data types represent different information; for example, peak-called protein binding data is interval data, while aligned sequencing read data is continuous. Some data, such as G-quadruplex-forming intervals, is binary, whereas other data, such as methylation frequency, is non-binary. It is vital to choose a correlation algorithm that is able to accommodate the data type being studied. Some types of correlation, such as Spearman or Pearson correlation, can provide a single value for the overall correlation of two tracks; spatial correlation yields information about the overall relationship between two tracks as a distribution of correlations; cross-correlation shows the relative position of two tracks' features; and partial correlation can be used to remove a confounder that affects both the inputs being compared. No algorithm is able to compute all of these correlations, so users ought to first consider what correlation type can best answer the question being asked. Lastly, not all algorithms to correlate genomic relationships have the built-in capacity to graph results, which may be an important factor to weigh in the data analysis pipeline.

deepTools *plotCorrelation* computes the Spearman or Pearson correlation between two or more files based on scores within genomic regions (default is 10 kb bins). The scores' correlation can only be computed between bins with overlapping intervals, so deepTools is blind to relationships between adjacent coordinates. However, the interval data analyzed can incorporate amplitude information such as total read coverage within the assigned bins. deepTools is unable to account for confounders affecting the tracks being correlated, but it does have the option to plot scatterplots or heatmaps to visualize the calculated correlation between files (47). *plotCorrelation* uses the output matrices from either *multiBamSummary* or *multiBigwigSummary*, so input data can only be in BAM or BigWig formats.

Since the CapR-predicted RNA structure data is generated as a BED file, we were unable to test how deepTools may be used to observe RNA structure and RBP binding correlation. However, we were able to test pairwise correlations between RBP eCLIP datasets. We input aligned BAM files from the intron-binding proteins PRPF8, RBM22, U2AF2, and BUD13 to see whether deepTools could observe the proteins' known colocalization patterns (*c.f.* Fig. 2.12A, B, and C). We input both replicate's BAM files for each protein as well as the size-matched inputs from RBM22 and U2AF2 as controls and visualized the correlations via *plotCorrelation* using the default arguments (Fig. 2.4). This method shows strong clustering of technical replicates but does not have conspicuous colocalization signal between U2AF2 and BUD13, which are known to both bind 3' intronic termini, and between PRPF8 and RBM22, which are known to both bind 5' intronic termini (23). Further, there is moderate to strong correlation between size-matched input files and RBP binding files (0.37-0.91), suggesting that size-matched input information may confound the RBP binding data.

In addition to analyzing overlapping intervals, BEDtools *reldist* can calculate the correlation between non-overlapping features (i.e. features upstream and downstream of one another). Developed from GenometriCorr, BEDtools *reldist* allows users to identify the relative distance between two sets of intervals with consistent (i.e. non-random) spacing or proximity (16). Unfortunately, BEDtools can only incorporate binarized interval data from BED, GFF, or VCF files in its correlation analysis (Fig. 2.3C). And, like deepTools, it is unable to discern the effect of potential confounders of the correlation. There is no built-in visualization functionality.

To test whether BEDtools could be used to recapitulate known RNA structure binding preferences of RBPs (*c.f.* Fig. 2.10), we input CapR-predicted RNA structure data and peak-called RBP binding data. Since BEDtools can only incorporate binarized interval data, we denoted a nucleotide as positive for a structure context (i.e. stem, hairpin, multibranch, bulge, internal, or exterior) if CapR predicted it to adopt that structure with at least 0.2 probability. We

used CLIPper-called peaks for LIN28B and STAU2 eCLIP datasets (Fig. 2.5A, B, and C); PUS1 has only 35 reproducible CLIPper-called peaks, so it was omitted from the analysis. Graphing the *reldist* output shows strong correlation of RBP binding and stem structure within the closest colocalization bin for both proteins and weaker correlation for the other five structure contexts. The graphs for LIN28B in HepG2 and K562 look more similar to one another than to the STAU2 graph at relative distance < 0.05, but additional conclusions about similarities or differences in the proteins' binding preferences are difficult (Fig. 2.5A, B, and C). Since BEDtools' output is binned into intervals of 0.01 relative distance, there is poor resolution between RBP binding and RNA structure. BEDtools also cannot show regions of unfavorable binding (the correlation is always positive) or differentiate between upstream and downstream binding geometries.

The resolution limit of BEDtools is apparent when observing the binding of PRPF8, RBM22, U2AF2, and BUD13: PRPF8 and RBM22 appear to bind close to each other and U2AF2 and BUD13 bind close to each other as expected (Fig. 2.5D and E); however, there is minimal difference in the correlation of U2AF2 with BUD13 versus U2AF2 with PRPF8 or RBM22 (Fig. 2.5E), even though PRPF8 and RBM22 are known to generally bind further from U2AF2 than BUD13. Since BEDtools does not differentiate binding geometries, it is unable to show that PRPF8 and RBM22 have enriched binding ~200 nt upstream of BUD13 and U2AF2 binding.

Whereas BEDtools' correlation of non-overlapping features is limited to binary profiles, StereoGene was extended to allow for spatial correlation of continuous values (15). In addition to overall correlation of two tracks (i.e. the degree to which two features are correlated across the tracks), as can be provided by deepTools and BEDtools, StereoGene also calculates positional cross-correlation to provide information about the relative position of two tracks. The cross-correlation calculation is ultimately the most valuable component of the output with regards to studying the relative position of two features, since it provides information about the

geometry of the features' positions. The statistical significance of correlations with StereoGene is evaluated by a permutation-based test that compares the correlation between tracks of matched versus randomly shuffled windows. If a variable is expected to confound the correlation between the two tracks being tested, StereoGene can account for genome-wide confounders by partial correlation with another explanatory input track. StereoGene outputs an R script to visualize the cross-correlation and correlation distributions automatically.

nearBynding wraps StereoGene in biologically relevant methods and expands its context of usability from genome-scale to transcriptome-scale analysis. nearBynding alters the default StereoGene variables so that the size of the cross-correlation windows suits a transcriptomic rather than genomic scale (10 kb rather than 100 kb) and the correlation occurs in a single-nucleotide sliding frame for maximal resolution (rather than 100 nt bins). nearBynding provides additional functionality by allowing users to select regions of the transcriptome of interest; users may select specific transcripts and specific regions of those transcripts to interrogate. If the strand of the track information is available, this information is preserved for correlation analyses. The transition from genome-scale to transcriptome-scale analysis improves colocalization calculation specificity and efficiency so that correlation can be conducted at single-nucleotide resolution on a personal computer. nearBynding also determines RNA sequences and wraps CapR, an *in silico* RNA structure prediction software (48), to provide RNA structure information as an optional track for correlation if only one data track (e.g. only protein binding position) is available. StereoGene only outputs the mean value at every position in the cross-correlation window of shuffled replicates; nearBynding can calculate and output the standard error of the shuffled null track to allow for a statistical assessment of foreground versus null signal. The StereoGene visualization function can only depict the correlation information for one replicate at a time; nearBynding replaces this with options to depict the mean cross-correlation signal with error bars for multiple replicates simultaneously and removes experimental background signal if

an input track is also provided. Further, if the user chooses to study correlation relative to the CapR-derived RNA structure contexts, the cross-correlation signal for all six contexts can be depicted on the same graph.

Benchmarking on simulated data demonstrates robust signal detection

The nearBynding algorithm is unique because it can visualize overlapping and adjacent feature coordinates, incorporate diverse transcriptomic data types, and analyze minimally processed (e.g. non-peak-called) data. Because of the uniqueness of this algorithm, we were unable to directly compare it to other algorithms available for spatial correlation of genomic tracks. Instead, to evaluate its performance we designed simulated tracks, which we term RBP binding and RNA structure but could instead represent any two transcriptomic data types, to benchmark a full range of biological variables that may impact performance (Fig. 2.6). Briefly, we tested three factors that may impact signal strength: peak concordance between tracks, foreground to background ratio, and peak width range. Each simulation contained a pairwise analysis of the cross-correlation between an RNA structure track [RNA] and a CLIP-seq track [CLIP], where a greater amplitude for cross-correlation density reflects better co-occurrence of the two tracks. The peak distances and heights of the RNA structure track were varied to simulate the range of predicted RNA structure probabilities and random distribution of these structures across the transcriptome. The RNA structure track consisted of 10,000 peaks of simulated base-paired regions 31 to 500 nucleotides apart (unless otherwise stated), 5 nucleotides in width, and 0.02 to 1 unit in height. The CLIP-seq track simulated signal from aligned CLIP-seq data and contained a mixture of both background and foreground signal. Since the width of RBP binding peaks cannot be narrower than the length of CLIP-seq sequencing reads, the CLIP-seq track contained 30-unit-wide peaks (unless otherwise stated) to simulate the 30-nt reads of CLIP-seq data deposited in the ENCODE portal (49). The CLIP-seq

track was also shifted 12 units to the left of and equal in height to the RNA structure track peaks.

First, we tested the impact of the frequency of an RBP binding to its target RNA structure across the transcriptome, which may be affected by the accessibility of the RNA structure and the binding strength of the protein. To simulate this effect, we varied the frequency of the foreground signal peak concordance of the simulated CLIP-seq track relative to the simulated RNA structure track (Fig. 2.6A). We hypothesized that tracks with a higher frequency of RBP binding to target RNA structure would provide stronger binding signals than RBPs with sparser target binding. Supporting this hypothesis, the result of the nearBynding algorithm for the simulated data showed that cross-correlation signal strength correlates positively with peak concordance.

Next, we simulated artifacts associated with collecting CLIP-seq reads, such as background signal from input, by varying the height of the background signal of the simulated CLIP-seq track relative to the foreground signal (Fig. 2.6B). We hypothesized that simulations with a greater foreground (dark grey) to background signal (light grey) would have stronger RBP binding signals. As expected, cross-correlation signal strength correlated positively with the ratio of foreground to background. Both peak concordance and foreground to background ratio greatly affected signal strength, with nearBynding requiring a foreground to background signal greater than 0.05 to detect the binding signal (Fig. 2.6C). Therefore, our algorithm performance may be superior when applied to data collected by protocols that minimize noise (e.g., via additional washing steps) rather than protocols that document all binding events at the expense of greater noise.

We further employed our simulated data to test the sensitivity of nearBynding to the uniformity of peak width. Specifically, we increased the range of the simulated CLIP-seq peak widths to accommodate the possibility that RBPs may have variable binding footprints (Fig.

2.6D). Though the shape of the cross-correlation density track changed to reflect greater variation in peak widths, the amplitude and position of the signal maximum did not since the peaks were still centered at the same location. Therefore, we conclude that differences in peak width have no effect on signal amplitude.

nearBynding creates a concatenated transcriptome made of only the regions being interrogated as an input for StereoGene. However, transcript annotations may not be evenly distributed across the concatenated transcriptome. Therefore, we tested the dependence of nearBynding on the distribution of peaks along a concatenated transcriptome by shifting the locations of the simulated peaks such that they were all uniformly distributed or clustered near either end of the CLIP-seq track (Fig. 2.6E). Compared to peak concordance and foreground to background ratio, only a negligible loss in signal amplitude was observed for the most extremely skewed data (Fig. 2.6E). Overall, our results demonstrate that the order in which transcripts are concatenated, which could possibly affect the distribution of peaks, has negligible effect on binding signal relative to other variables tested.

Called peaks or aligned tracks for RBP binding produce similar binding profiles

Current practice for analyzing CLIP data is to call RBP-bound peaks using commonly used algorithms such as Piranha (44) and CLIPper (43), as well as recently developed algorithms such as CLAM (45), or omniCLIP (42). The ENCODE portal (49) has eCLIP datasets for 103 RBPs in HepG2 and 120 RBPs in K562, with each dataset containing two replicates and an input control. We selected 29 different RBPs in HepG2 and K562 cells that demonstrate strong, reproducible binding signals at 3'UTRs based on analysis from Van Nostrand et al., 2020b (Fig. 2.7A), which came to 40 unique cell type–RBP combinations. We used these high-confidence datasets to test whether nearBynding can produce comparable peak binding profiles from peak-callers and aligned reads. We collected eCLIP aligned reads and binding peaks called by two of the most commonly-used peak callers, Piranha and CLIPper, for these RBPs.

We ran Piranha on all replicates with parameters as described in the original publication (44). We also downloaded CLIPper-derived peaks of the eCLIP data from the ENCODE data portal (43). These three different inputs—aligned eCLIP tracks, Piranha peaks, and CLIPper peaks—were run through nearBynding, and binding was assessed for 3'UTR-annotated regions of the transcriptome. We used the Wasserstein distance (41) to determine the amplitude and distance required to transform one RBP binding profile into another. We calculated the sum of Wasserstein distances between the cross-correlation density tracks of all RNA structure contexts for all 40 unique cell type–RBP combinations across the three different input types each. Visualizing their distances in 2D on a multidimensional scaling plot (Fig. 2.7B) and comparing binding profiles across input types (Fig. 2.7C) showed only minor differences in the binding profile for RBPs based on the input source. Iterative sampling of binding profiles further indicated that the binding profiles for the three input sources of the same protein are more closely clustered compared to randomly chosen binding profiles from other proteins in the same cell line ($p = 5.59 \times 10\text{-}7$ in K562 cells and $p = 2.51 \times 10\text{-}10$ in HepG2 cells, Kolmogorov-Smirnov test, Fig. 2.7D). Therefore, the difference between profiles for different RBPs is greater than the difference within the same experiment queried via different inputs.

We next tested whether technical replicate reproducibility is comparable between peak-called and aligned read inputs by calculating the Wasserstein distance between replicates for the same 40 unique cell type–RBP combinations using the three different input track types. The distance between aligned read tracks for technical replicates was smallest in 12 of 40 cases (30%); Piranha peak-called replicate tracks were closest in 23 of 40 cases and CLIPper peak-called replicate tracks were closest in 5 of 40 cases (Fig. 2.7E). Further, the distances between technical replicates for peak-called and aligned read tracks are quantitatively similar for the majority of cell type–RBP combinations tested (Fig. 2.7F). These data cumulatively suggest that

aligned reads generate similar outputs for technical replicates via nearBynding when compared to peak-called tracks.

Cross-correlation tracks reproducibly cluster RBP data across biological replicates

The context-dependence of RNA binding can be expected to lead to variable signal concordance in binding predictions from CLIP-seq data with the same RBP. Replicates from the same cell type would likely manifest technical differences whereas analyses of the same RBP across different cell types may depict biological differences in RBP binding. Analyses of the same RBP between replicates within the same cell type can be expected to have greater concordance than analyses from different cell types. We sought to assess the fidelity of nearBynding's ability to reproducibly identify such RBP binding context by clustering RBP binding profiles.

Within the ENCODE datasets, 73 RBPs are common across both cell lines. Genome-wide RNA structure profiling showed that 3' untranslated regions (UTR), which are targets for many RBPs, are generally highly structured in cells (50,51). Therefore, in order to test our algorithm on a robust dataset, we restricted our analysis of RBP binding to 3'UTR regions. We collected isoform information of all 3'UTRs expressed in HepG2 and K562 using RNA-seq data from ENCODE (49). We generated cell type-specific binding profiles by selecting eCLIP reads that aligned to isoforms expressed in the corresponding eCLIP cell type. The most abundant 3'UTR isoforms for the expressed transcripts were then submitted to nearBynding to determine RBP binding preferences for these regions.

First, we wanted to test how well biological replicates of the same RBP in the same cell type clustered. We calculated the sum of Wasserstein distances between the cross-correlation density tracks of all RNA structure contexts for every sample within each cell type. 71 of 206 replicates in HepG2 (34%, p = 9.4 x 10-109 for a one-tailed binomial test assuming random

chance) and 115 of 240 replicates in K562 (48%, p = 1.3 x 10-203, one-tailed binomial test) most closely clustered in pairs with their corresponding biological replicate (Fig. 2.8A). For those RBPs with a large proportion of their binding events in 3'UTRs, they tend to cluster more closely based on Wasserstein distance than other RBPs (Fig. 2.8B), which is reasonable since the RBP binding profiles were generated from 3'UTR regions of the transcriptome (21). Among these 3'UTR-binding RBPs, 33 of 42 replicates in HepG2 (79%, p = 2.2 x 10-45, one-tailed binomial test) and 37 of 44 replicates in K562 (84%, p = 1.2 x 10-53, one-tailed binomial test) most closely clustered in pairs with their corresponding biological replicate, and 93% and 91% of replicates were clustered within the top two distances in K562 and HepG2, respectively (Fig. 2.8C). These analyses demonstrate that biological replicates largely cluster together under Wasserstein distance.

Next, we interrogated the reproducibility of RBP binding profiles across cell lines. The cross-correlation densities of biological replicates for each RBP were averaged, and these averaged values were used to calculate the Wasserstein distances for all RNA structural contexts. For every RBP in K562 cells, we ranked how similar its binding profile was to RBPs in HepG2 cells. 15 of 73 RBPs (21%) clustered closest with their counterparts in the other cell line, and 21 of 73 RBP counterparts (29%) were within the top three closest distances in the other cell line (Fig. 2.8D and E). Interestingly, there was no difference in clustering distances across cell lines between 3'UTR-binding RBPs and all other proteins (Fig. 2.9A). The inverse comparison—the distance of HepG2 RBPs against all K562 RBPs—also had 29% of RBPs cluster within the top three distances of their counterparts (Fig. 2.9B), suggesting poorer concordance across cell types than between biological replicates in the same cell line.

RNA structure cross-correlation signal is specific

We wanted to analyze negative control datasets to test nearBynding's specificity. Chromatin immunoprecipitation (ChIP) sequencing data depicts where proteins bind DNA and

can be mapped to the transcriptome, but the protein-bound DNA is expected to have virtually no association with the RNA structure transcribed from the bound genomic regions, making it a good negative control for RNA structure binding preferences. We selected 3'UTR-binding proteins for which there is eCLIP and ChIP data for the same cell line in the ENCODE portal (49). The eCLIP data were predicted to have significant preference for one or multiple RNA structure contexts while the ChIP data was expected to have no significant binding preference signal. All tested ChIP datasets did not reach statistical significance based on a +/- three standard error confidence interval, whereas the eCLIP datasets for the same proteins demonstrated statistically significant RNA structure binding preferences (Fig. 2.9C, D, and E). These negative control results suggest that nearBynding's detection of RNA structure binding preferences is specific.

RBP binding profiles recapitulate known structural preferences

Next, we tested whether the binding profiles generated by nearBynding reflect known RBP structural binding preferences. We selected four diverse RBPs to represent RBPs with a well-characterized sequence preference (PUM2), enzymatic activity (PUS1), multiple types of RNA-binding domains (LIN28B), and a well-characterized structure preference (STAU2).

First, we selected PUM2 for analysis because its modular structure of eight tandem repeats is known to recognize RNA in a sequence-specific manner (52). Although PUM2 preferentially binds 3'UTRs in a sequence-specific manner, there is evidence that PUM2 also has a structural component to its binding preferences: in vitro analysis shows that PUM2 dissociates from double-stranded regions faster than single-stranded regions and that it stably binds regions flanked by stem structures (39). The PUM2 binding profile (Fig. 2.1C) showed that PUM2 has minimal structure preference at the point of binding (distance = 0), but it does prefer stem context upstream and downstream of its point of contact.

We then studied PUS1, an enzyme that adds a pseudouridine modification to target RNAs (19). PUS1 has a weak trinucleotide binding sequence motif and modifies nucleotides at the 5' end of stem loops flanked by single-stranded runs for the vast majority of its high-confidence targets (19). Consistent with PUS1 binding and modifying the 5' base of stems, its binding profile showed a preference for single-stranded regions at the end of the transcript (exterior context) upstream and double-stranded (stem) context downstream of PUS1 binding (Fig. 2.10A).

We next selected LIN28B because it has two RNA-binding domains: a cold shock domain (CSD) and tandem zinc-binding motifs (ZFs). Although LIN28 has a preference for binding GGAGA motifs, target motifs are generally single-stranded (28). NMR spectroscopy suggests that although LIN28 binds stem-rich regions, the CSD binds hairpins and the ZFs bind bulges containing the sequence motif associated with the stem (53). These same results were apparent in the binding profile, which showed enrichment for stem, bulge, and hairpin contexts at or proximal to the LIN28 binding site (Fig. 2.10B).

STAU2 binds stretches of base-paired sequences of variable lengths via its three double-stranded RNA-binding domains (dsRBDs) (54). Although the dsRBDs bind tightest to perfectly complementary stem structures, they are able to bind stems that contain bulges (54). Consistent with expectations, the binding profile of STAU2 was strongly enriched for stem context, had slight enrichment for bulge context, and was generally depleted for single-stranded contexts such as hairpin, multibranch, and exterior (Fig. 2.10C).

There is a range of nucleotide resolution derived from the various CLIP-seq techniques. For example, eCLIP provides 30-nucleotide reads surrounding the protein–RNA cross-linking site, whereas better resolution can be achieved with techniques such as individual-nucleotide resolution CLIP (iCLIP) and RNA hybrid iCLIP (hiCLIP) that are able to identify the protein–RNA cross-link site with single-nucleotide resolution. The resolution of nearBynding's profiles reflects

the resolution of the input data. For example, by using hiCLIP cross-link sites of STAU1 (55), which binds dsRNA similar to STAU2, nearBynding was able to demonstrate that STAU1 contacts ssRNA—preferably hairpin context but there was enrichment for all ssRNA contexts at the binding point—but that this ssRNA was directly 3' of the double-stranded stem context (Fig. 2.10D). Consistent with our nearBynding analyses, the authors of the hiCLIP data hypothesized that cross-linking sites were enriched at ssRNA because bases within the duplexes are inaccessible for protein–RNA cross-linking (55). Further, the cross-linking site was often 3' of a stem-hairpin-stem structure. Overall, although there are only a few experimentally-confirmed RNA structure binding preferences for us to use as true positives, nearBynding-generated RBP binding profiles effectively recapitulate documented preferences for RNA structures.

Differentiating wild-type and mutant RBP structural preferences

Besides investigating wild-type (WT) protein binding relative to null signal, nearBynding can be applied to researching mutant RBPs by comparing WT and mutant protein binding. Whereas a comparison of WT versus input control depicts the full complement of RBP binding across the transcriptome, a comparison of WT versus a mutant allows visualization of the function-dependent binding of an RBP. For example, binding data is available for the processive RNA helicase UPF1, which is involved in many RNA decay pathways (56), as well as for two helicase-dead UPF1 mutants, K498A and DEAA, which are deficient in ATP binding and hydrolysis, respectively (57). Both helicase-dead UPF1 mutants retain the ability to bind RNA, but they exhibit a complete loss in target discrimination (57). The WT-only UPF1 binding profile and the profiles corrected for helicase-dead mutant signals are all highly symmetrical (i.e., similar RNA structure binding preferences upstream and downstream of binding), but the corrected profiles indicate a broader span of structure signals (Fig. 2.10E, F, and G). The mutant-corrected profiles suggest that UPF1 requires helicase activity to occupy stem contexts and select against the unstructured multibranch and exterior contexts. The symmetry of protein

binding profiles within 3'UTRs correlates with the strength of a protein's preference for stem context

Probe protein binding profile symmetry within 3'UTRs

Symmetrical RNA binding signal within 3'UTRs that is strongest at the position of maximal protein-binding is apparent for some proteins (e.g., UPF1; Fig. 2.11A), whereas other proteins seem to have asymmetrical RNA binding preferences (e.g., ZNF800; Fig. 2.11B). In order to quantify the asymmetry of protein binding, we created functions in nearBynding, *symmetryCapR* and *symmetryContext*, that calculate the Wasserstein distance between the binding profile upstream of protein binding (negative values to position 0) and the inverted binding profile downstream of protein binding (position 0 and positive values). Each of the RNA structure contexts' amplitudes are normalized to the maximum value within each track within *symmetryCapR* so that high-amplitude contexts do not disproportionally contribute to the symmetry signal. All six contexts' normalized Wasserstein distances are summed to provide a single symmetry metric where a value of 0 represents perfect symmetry and increasingly large values represent increasing asymmetry. For example, ZNF800 binding in 3'UTRs is highly asymmetrical (symmetry = 2.4, Fig. 2.11A), whereas UPF1 binding is symmetrical (symmetry = 0.4, Fig. 2.11B).

We wanted to determine whether the RNA contexts a protein prefers to bind has any influence on the protein's binding symmetry, so we calculated the binding symmetry of every protein binding profile within 3'UTRs (Fig. 2.11C). We next found the greatest absolute value amplitude across all positions for the six RNA structure contexts, termed maximal correlation signal (e.g. in UPF1 stem maximal amplitude is 3.3, hairpin is 0.5, and multibranch is 1.0 per Fig. 2.11B). We correlated symmetry and maximal correlation signal amplitude and found that signal amplitude for bulge and stem binding within 3'UTRs correlates negatively with the binding profile symmetry (Fig. 2.11D); that is, there is a weak association between strong protein

binding preference for the stem and/or bulge contexts and symmetrical protein binding profiles. However, since bulge context maximal amplitude correlates with that of stem (R^2 = 0.45, Fig. 2.11E), we conducted a multivariate regression with all RNA structure contexts and found that only stem context explains protein binding symmetry (Table 2.2).

Illustrate relative binding positions of RBPs

nearBynding can also use RBP binding as a track against another RBP binding track, allowing the user to assess binding preference of one RBP relative to another. To exemplify this functionality, we chose proteins known to occupy the 5' and 3' ends of introns and used nearBynding to observe the position and density of their preferred binding relative to each other across unspliced transcripts in K562 cells (Fig. 2.12A). We studied four proteins important for pre-mRNA splicing via their roles in the spliceosome: PRPF8 and RBM22, both of which bind 5' intronic termini, and BUD13 and U2AF2, which bind 3' intronic termini (23). As expected, PRPF8 and RBM22 colocalized, and BUD13 and U2AF2 also colocalized (Fig. 2.12B and C). PRPF8 and RBM22 have comparatively weak, broad signals roughly 100-300 nucleotides upstream of U2AF2 (Fig. 2.12C), which corresponds to the ~10^2 bp intron length of the significant minimal intron peak common in mammalian genomes (58). Though these tests only reproduce known binding geometry, additional pairwise analyses of RBP binding using nearBynding could provide deeper insights into the arrangements of proteins relative to one another across the transcriptome.

Inform RBP binding preferences using experimentally-derived RNA annotations

Analyzing RBP binding to G-quadruplexes

nearBynding is not restricted to *in silico* RNA structure prediction input, so we next interrogated RBP binding profiles with experimentally-derived RNA structure data. Guanine-rich RNA sequences can interact via Hoogsteen base-pairing and fold into non-canonical structural

motifs called G-quadruplexes (G4s) (59). Although many tools are available to predict putative G4s, they are prone to false-positives, since G4 folding is often dependent on the wider context of the RNA sequence and RBP regulation (51,60). We therefore used rG4-seq data (61) to map G4s that form in cells. Although the rG4-seq data was collected from HEK293 cells and ENCODE provided RBP binding data from HepG2 and K562 cells, we reasoned that these cell lines would have enough G4s in common that we could observe general G4 binding trends. Indeed, we observed strong RBP binding at G4s across the exome for multiple published G4-binding proteins such as NONO, FUS, GRSF1, DROSHA, and DDX42 (Fig. 2.13A and B) (62–66). Additionally, many of the RBPs that exhibited the strongest G4-binding signal—PRPF4, GTF2F1, FAM120A, CSTF2T, and DDX6—have recently been shown to bind at putative G4 sites in mRNA UTRs (67). However, some published G4-binding proteins such as FMR1 did not exhibit a robust signal, perhaps due to cell-type-specific variations in binding (Fig. 2.13A and B and Table 2.3). Our analysis also identified RBPs such as YBX3, PRPF8, ZNF800, PPIG, and NOLC1 that are depleted for G4s at their binding sites in HepG2 and K562. However, these proteins have not previously been documented for their preference against G4 binding, which warrants future investigation.

Since many of the tested RBPs have been shown to occupy intronic regions of the transcriptome and intronic G4s have been implicated in RBP regulation (68), we tested whether binding profiles for RBPs differed between exonic and unspliced transcriptomic contexts (Fig. 2.14A). We used the maximum RBP binding to G4 density signal—positive or negative—from every RBP's binding profile as a metric for G4 binding preference and found a strong correlation ($R^2 = 0.81$) between exonic and unspliced transcripts, with high-signal RBPs (|maximum signal density| > 2.5 x 100) correlating better than RBPs with weak preference for G4s (|maximum signal density| < 2.5 x 100, Fig. 2.14B).

We compared protein binding signals between exonic and unspliced transcriptomic analyses for intron-binding versus all other proteins and found that predominantly intron-binding proteins have a greater residual than other RBPs when comparing exonic and unspliced transcriptomic analyses ($p < 0.001$, one-tailed t-test, Fig. 2.14C) (21). This demonstrates that for RBPs that do not bind introns, there is only minor differences between the maximal G4-binding signal calculated by nearBynding using exonic versus using unspliced regions of the transcriptome. In contrast, there is a greater difference between the maximal G4-binding signal calculated using exonic versus using unspliced regions of the transcriptome for intron-binding RBPs. This is reasonable, since intron-binding RBPs would be expected to show greater shifts in their binding profile upon the inclusion or exclusion of intronic regions for the correlation calculations (*c.f.* Fig. 2.8B: 3'UTR-binding protein profiles are more replicable than non-3'UTR-binding profiles when only observing binding within 3'UTRs).

Next, we aimed to determine whether the position of the maximal G4 binding preference signal shifted when comparing exonic versus unspliced transcriptomic regions for high-signal RBPs, since the maximal signal may remain the same even if the distance of maximal signal from protein binding changes. All high-signal RBPs had a maximal signal downstream of RBP binding, and the distance of maximal signal from RBP binding was extremely similar for both intronic ($R^2 = 0.94$) and non-intronic ($R^2 = 0.97$) RBPs between exonic and unspliced binding profiles (Fig. 2.14D). This suggests that even if the strength of the maximal signal varies between nearBynding analyses of exonic and unspliced regions of the transcriptome, the relative geometry of the RBPs and the G4s does not vary.

These data cumulatively suggest that even among RBPs that primarily bind intronic regions of the transcriptome, overall G4 binding profiles are consistent when analyzing RBP binding in exonic and unspliced transcriptomic regions for RBPs that have moderate to strong G4 binding preference. However, intron-binding proteins are most affected by whether or not

introns were included in the analysis, demonstrating that it is likely best, though not necessary, to include preferred binding regions of the transcriptome in the analysis.

We hypothesized that our protein-level data can help identify domains that play a role in G4 binding preference. We pooled the exonic HepG2 and K562 data and used the Pfam database (69) and protein sequence information to identify protein domains present within the RBPs. Across 13 common protein domains identified, most did not affect G4 binding (Table 2.4, Fig. 2.14E). RGG repeats are the most common motif in G4-binding RBPs (e.g., FUS) (70) and, based on our analysis, RBPs with RG-rich domains did demonstrate increased G4 binding. Proteins that contain SAP, dsRBD, or G-patch domains also had increased G4 binding, though there is no literature evidence of this preference. In contrast, RBPs that contain one or more armadillo domains had significantly decreased G4 binding, with 6 out of 8 armadillo-containing proteins demonstrating G4 depletion in their binding preference.

Analyzing RBP preferences for RNA modification

Lastly, m6A modification is an abundant RNA modification that affects RNA structure (71). Since m6A modifications affect RNA folding but are not considered in currently-available *in silico* folding algorithms, we tested whether m6A-iCLIP (miCLIP-seq) data could be used as an input for nearBynding to observe protein binding contexts relative to m6A modification. Multiple RBPs are involved in the writing, reading, and erasing of m6A, such as RBM15, IGF2BP1/3, and FTO, respectively (72–74). We used miCLIP-seq data (75) and eCLIP data (49) from HepG2 to determine whether these m6A-interacting RBPs show binding preferences relative to m6A modification. As expected, RBM15, IGF2BP1, and IGF2BP3 all demonstrated a preference for binding m6A-modified RNA (Fig. 2.15A). In contrast, FTO did not seem to occupy m6A-modified regions of the transcriptome, perhaps reflecting its role as an m6A eraser. Despite their modest density amplitudes, likely due to a small signal to noise ratio in the miCLIP-seq data, these signals are significant given a +/- three standard error confidence interval above the

shuffled null track (Fig. 2.15B). Additional miCLIP-seq replicates or reproduction of the study in other cell types would be the best way to discern whether this signal is biologically meaningful. In sum, these reproductions of previously-observed results demonstrate the diversity of data types that can accurately inform RBP binding contexts using nearBynding.

Discussion

The nearBynding algorithm, which is also available as a pipeline in the R/Bioconductor package of the same name, provides a new tool to discern colocalization of transcriptomic data features within regions of the transcriptome defined by the user. As an example, we illustrated using nearBynding to identify RNA structure and modification at and proximal to the site of protein binding. Protein-binding data can be input as either aligned CLIP reads or peak-called files, and RNA structure can be input based on in silico sequence-based prediction or derived from experiments. Cross-correlation profiles can be visually and quantitatively compared across multiple formats. Using simulation datasets, we verified that nearBynding is capable of discerning signal even with a foreground to background ratio of 0.05 and peak concordance less than 40%. Using extensive ENCODE eCLIP datasets, we established that nearBynding reproduces binding profiles between biological replicates and across cell lines. In addition, these binding profiles recapitulated known preferences for RNA structure and modification as well as identified protein domains that may bind G4.

Our analyses revealed that the concordance of RBP binding profiles across cell lines were lower than biological replicates of the same cell lines. This difference may be because (i) the eCLIP data is cleaner for one cell type (as supported by the fact that there is poor agreement between some technical replicates), (ii) the transcripts expressed and therefore available for binding differ, (iii) differential expression of competitive binders, modifiers, or cofactors, or (iv) the RNA folds differently between cell types. Prior analysis of these datasets has also uncovered cell type-specific differences in preferred transcript region binding for some

of the RBPs interrogated (Van Nostrand et al., 2020a). There are many examples of RBPs that shift their binding preferences as a result of post-translational modifications or cofactor binding (Schmidt et al., 2017; Timchenko et al., 2006). For some RBPs, their cellular localization and therefore binding opportunities rely on cofactor binding (Heininger et al., 2016), whereas others' binding opportunities depend on the expression of competitive binders (Liu et al., 2015). Variations in protein expression and post-translational modification frequencies across cell lines may therefore drive differences in protein binding profiles. Molecular validation would be required to prove cell-specific binding preferences for an RBP; in the absence of this proof, we suggest using signal similarities across replicates and cell types to bolster confidence in predictions of structure binding preference for a given RBP but hesitate to interpret the differences since they may not be biologically meaningful. Protein binding profile intervals where the context signal and null track error bars do not overlap are statistically significant, but there may be circumstances where the results are statistically but not biologically significant.

A statistically significant cross-correlation signal between RBP binding and RNA structures implies that an RBP binds that specific structure, but there could be an alternative explanation. It is possible that these RBP-associated sequences are prone to adopt a particular RNA structure only when it is not bound by the RBP. DROSHA, for example, binds G4-forming regions only when these regions are unfolded (Rouleau et al., 2017). Because many G4-forming sequences are actively unfolded in vivo, we cannot know without further molecular experimentation whether an RBP binds to G4s or RBP-associated sequences are prone to forming G4s. We speculate that a phenomenon similar to DROSHA's binding drove the enrichment of dsRBD-containing RBPs among the higher G4 signals (Table 2.4), since G4-forming sequences are necessarily GC-rich and likely form stable regions of dsRNA. Biochemical experimentation such as kinetics assays or crystal structures are necessary to definitively ascertain RBP binding.

nearBynding may use non-interval data, such as the amplitudes of aligned CLIP-seq reads, to quantify RBP affinities for different segments of the transcriptome, and therefore does not require peak-calling in RBP binding data analysis. It has been previously observed that modeling RBP binding as a list of bound regions across the transcriptome provides only a coarse approximation of RBP binding motifs (Maticzka et al., 2014), and this nuanced read amplitude information enables us to identify preferred RNA structure motifs based on RBP binding frequency. Some differences may exist in RBP binding profiles when CLIPper-called peaks, Piranha-called peaks, or aligned eCLIP reads were served as the input for the same experiment. However, the differences between binding profiles for different RBPs is far greater than the differences for the same RBP interrogated using different inputs, demonstrating the similarity in results between interval and continuous datatypes and allowing for the possibility of omitting the step of peak calling for RBP binding analysis.

Most state-of-the-art algorithms that incorporate RNA structure into predictions of RBP binding motifs rely on RNA sequence alone to predict RNA secondary structure (Guo and Bartel, 2016). Similarly, nearBynding can call CapR to predict RNA structures. All these algorithms, however, assume that the mRNA being folded is naked and unmodified, with only the queried RBP binding it. The recently-developed deep learning tool, PrismNet, works to address these limitations by integrating cell type-matched in vivo icSHAPE and CLIP data to predict RBP binding across cell conditions (Sun et al., 2021). nearBynding provides the flexibility for users to input an even broader range of experimentally-derived RNA structure information, which could be used to study the binding of non-canonical RNA structures (e.g., G4s, triple helices) and RNA modifications (e.g. A-to-I editing, m6A, or N4-acetylcytidine). In addition, users can improve the study of canonical RNA structure binding by incorporating structural information collected via chemical probing (e.g., SHAPE or dimethyl sulfate [DMS]). Future work

will characterize the impact of chemical probing-informed RNA structure data on RBP profiles relative to in silico-derived RNA structure.

nearBynding allows for the comparisons of various interval or continuous features across user-selected regions of the transcriptome and may be adapted to correlate many other transcriptome-related features beyond the ones exhibited in this work. Users can study the binding of an RBP relative to RNA structure, one RBP relative to another, an RNA modification relative to RNA structure, or any pair of interval or continuous features so long as they can be annotated on a transcriptome. Further, mutant and wild-type data can be directly compared to understand how genetic changes affect elements such as RBP binding preferences or RNA modifications. Future work will take advantage of this flexibility to characterize RBP complexes relative to RNA structure via wider-ranging approaches than previously possible.

Methods

Resource Availability

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Anthony Leung (anthony.leung@jhu.edu). This study did not generate new unique reagents (Table 2.5).

All orginical code for nearBynding is available at Bioconductor (https://bioconductor.org/packages/nearBynding/, BioC 3.13). nearBynding v1.3.1 was used for all data analysis and the latest updates are available in the GitHub repository (https://github.com/vbusa1/nearBynding). nearBynding is entirely coded in R (v4.0) and its license is Artistic-2.0. nearBynding was run on macOS Catalina (10.15.4) but has been effectively run on Windows and Linux OS as well. Some processed data and the code necessary to generate the simulated data and figures presented in this paper are available in the GitHub repository (https://github.com/vbusa1/nearBynding_manuscript). CapR (40) (v1.1.1)

is available at https://github.com/fukunagatsu/CapR , StereoGene (15) (v2.20) is available at http://stereogene.bioinf.fbb.msu.ru/ , and BEDtools (76) (v2.28.0) is available at https://bedtools.readthedocs.io/en/latest/.

All eCLIP BAM files and CLIPper peak files are available in the ENCODE repository (v102 https://www.encodeproject.org/) (49). STAU1 hiCLIP data is available in the iMaps repository (https://imaps.genialis.com/iclip/search/collection/hi-clip-reveals-m-rna-secondary-structures) (55). UPF1 WT, K498A, and DEAA CLIP-seq data is from Lee et al., 2015 (accessible through the Gene Expression Omnibus (GEO) accession number: GSE69586). Peak calling was conducted using Piranha (44) v1.2.1 (http://smithlabresearch.org/software/piranha/) as described in the original paper: bin size of 36 nucleotides and the single covariate being the log of the read counts from the control input. We used FASTA and GTF files from the Ensembl release 100 Homo sapiens GRCh38 genome (ftp://ftp.ensembl.org/pub/release-100/) (77). HepG2 and K562 isoform information was derived from RNA-seq data available in the ENCODE repository (identifiers ENCSR181ZGR and ENCSR885DVH, respectively); HEK293 isoform information was derived from WT RNA-seq data (78) (GEO: GSE122425). In the case of multiple expressed isoforms, we chose the highest-expressed isoform based on the RNA-seq data to prevent overlapping data mapping. Single-nucleotide resolution profiling of m6A (miCLIP-seq) in HepG2 cells is from Huang et al., 2019 (GEO: GSE121942). rG4-seq data in HEK293 cells is from Kwok et al., 2016 (GEO: GSE77282). Both of these datasets were aligned to hg19 and so were lifted over to hg38 using a UCSC chain file before input into nearBynding (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz) (79). The protein domain data used to compare G4 binding preferences is available in the Pfam repository (69) (v33.1 http://pfam.xfam.org/). RGG-repeat-containing proteins are listed in the supplementary data of ref. Ozdilek et al., 2017.

Workflow details

nearBynding inputs

In order to provide predicted RNA structure context for RBP binding, nearBynding requires the following pieces of input data: (1) CLIP-seq alignment tracks for the RBP of interest, (2) an annotated genome and associated FASTA sequence, and (3) a list of transcripts of interest. It is recommended that all transcripts selected are expressed in the cell type used for the CLIP-seq experiment. Alternative RNA structure information can optionally be included, and it is recommended that the data is derived from the same cell type. All data must be converted to bedGraph format, and the functions *CleanBEDtoBG* and *CleanBAMtoBG* help users do this.

Map data to pseudochromosomes

Users must first choose which regions of the transcripts of interest to interrogate (e.g. UTRs, exons, whole transcript), based on annotations available in the genome GTF file. nearBynding creates (1) a chain file that will be used to map the selected regions of transcripts end-to-end, excluding the intergenic regions and undesired transcripts that comprise the majority of the genome, via the function *GenomeMappingtoChainFile* and (2) a file detailing the names and sizes of all the chain file's pseudochromosomes via the function *getChainChrSize*. The chain file can then be used to transfer genome references of the CLIP-seq data from the whole-genome alignment to the transcriptome alignment of interest via *liftOverToExomicBG*. If external RNA structure data is being studied, its genome references would need to be transferred to the transcriptome alignment as well using the same chain file. Chain files cannot accommodate overlapping intervals since they cause ambiguous regions in the transfer process, so it is recommended that users supply the highest-expressed isoform of every gene expressed in the cell type of the CLIP-seq data to create the concatenated pseudochromosomes.

CapR RNA structure prediction

nearBynding pulls the sequences of selected regions of transcripts of interest from the genome FASTA file based on genome annotations using BEDtools (76) wrapped in the function *ExtractTranscriptomeSequence*. Probabilistic RNA structure for the selected regions are derived from *in silico* folding predictions by CapR, which includes RNAfold software in its structure predictions (40), via the function *runCapR*. Each nucleotide is scored as having a 0 to 1 probability of adopting one of six different contexts by CapR. The data for the six different folding conformations are then separated and the transcript fragments are concatenated into pseudo-chromosomes via *processCapRout*. In secondary structure representation, RNA bases are depicted as vertices of polygons with edges of RNA backbone or hydrogen bonds (Fig. 2.1B). The six different RNA structure contexts are defined thus: stem context is if a base participates in hydrogen-bonding with another base; exterior context is if a base does not form a polygon such as that the end of a transcript; hairpin context is if a single-stranded base is involved in a polygon with only one hydrogen-bonding edge; bulge context is if a single-stranded base is involved in a polygon with two hydrogen-bonding edges and where all stem context vertices are contiguous in the polygon; internal context is if a single-stranded base is involved in a polygon with two hydrogen-bonding edges and where stem context vertices are not contiguous in the polygon; multibranch context is if a single-stranded base is involved in a polygon with at least three hydrogen-bonding edges (40).

Relative binding position calculation

To visualize the RNA structure landscape surrounding protein binding, StereoGene (15), wrapped within the functions *runStereogene* and *runStereogeneOnCapR*, is used to calculate the cross-correlation between RNA structure and protein binding. The *get_error* argument for *runStereogene* and *runStereogeneOnCapR* allows users to also get the standard error for the shuffled null track, and the *nShuffle* argument determines how many times the null track is

shuffled; calculating the shuffled null standard error is optional because it requires substantially more computational time, especially in cases where many shuffling iterations are requested. nearBynding analyzes local structure in single-nucleotide frames, which sacrifices some of the computational efficiency of StereoGene but maximizes RBP binding resolution. Cross-correlation densities are within the range -1 to +1, where -1 suggests perfect depletion of an RBP for a tested RNA structure context, 0 represents no binding preference, and +1 suggests perfect RBP binding for a tested RNA structure context. Since actual experimental correlation densities are far more modest, they are reported as density x 100 for visualization, which is conducted using the functions *visualizeStereogene* and *visualizeCapRStereogene*.

nearBynding output analyses

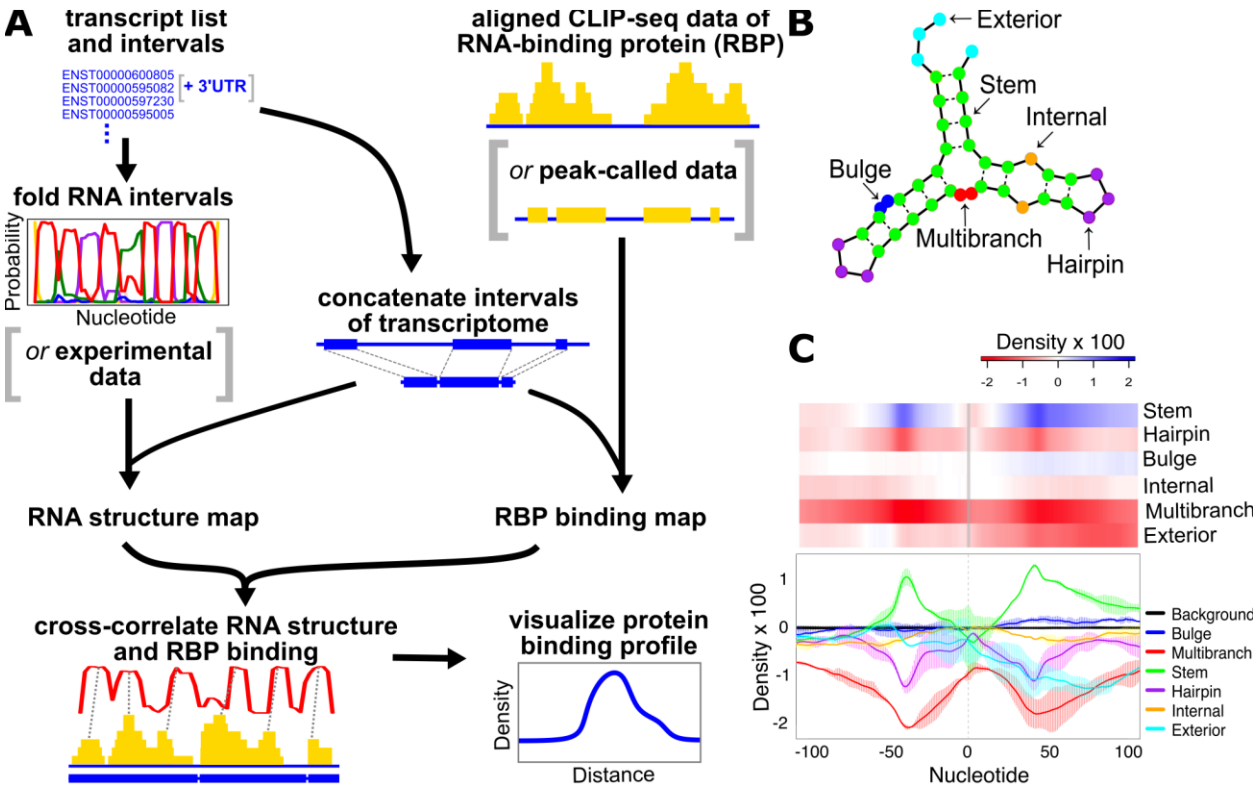
nearBynding allows users to calculate the similarity between output binding profiles via Wasserstein distance via *bindingContextDistance* and *bindingContextDistanceCapR*, where small values indicate greater similarity. Users can also assess the grouping of different categories of points via bootstrapping and the Kolmogorov–Smirnov test (Fig. 2.7D) using *assessGrouping*.

Quantification and Statistical Analysis

nearBynding tracks were considered significant if +/- three standard error intervals of foreground signal computed from technical replicates did not overlap with +/- three standard error intervals of the shuffled null distribution (1,000 null iterations). Functions from the stats package were used to calculate binomial tests (*binom.test*), t-tests (*t.test*), and Kolmogorov-Smirnov tests (*ks.test*); *ks.test* is wrapped in nearBynding's *assessGrouping* function. The lsr package (81) was used to calculate Cohen's *d* (*cohensD*). The transport R package function *wasserstein1d* (41), wrapped in nearBynding's *bindingContextDistance* and *bindingContextDistanceCapR* functions, was used to calculate Wasserstein distances.

Figures and tables

Figure 2.1 Overview of nearBynding

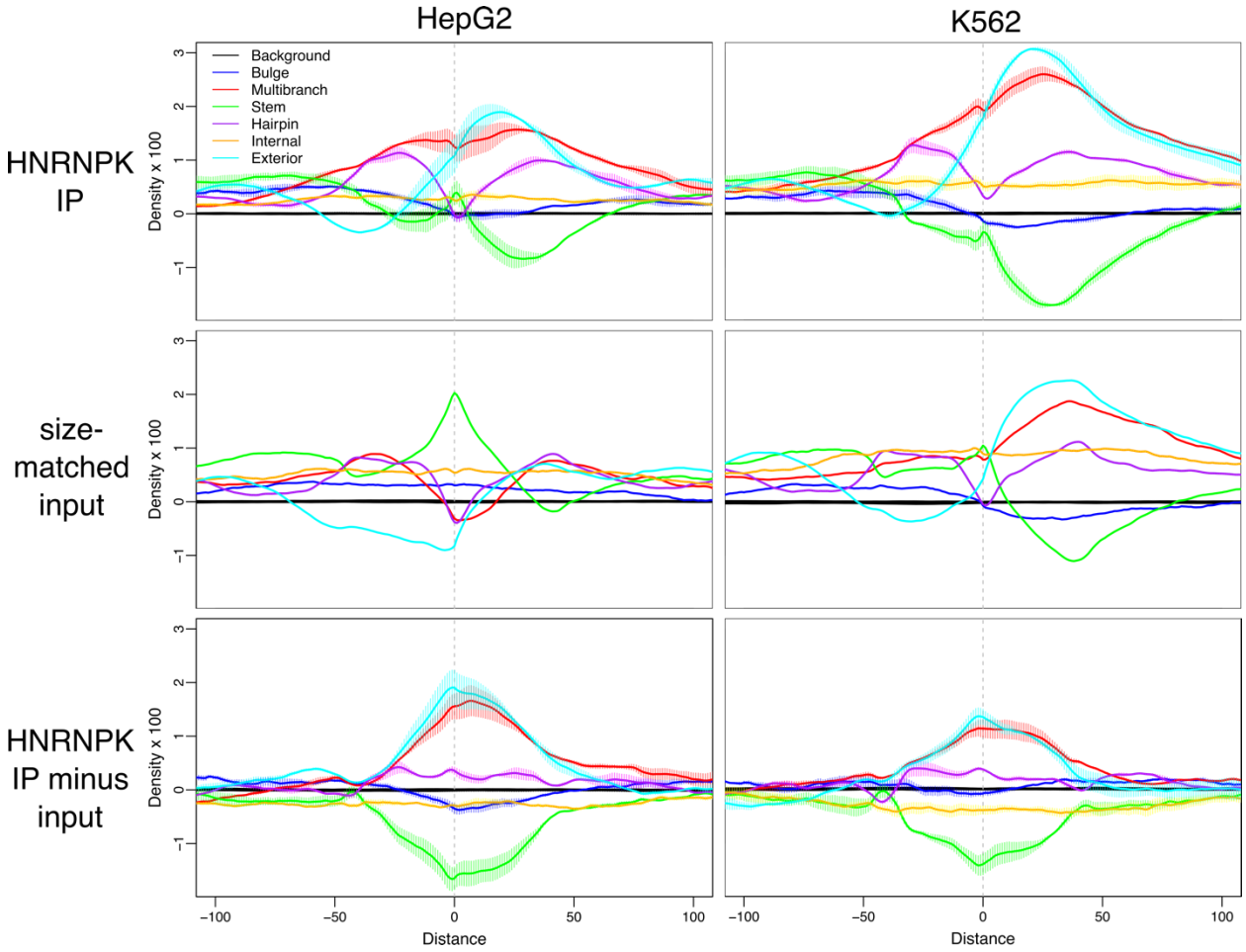


A. The user inputs CLIP-seq data (aligned reads or called peaks), a list of transcripts, and an annotated genome. Optionally, in silico folding of RNA intervals can be replaced by an experimentally-derived RNA structure track. RBP binding and RNA structure data is mapped to the concatenated transcriptome and cross-correlated. nearBynding outputs cross-correlation densities and their distributions to estimate RNA binding.

B. Examples of six RNA structure contexts predicted by CapR (40) for which nearBynding can be applied.

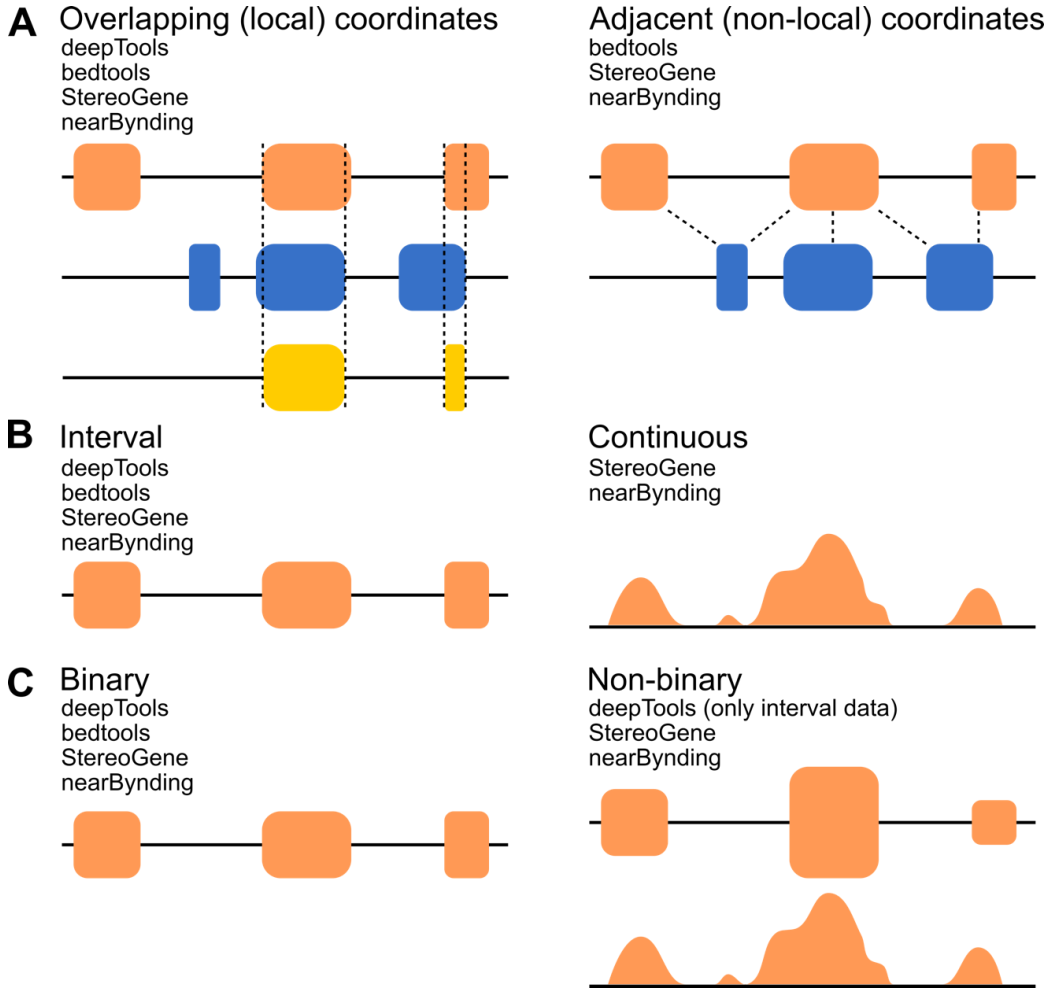
C. Example heatmap and line plot visualizations of PUM2 binding from eCLIP data in two replicates of K562 cells estimated from the cross-correlation densities and visualized as part of the nearBynding software. The line plot shows the average signal as a dark line and error bars corresponding to +/- three standard errors as lighter-colored shading; the black line and error bars represent null signal (labelled as background) derived from recursively shuffled foreground signal. The heat map only shows the average value at every position if multiple samples are used to calculate density values to model binding.

Figure 2.2 Removing cell line-specific background from binding profiles



Binding profiles of HNRNPK immunoprecipitation alone (top), size-matched input alone (middle), or HNRNPK IP minus size-matched input signal (bottom) for HepG2 (left) and K562 (right) cell lines. Error bars represent the confidence interval for +/- one standard error.

Figure 2.3 Graphical representations of different data types

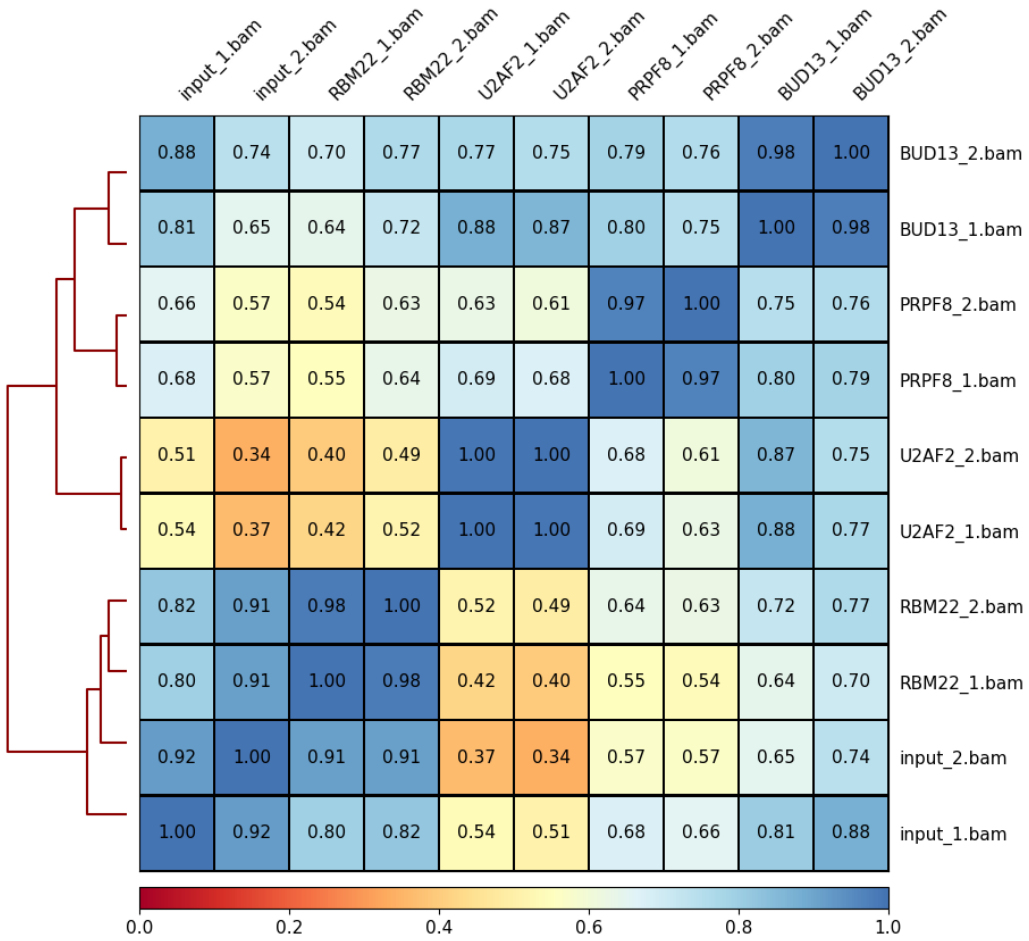


A. Overlapping coordinates (left) have positions in common, shown in yellow, for all tracks being compared, shown in blue and orange. Adjacent coordinates (right) may or may not have positions in common, and so analysis instead focuses on the distances between data, shown as dotted lines, for all tracks being compared, shown in blue and orange. Programs able to accommodate the indicated data type are listed below.

B. Interval data (left) have discrete start and end coordinates, with the same amplitude for every position between the terminal coordinates. Continuous data (right) may also have start and end coordinates, but can have diverse amplitudes for every position along the track.

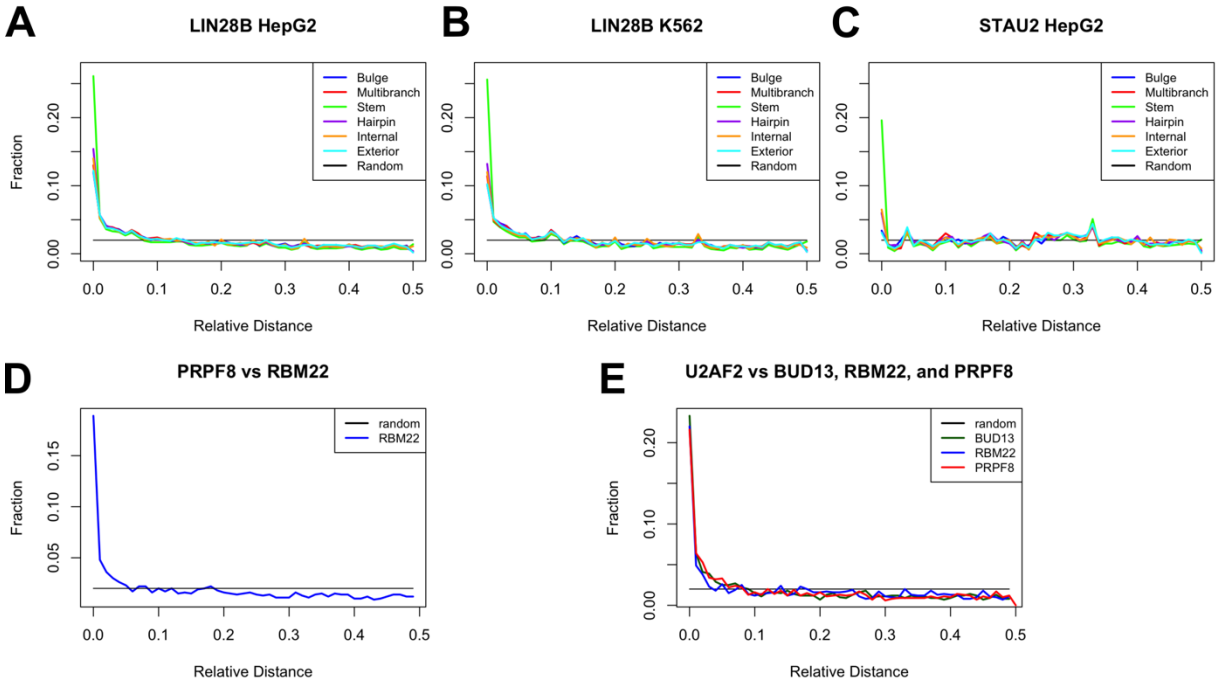
C. Binary data tracks (left) only include Boolean values, where each position either has a signal or does not. Non-binary data tracks (right) can include a variety of amplitudes along a track, either as interval data (top) or continuous data (bottom).

Figure 2.4 deepTools output



deepTools output correlating BUD13, PRPF8, U2AF2, RBM22, and two size-matched input BAM files from K562 eCLIP experiments. Colors and numbers inside boxes represent Pearson correlation values between files. Pearson correlation and visualization were performed using default arguments.

Figure 2.5 Graphed BEDtools *reldist* outputs



A. LIN28B binding in HepG2 cells.

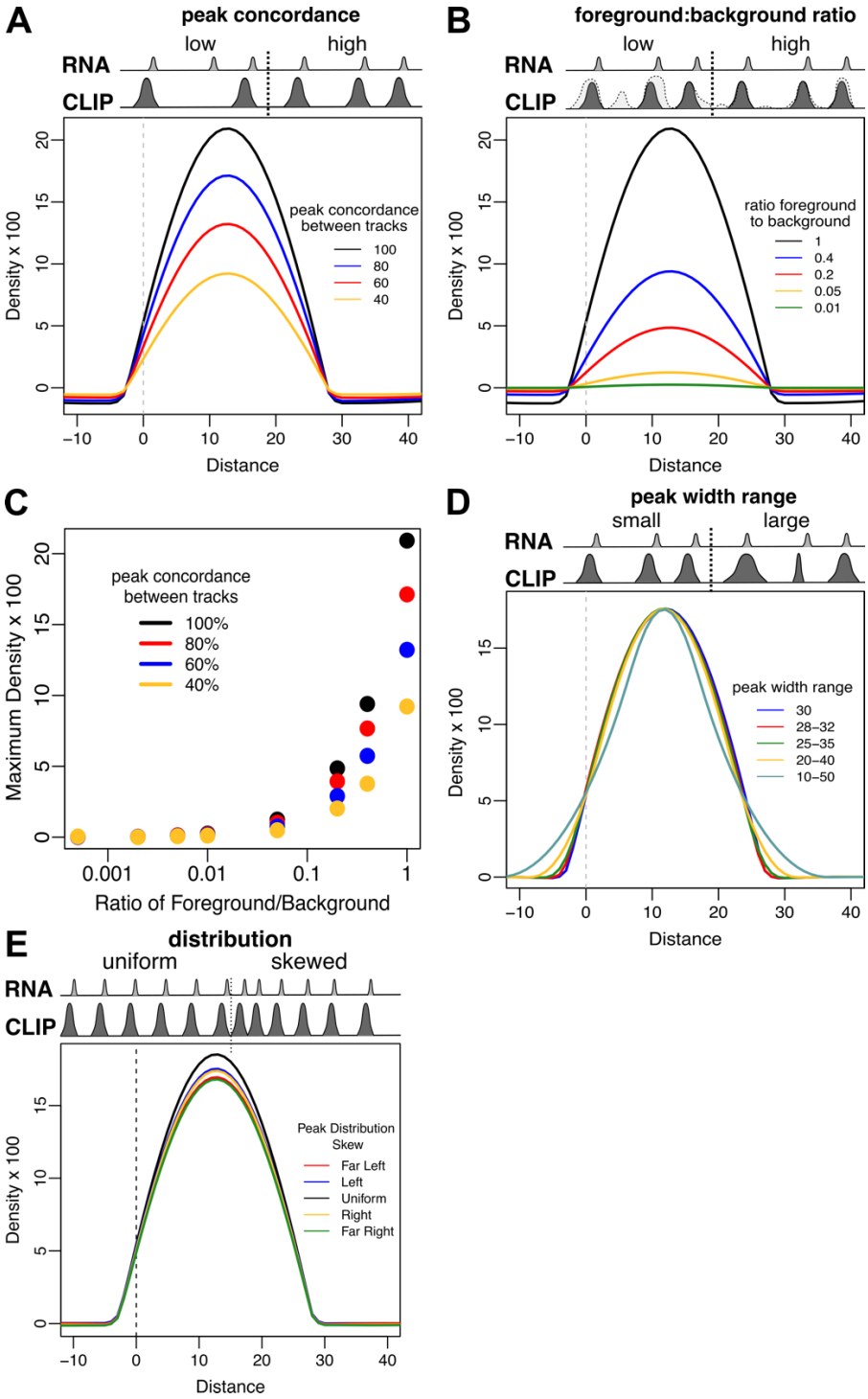
B. LIN28B binding in K562 cells.

C. STAU2 binding in HepG2 cells.

D. PRPF8 versus RBM22 binding in K562 cells.

E. U2AF2 versus BUD13, PRPF8, and RBM22 binding in K562 cells.

Figure 2.6 Cross-correlation distribution tracks of simulated RNA binding data to benchmark the performance of nearBynding



In the simulations, the RNA structure track [RNA] is shifted twelve units to the right relative to the CLIP track [CLIP] to model proximal RNA structure, as reflected by the maximum cross-

correlation density at distance=12. For A, B, and D, the middle grey peaks represent RNA structure data and the dark grey peaks represent signal from CLIP simulation data (i.e. foreground).

A. Cross-correlation distribution tracks with differing peak concordance.

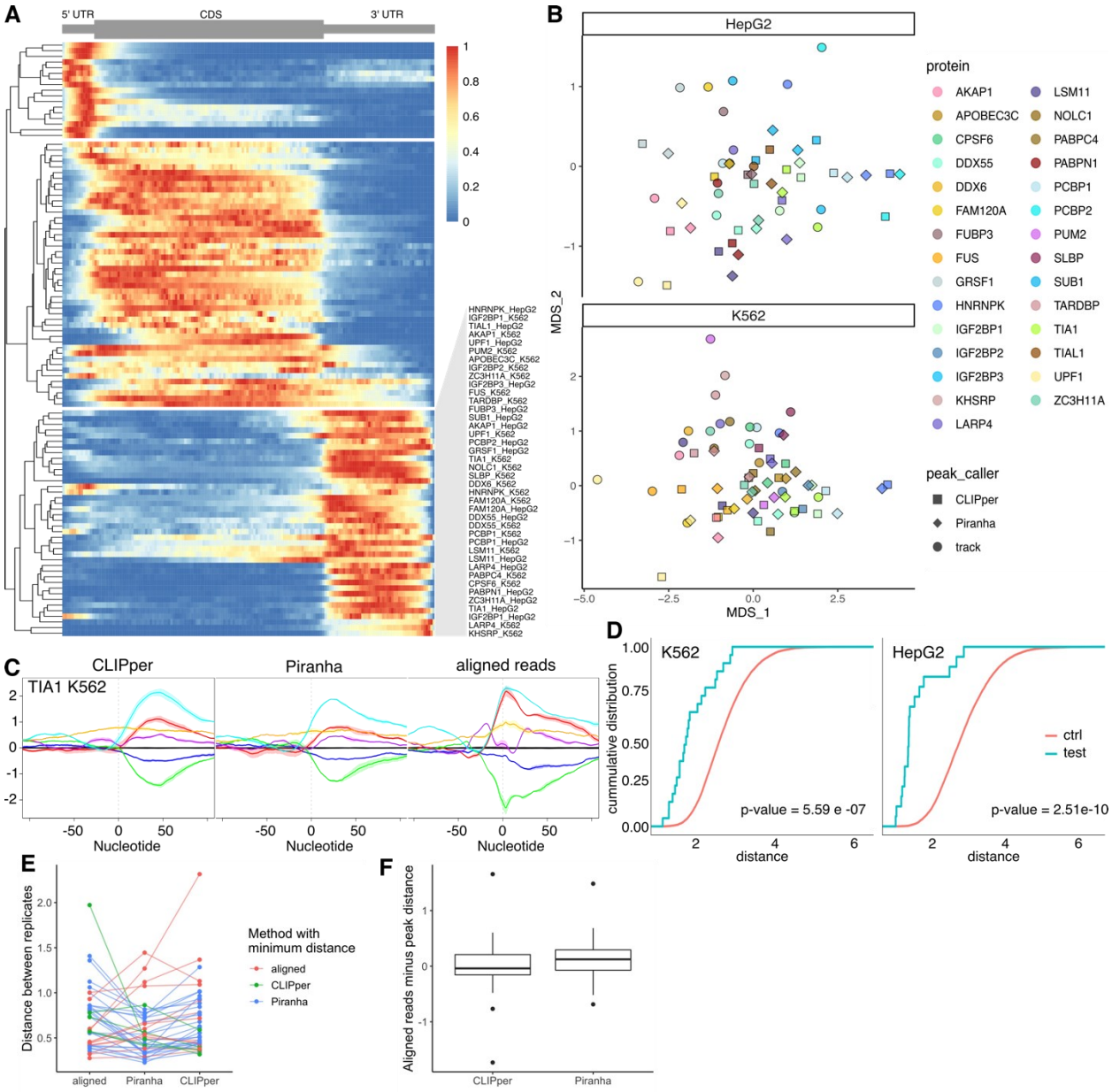
B. Cross-correlation distribution tracks with differing foreground to background signal ratios. Foreground signal (dark grey) does not change, but different amounts of randomly distributed background signal is added to the foreground, as represented by the lightest grey regions of peaks.

C. Maximum cross-correlation density values for pairs of tracks with varying peak concordance and foreground to background signal ratios.

D. Cross-correlation distribution tracks with differing peak width range.

E. Graphical depiction of simulated skewing and cross-correlation distribution tracks of simulated data with differing peak skews.

Figure 2.7 Comparison of aligned reads and called peaks as nearBynding inputs



A. Heat map of high-signal eCLIP samples across mRNAs, generated as in ref. (50). RBP and cell type are listed for RBPs that primarily bind 3'UTRs.

B. Distances between binding profiles with RBP binding defined by peak-callers or aligned reads mapped into Cartesian space via multidimensional scaling. Signal was tested across all six predicted RNA structure contexts in HepG2 (top) and K562 (bottom) cell lines.

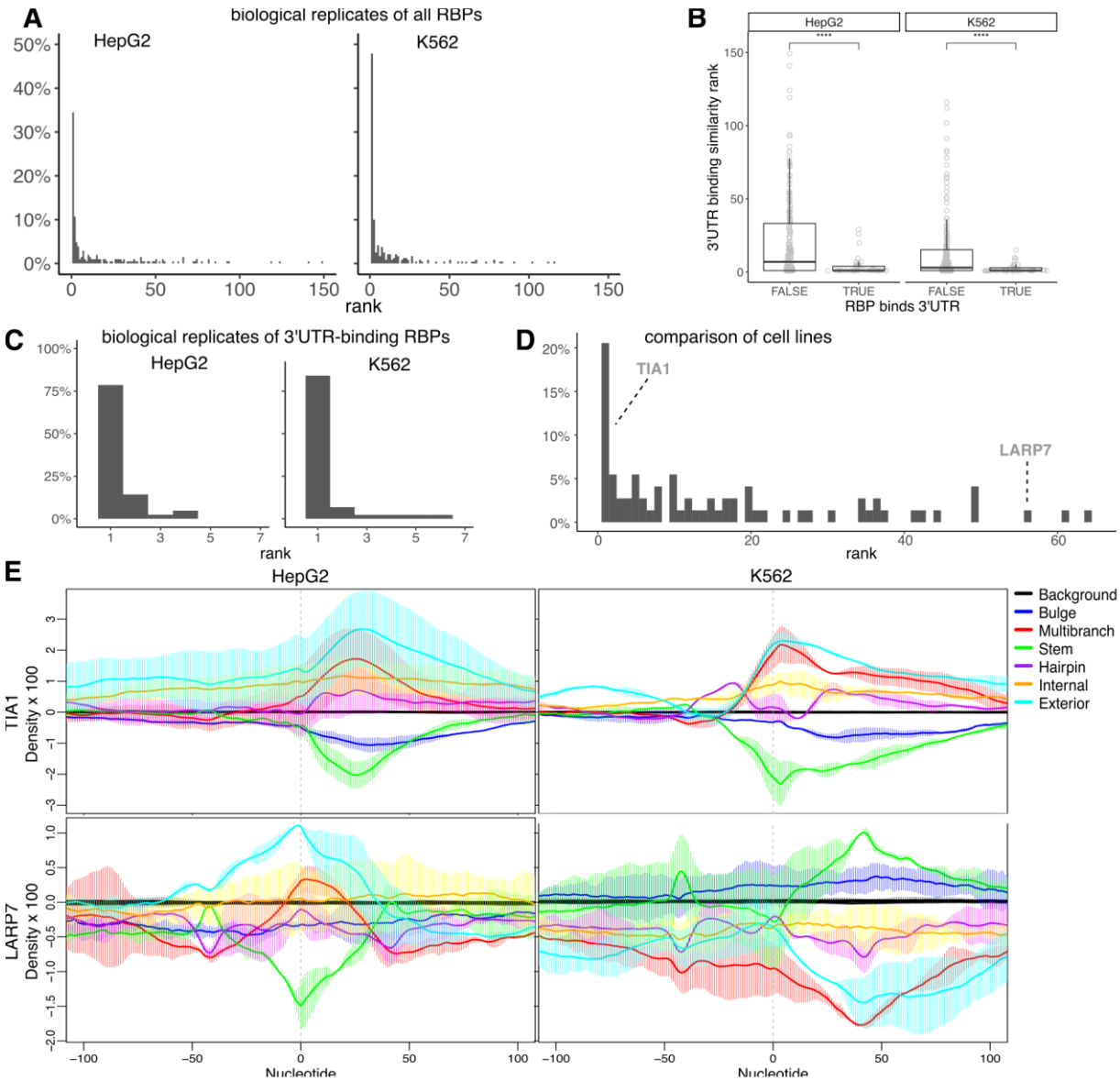
C. nearBynding output binding profiles of TIA1 in K562 cells with input derived from (left to right) CLIPper-called peaks, Piranha-called peaks, or aligned sequencing reads. Error bars represent the confidence interval for +/- one standard error.

D. Kolmogorov-Smirnov test comparing the mean distance between binding profiles for Piranha, CLIPper, and aligned-read inputs of the same protein versus bootstrapping for three random binding profiles in K562 (left) and HepG2 (right) cell lines.

E. Wasserstein distances of nearBynding outputs between technical replicates for the same samples as calculated from aligned reads, Piranha-called peaks, and CLIPper-called peaks. RBP and cell type samples are colored by which method generates the most similar technical replicate binding profiles for them.

F. The difference in Wasserstein distances for technical replicates between peak-calling methods tested and aligned reads. Positive values represent samples for which peak calling generated more similar outputs between technical replicates than the aligned reads tracks.

Figure 2.8 Binding profiles of all RBPs with eCLIP data in ENCODE clustered using Wasserstein distance



A. Histogram of ranks for Wasserstein distances of paired biological replicates in HepG2 (left) and K562 (right) cells.

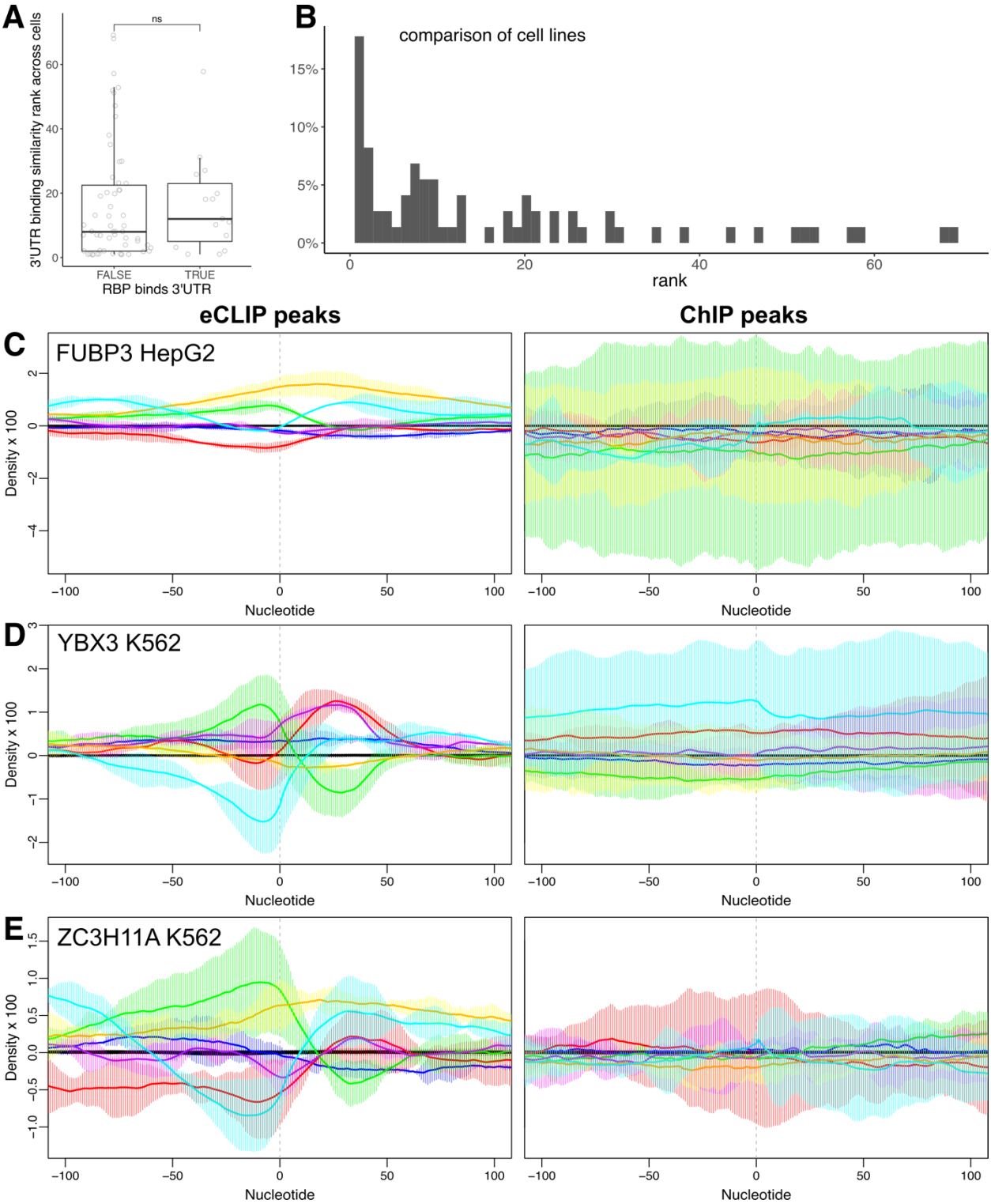
B. Boxplots of biological replicate binding profile similarity rankings for RBPs that do (TRUE) or do not (FALSE) preferentially bind 3'UTRs compared via t-test ($p < 0.0001$ for both) in HepG2 (left) and K562 (right) cells.

C. Histogram of similarity rankings for Wasserstein distances of paired biological replicates in HepG2 (left) and K562 (right) cells for RBPs that preferentially bind 3'UTRs.

D. Histogram of Wasserstein distance similarity rankings for the same RBP across HepG2 and K562 cell lines. The rankings of TIA1 and LARP7 across cell lines are indicated.

E. Example binding profiles for TIA1 (top), an RBP that is similar across cell types, and LARP7 (bottom), an RBP that is dissimilar across cell types, in HepG2 (left) and K562 (right) cells. Error bars represent the confidence interval for +/- three standard errors.

Figure 2.9 Benchmarking reproducibility and specificity based on ENCODE data

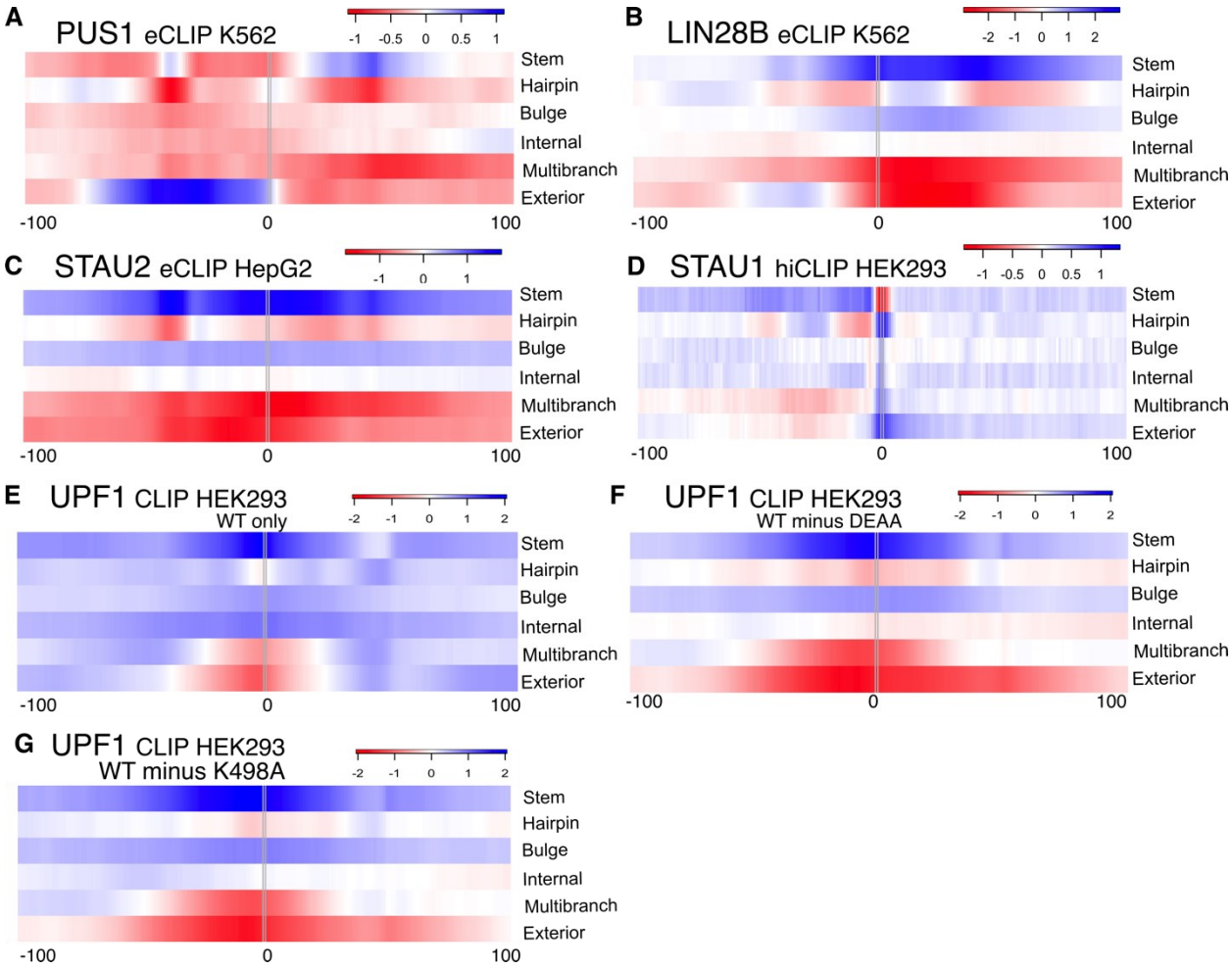


A. Boxplots of K562 versus HepG2 binding profile similarity rankings for RBPs that do (TRUE) or do not (FALSE) preferentially bind 3'UTRs assessed via t-test.

B. Histogram of Wasserstein distance similarity ranks for the same RBP in HepG2 versus K562 cell lines.

Binding profiles for 3'UTR-binding and DNA-binding proteins FUBP3 (C), YBX3 (D), and ZC3H11A (E) for eCLIP (left) and ChIP (right) data. Error bars represent the confidence interval for +/- three standard errors.

Figure 2.10 Application of nearBynding to analyze RBP binding profiles for proteins with known structure preference



Binding profiles for (A) PUS1, (B) LIN28B, and (C) STAU2 from eCLIP data.

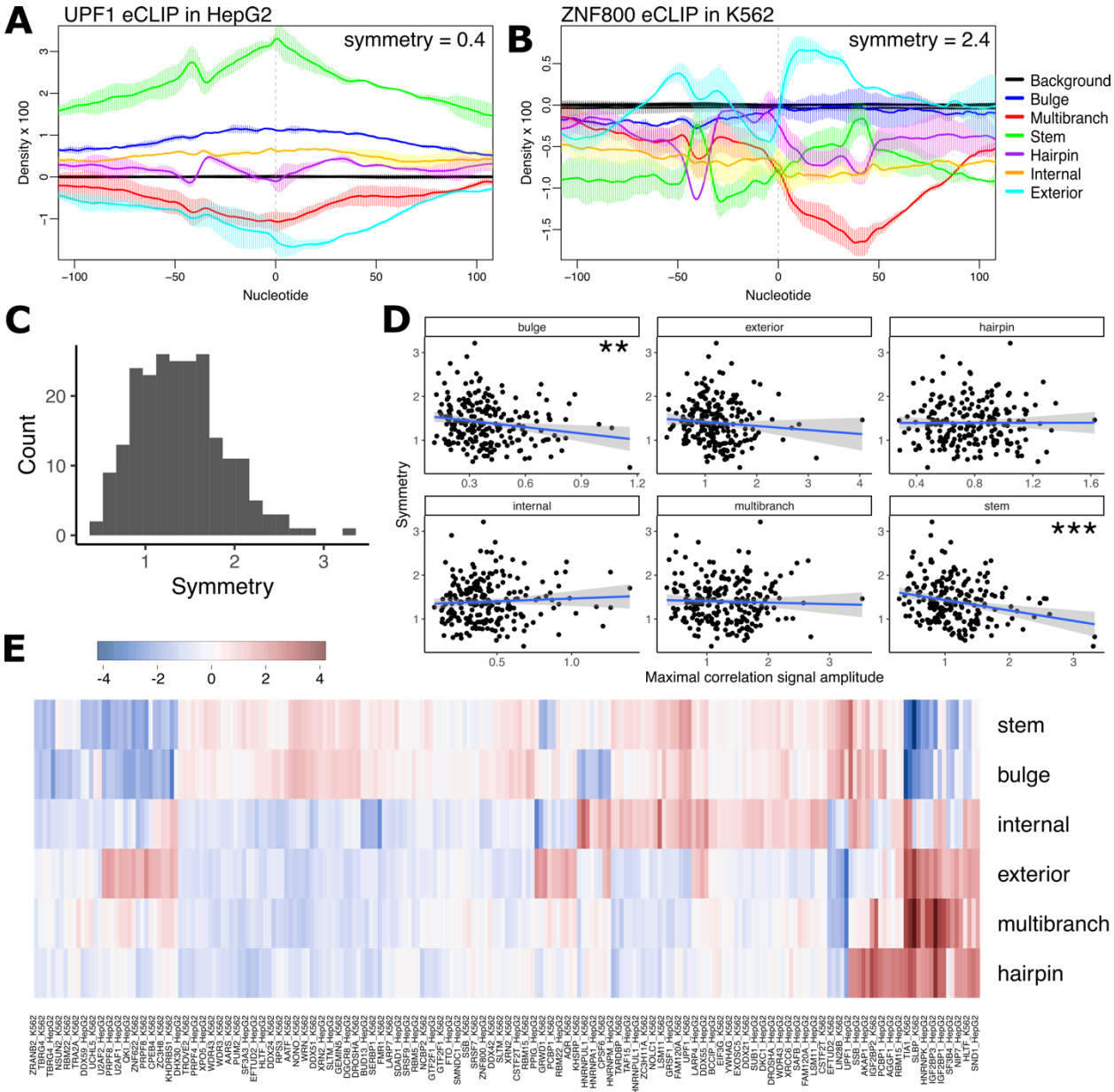
D. Binding profile for STAU1 from hiCLIP cross-link site data.

E. Binding profile for WT UPF1 binding from CLIP data.

F. Binding profile of helicase-dependent UPF1 binding based on subtraction of DEAA UPF1 signal from WT.

G. Binding profile of helicase-dependent UPF1 binding based on subtraction of K498A UPF1 signal from WT

Figure 2.11 The symmetry of protein binding profiles within 3'UTRs correlates with the strength of a protein's preference for stem context



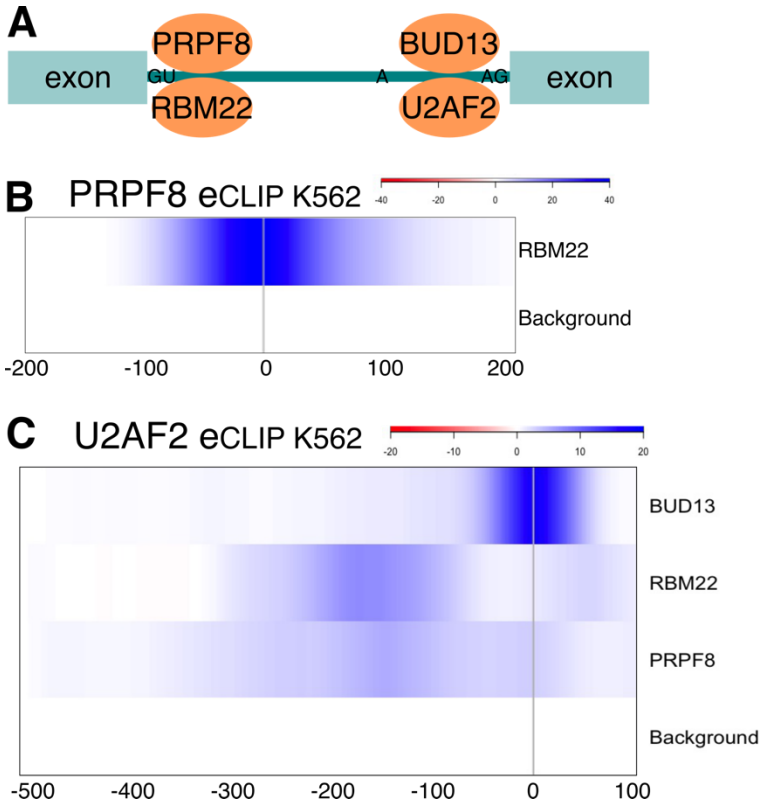
A. Binding profile of UPF1 in HepG2 from eCLIP reads within 3'UTRs.

B. Binding profile of ZNF800 in K562 from eCLIP reads within 3'UTRs. Error bars for A and B represent 3 standard deviations from the mean signal derived from two technical replicates and one background eCLIP track.

C. Distribution of protein binding symmetry values across all proteins in HepG2 and K562, where lower symmetry values represent more symmetrical binding profiles. D. Maximal absolute value signal amplitude versus protein binding profile symmetry for all proteins across each RNA structure context. Blue lines are the simple linear regression model with 95% confidence interval shaded.

E. Clustered heat map of largest correlation signal amplitude—positive (red) or negative (blue)—for each RNA binding context across all proteins. Signals are scaled within RNA structure contexts to maximize the visibility of differences.

Figure 2.12 Recapitulating known RBP binding preferences

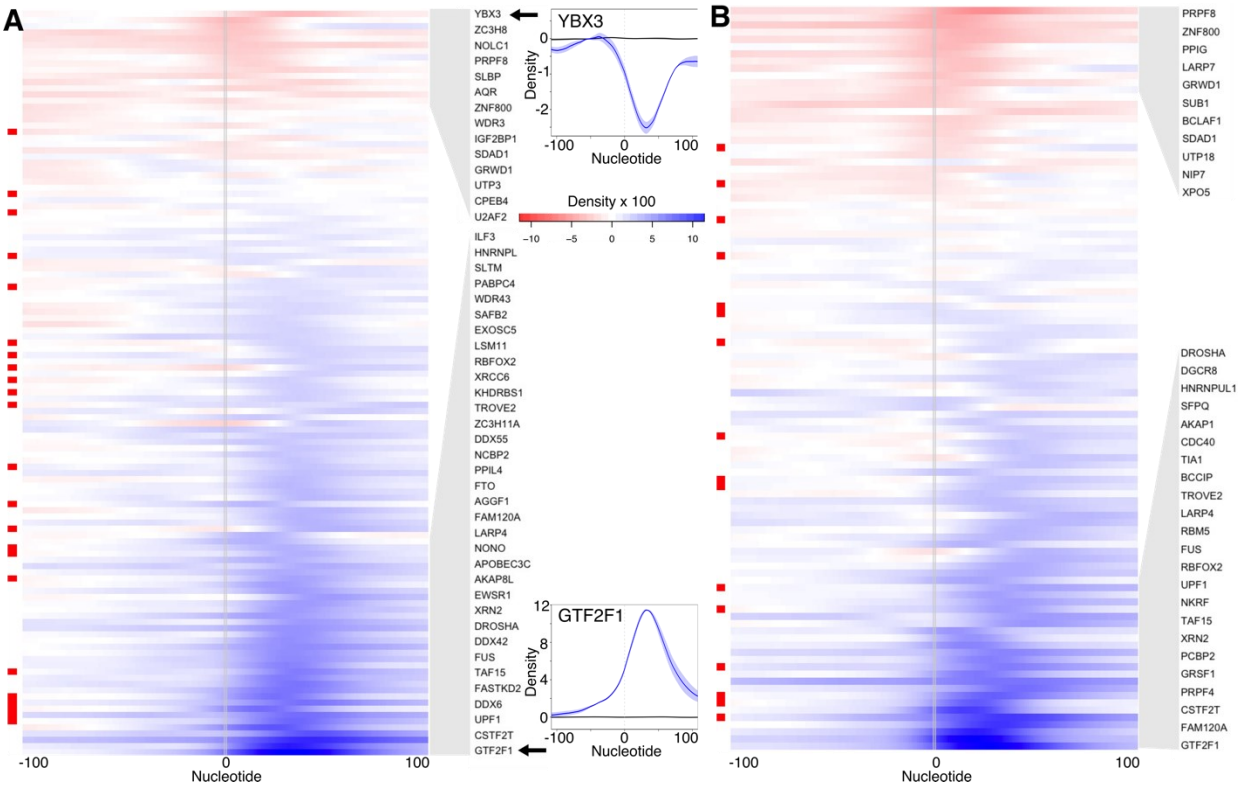


A. Graphical representation of relative binding locations within introns of four RBPs: PRPF8, RBM22, BUD13, and U2AF2.

B. Binding of RBM22 relative to PRPF8.

C. Binding of BUD13, RBM22, and PRPF8 relative to U2AF2.

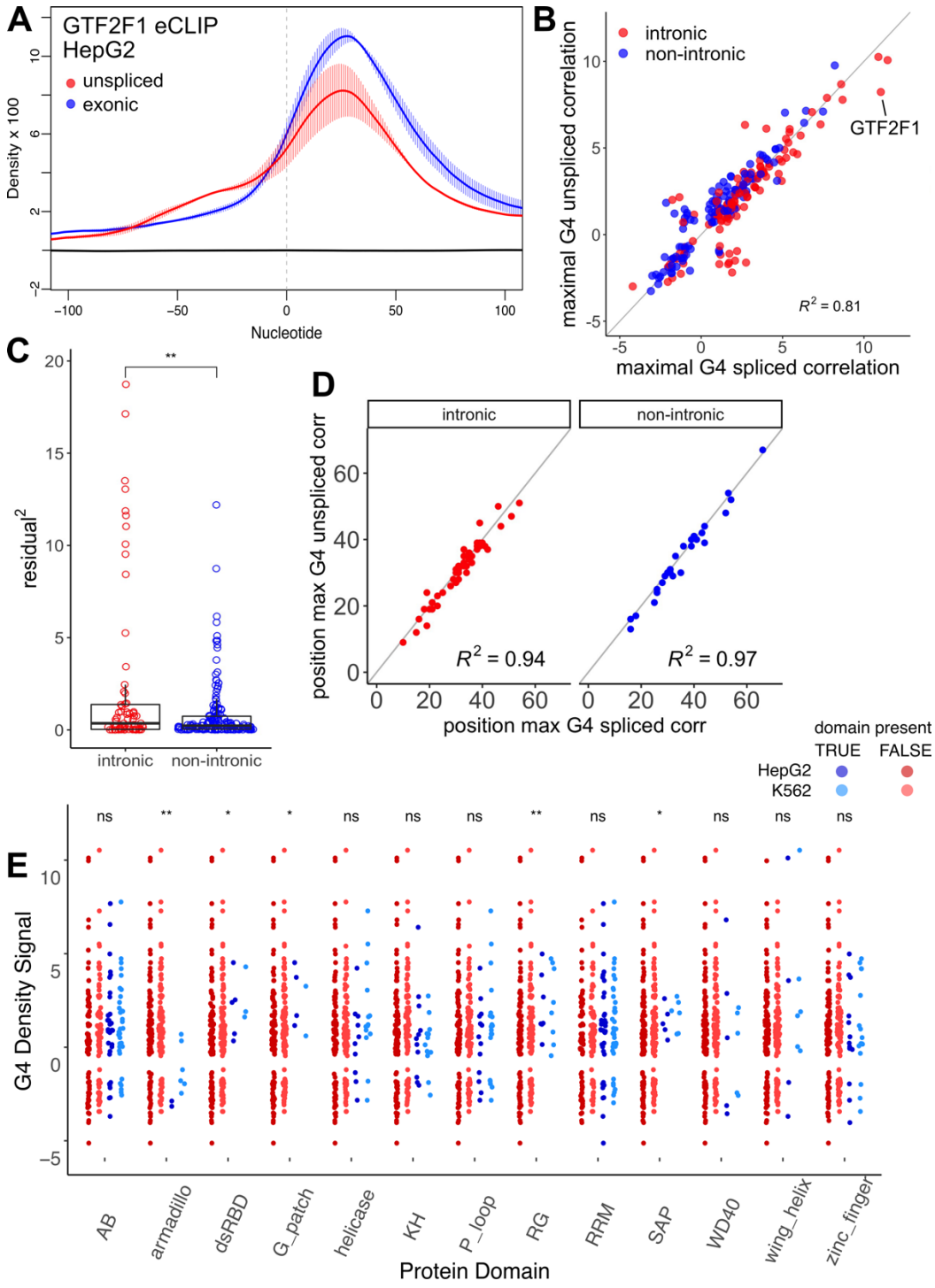
Figure 2.13 RBP binding to experimentally-derived exonic G4s



A. G4 binding profiles for all 120 proteins with eCLIP exome data in K562 cells. RBPs with molecular evidence of G4 binding in the literature are indicated in red on the left. RBPs with the most positive and most negative correlation signals are highlighted by the grey blocks and listed on the right, with line graph examples to the far right.

B. G4 binding profiles for all 103 proteins with eCLIP exome data in HepG2 cells.

Figure 2.14 G4 binding signal strength relative to transcriptomic regions and protein domains



A. Example G4 binding preferences for GTF2F1, an intron-binding protein, in HepG2 cells when exonic (blue) and unspiced (red) transcriptomic regions are analyzed by nearBynding.

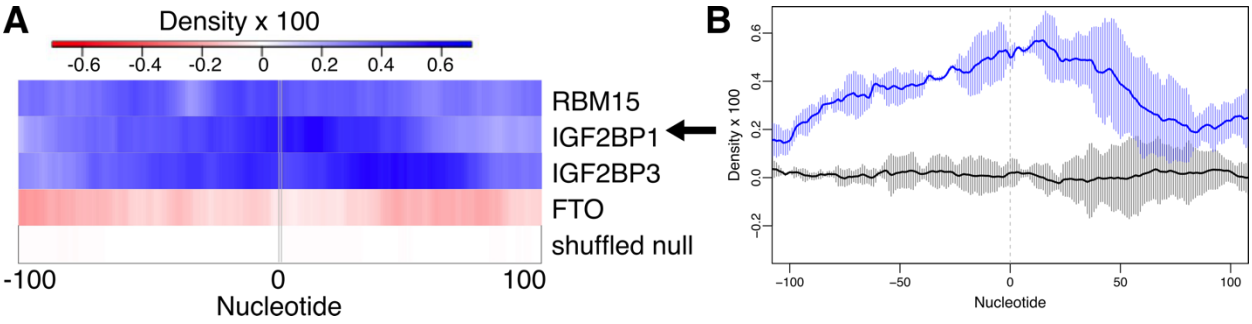
B. Correlation of nearBynding outputs for maximal G4 cross-correlation signal between exonic and unspliced transcriptomes. Proteins known to bind introns are shown in red. The grey line is y = x, and residuals are the distance of any point from this line.

C. Residuals of maximum G4 correlation signal between exonic and unspliced transcriptomic regions for intron-binding and non-intron-binding proteins.

D. Correlation of nucleotide positions of maximal G4 binding signal between exonic and unspliced transcriptomic regions for intron-binding (left) and non-intron-binding (right) RBPs.

E. Sina plots comparing maximal G4 binding signals that contain or do not contain the protein domains indicated for K562 and HepG2 proteins. Protein domains tested are alpha-beta (AB), armadillo, double-stranded RNA-binding (dsRBD), G-patch, helicase, K homology (KH), phosphate-binding loop (P-loop), RGG-repeats (RG), RNA-recognition motif (RRM), SAP, WD40, winged helix, and zinc finger. Significance was assessed via t-test and indicated as ns ($p > 0.05$), * ($p < 0.05$), or ** ($0.01 > p > 0.001$).

Figure 2.15 RBP binding profiles for experimentally-derived m6A



A. m6A binding profiles for RBM15, IGF2BP1, IGF2BP3, FTO, and shuffled null track based on miCLIP-seq data.

B. Line plot of binding profile for IGF2BP1 and shuffled null track with error bars representing a confidence interval of +/- three standard errors.

Table 2.1 Comparison of functionality for correlation analysis programs

		deepTools *plotCorrelation*	bedtools *reldist*	StereoGene	nearBynding
Data type	Optimized for single-nucleotide resolution				+
	Analyze interval data	+	+	+	+
	Analyze non-binary data	+		+	+
	Analyze continuous data			+	+
Correlation capabilities	Partial correlation			+	+
	Cross-correlation			+	+
	Correlate non-local features		+	+	+
	Correlation visualization	+		+	+
Transcriptome tools	Retains strand information				+
	Select regions of transcriptome				+
	Generate and integrate RNA structure predictions				+

Table 2.2 Multivariate regression model predicting protein binding profile symmetry from all maximal RNA structure context preference signals

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.551	0.130	11.885	2.146E-25
|stem|	-0.319	0.104	-3.065	2.452E-03
|bulge|	-0.118	0.241	-0.490	0.624
|internal|	0.203	0.144	1.410	0.160
|exterior|	0.074	0.106	0.698	0.486
|multibranch|	0.094	0.102	0.917	0.360
|hairpin|	-0.057	0.176	-0.324	0.746

Table 2.3 Literature sources of RNA G4-binding proteins

Protein	**PMID**
DDX21	29906500; 28472472; 23381195
DDX3X	30256975
DDX42	31287417
DHX30	23381195
DROSHA	29171334; 26554903
FMR1	11532944; 25692235; 19396385; 21642970; 26374839; 24876161; 32510328
FUS	23381195; 29800261; 32118033; 24251952; 23521792
FXR1	30256975; 27606879
FXR2	30256975
GRSF1	29967381; 30256975; 29129743
HNRNPA1	15302914; 18776302; 28510424; 29361764
HNRNPK	23381195
HNRNPL	23381195
HNRNPU	29129743; 24771345
ILF3	23381195
NONO	32496517; 23381195
NPM1	22707729
SAFB	23381195
SAFB2	23381195
SFPQ	23381195
SRSF1	24771345
TAF15	18776329
TARDBP	24920338
WRN	30279242
XRN2	21908404

Table 2.4 Influence of RBP domains on G4 binding signal

Domain	G4 Signal Difference	Effect Size	p-value
armadillo	-2.41	0.913	0.00147
RGG-repeat	1.57	0.592	0.00612
SAP	0.836	0.312	0.0123
dsRBD	1.57	0.589	0.0219
G-patch	1.77	0.662	0.0385
alpha-beta	0.476	0.178	0.198
winged-helix	1.88	0.707	0.238
RRM	0.385	0.144	0.301
K-homology	-0.384	0.143	0.402
helicase	0.428	0.16	0.436
WD40	-0.479	0.179	0.66
P-loop	0.0385	0.0143	0.935
zinc finger	0.000907	0.000338	0.999

Statistics from pooled HepG2 and K562 binding profiles. G4 signal difference is the difference in mean G4 binding signal between proteins with and without the indicated domain. Effect size is Cohen's *d* and the p-value is of a t-test comparing G4 signal of proteins with and without the indicated domain.

Table 2.5 Key Resources

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
ENCODE eCLIP, ChIP, and RNA-seq	(49)	https://www.encodeproject.org/
iMaps STAU1 hiCLIP	(55)	https://imaps.genialis.com/iclip/search/collection/hi-clip-reveals-m-rna-secondary-structures
UPF1 WT, K498A, and DEAA CLIP-seq	(57)	GEO: GSE69586
Ensembl FASTA and GTF files	(77)	ftp://ftp.ensembl.org/pub/release-100/
UCSC hg38 chain file	(79)	http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz
Pfam protein domains	(69)	http://pfam.xfam.org/
RGG-repeat-containing proteins list	(80)	Supplementary data
HepG2 miCLIP-seq	(75)	GEO: GSE121942
HEK293 rG4-seq	(61)	GEO: GSE77282
HEK293 RNA-seq	(78)	GEO: GSE122425
Software and algorithms		
nearBynding	This work (82)	https://bioconductor.org/packages/nearBynding/ ; https://github.com/vbusa1/nearBynding
CapR	(48)	https://github.com/fukunagatsu/CapR
StereoGene	(15)	http://stereogene.bioinf.fbb.msu.ru/
BEDtools	(76)	https://bedtools.readthedocs.io/en/latest/
Piranha	(44)	http://smithlabresearch.org/software/piranha/
deepTools	(47)	https://deeptools.readthedocs.io/en/develop/index.html#

Chapter 3: Linking structure-mediated RNA decay and racial disparities in breast cancer

Highlights

- A novel metric to assess global RNA structure negatively correlates with G3BP1 expression across cell types
- African American cancer samples have greater global RNA structure and lower G3BP1 expression than European American samples across many cancer types
- Five candidate miRNAs may drive the differential regulation of global RNA structure in breast cancer between African American and European American samples
- The race-based global RNA structure difference is attributable to expression of highly-structured RNA with functions associated with cell migration, adhesion, and proliferation

Introduction

RNA encodes information not only in its nucleotide sequences but also in its folded structures formed through base-pairing. Genome-wide RNA structure profiling reveals that a sizable fraction of mammalian RNA is highly structured in vivo and that RNA structures are prevalent across untranslated mRNAs and at 3'UTRs of translating mRNAs (83). mRNA decay is often regulated by proteins that recognize specific single-stranded sequences within the 3' UTR of mRNAs (84) or by proteins that recognize discrete RNA structures within the 3' UTR, such as iron response elements and stem-loop structures (85–87). Structure-mediated RNA decay (SRD) is an RNA decay pathway that depends on overall base-pairing density, which is quantified as the overall change in free energy due to RNA folding normalized to transcript length (-**Δ**G/nt) (32). SRD relies on two double-stranded RNA binding proteins: UPF1, a processive RNA helicase (57,88), and G3BP1, a putative endonuclease (89). SRD decays both coding and non-coding RNAs and is predicted to regulate roughly one-third of mRNAs across the transcriptome (32). Combinatorial genetic depletion assays revealed that neither UPF1 nor

the G3BP1 homolog G3BP2 can regulate mRNAs without G3BP1, indicating that G3BP1 is epistatic to UPF1 and G3BP2 in SRD (32).

G3BP1 has been implicated in numerous cancer types and is overexpressed in many, including breast, lung, esophageal, hepatocellular, colon and pancreatic cancer (90–95). G3BP1 affects cancer migration, proliferation, and growth by influencing cell signaling and mRNA expression levels (31). However, individuals with breast cancer that had high G3BP1 expression have been shown to have longer recurrence-free survival than individuals with low expression in, for example, breast cancer (95) (Fig. 3.1A). We hypothesize that G3BP1's role in many types of cancer is also linked to its involvement in SRD. We investigated the correlation between RNA structure and various phenotypes in publicly-available cancer datasets and cancer cell lines in order to determine if SRD plays a role in cancer.

We used WT and G3BP1 KO or KD cell culture data to create a metric to quantify a sample's global RNA structure. Using the metric, we found a race-based discrepancy in global RNA structure between African American and European American samples in cancer samples that inversely correlates with G3BP1 expression. Race-based discrepancies in global RNA structure and G3BP1 expression are absent in healthy samples, suggesting that G3BP1, via SRD, may contribute to racial differences in cancer progression and prognosis. The RNA structure discrepancy persists in breast cancer samples even after correcting for G3BP1 expression, suggesting G3BP1 contributes to but does not explain the entirety of race-based global RNA structure differences. We further identify miRNAs and transcripts that may regulate SRD in cancer and contribute to race-based discrepancies in global RNA structure.

Results

-**Δ**G/nt correlates with G3BP1-dependent decay for isoforms with highly-structured 3'UTRs

We defined highly-structured 3'UTRs (HSUs) as transcripts for which the 3'UTR -**Δ**G/nt > 0.25 and poorly-structured 3'UTRs (PSUs) as transcripts for which the 3'UTR -**Δ**G/nt < 0.25. Prior studies have focused on a small subset of transcripts for study of SRD and adopted more stringent cut-offs to separate highly-structured and poorly-structured transcripts (32), so we aimed to determine whether HSUs and PSUs are differentially affected by G3BP1 expression across the transcriptome based on our -**Δ**G/nt cut-offs. DLD1 cells demonstrate SRD similar to other cell lines tested, so we postulate that any SRD observations made in DLD1 cells will be representative (32). We therefore used RSEM quantile normalized transcript per million (tpm) RNA-seq data from wild-type (WT) and G3BP1 knock-out (KO) DLD1 cells that have been harvested 0 to 16 hours post-actinomycin-D (actD) treatment (32). Since actD halts transcription, any differences in mRNA expression between 0- and 16-hour time points can be attributed solely to RNA decay. By comparing mRNA levels between 0 and 16 hours post-actD treatment between WT and G3BP1 KO cells, we can calculate G3BP1-dependent differences in mRNA decay for all transcripts, or decay odds ratio (OR), where a large decay OR suggests a transcript's decay rate is highly affected by G3BP1. We filtered for transcripts that had an average of at least one tpm across all samples and that demonstrated differential expression between 0 and 16 hours post-actD treatment (t-test $p < 0.05$) to ensure that we had high-confidence decay data.

$$\text{Decay OR} = \frac{\text{WT 0 hrs post-actD} / \text{WT 16 hrs post-actD}}{\text{G3BP1 KO 0 hrs post-actD} / \text{G3BP1 KO 16 hrs post-actD}}$$

There was no effect of G3BP1 expression on decay ORs among PSUs (linear model $p = 0.91$, Fig. 3.1B). That is, the effect of G3BP1 on transcript decay rate is the same among PSUs, regardless of the -**Δ**G/nt. In contrast, decay OR positively correlates with -**Δ**G/nt among HSUs

(linear model $p = 3.35 \times 10^{-11}$, Fig. 3.1B), suggesting that there is increasing effect of G3BP1 on transcript decay rate as HSU structure increases.

G3BP1 expression is a predictor of novel global RNA structure metric

We aimed to develop a metric that could serve as a single-value indicator of SRD activity within a cell. Since an isoform's -**Δ**G/nt is correlated with how much its expression is affected by G3BP1, we multiplied each isoform's expression by its -**Δ**G/nt to more heavily weight transcripts with more structure. We then took the ratio of the sum of normalized read counts of mRNA isoforms with HSUs to the sum of normalized read count of mRNA isoforms with PSUs. This HSU/PSU ratio represents the global RNA structure of a cell.

$$\text{HSU/PSU ratio} = \frac{\Sigma \text{ HSU isoform mRNA count}^* \ \Delta G/nt}{\Sigma \text{ PSU isoform mRNA count}^* \ \Delta G/nt}$$

To assess whether the HSU/PSU ratio is a reliable metric to represent SRD within cells, we tested whether G3BP1 expression can predict HSU/PSU ratio; we hypothesized that as G3BP1 expression increases, it causes a decrease in HSU expression via SRD, and therefore would cause a decrease in the HSU/PSU ratio. We used RNA-seq data from WT and G3BP1 KO DLD1 cells as well as SHSY-5Y cells that had been treated with scrambled or G3BP1 shRNA (32). G3BP1 expression was lower in the G3BP1 KO and shG3BP1-treated cells than the WT and shScram-treated cells, as expected (Fig. 3.2A). We calculated the HSU/PSU ratio for each sample and observe a strong negative correlation between HSU/PSU ratio and G3BP1 expression (linear model $R^2 = 0.79$, $p = 0.0196$, Fig. 3.2B), in alignment with our hypothesis.

To test the global RNA structure metric in tissue samples, we obtained isoform counts from RNA-seq data for cancer patients from The Cancer Genome Atlas (TCGA, https://www.cancer.gov/tcga) and calculated the HSU/PSU ratio for each sample. 19 of the 25 cancer types available demonstrate a negative correlation between G3BP1 expression and HSU/PSU ratio, including all 15 cancers for which there are at least 300 patient samples (linear

model $p < 0.05$, Table 3.1, Fig. 3.2C). These data cumulatively suggest that the HSU/PSU ratio is a robust metric to study SRD, since it is derived from a sample's RNA structure and is reproducibly predicted by the limiting protein of SRD, G3BP1.

HSU/PSU ratio separates African American and European American cancer samples

We investigated whether this new SRD metric is able to separate cancer phenotypes using data from TCGA. Five-year relative survival is higher in White than Black Americans for over a dozen cancer types (96), so we tested whether SRD-- which acts in all cell types that have thus-far been tested-- varies across cancer samples based on their race. We compared the HSU/PSU ratio between Black or African American (AA) and White or European American (EA) patients for all cancer types available in TCGA for which at least 20 samples were available for both races, and 5 out of 15 cancer types have a different HSU/PSU ratio for AA versus EAs (t-test $p < 0.05$, Fig. 3.3A), though the effect is subtle. All cancer types with a statistically significant difference in HSU/PSU ratio had a higher HSU/PSU ratio in AA patients compared to EA patients.

We focused on breast invasive carcinoma (BRCA) patients since they make up the largest cohort of samples and found that HSU/PSU ratio is weakly inversely correlated with months of disease-specific survival time (linear model, $p = 0.017$, Fig. 3.3B). AA primary tumors have a higher HSU/PSU ratio than EA tumors at all cancer stages and HSU/PSU increases with increasing cancer stages (multiple linear regression test $p = 6.6 \times 10^{-20}$ for racial difference, $p = 0.014$ for cancer stage difference, Fig. 3.3C). Interestingly, G3BP1 expression does not change across cancer stages, but it is lower in AA than EA tumor samples at every cancer stage (multiple linear regression test $p = 2.0 \times 10^{-7}$ for racial difference, $p = 0.36$ for cancer stage difference, Fig. 3.3D). Yet, this disparity in HSU/PSU ratio and G3BP1 expression is absent in normal breast tissues based on Genotype-Tissue Expression data (GTEx, t-test $p = 0.21$ for HSU/PSU ratio and $p = 0.97$ for G3BP1 expression, Fig. 3.3E and F). This suggests that SRD in

cancerous breast tissue differs between AA and EA samples, whereas healthy tissue SRD does not, and that the HSU/PSU ratio may be able to serve as a predictive marker for cancer prognosis.

SRD drives HSU/PSU difference between European American and African American breast cancer samples

To test whether G3BP1 expression alone could account for the race-based difference in global RNA structure, we removed differences attributable to G3BP1 expression via regression and re-assessed HSU/PSU ratio between AA and EA TCGA BRCA samples. Although there was a reduction in the HSU/PSU ratio difference, AA BRCA samples still showed greater global RNA structure than EA samples (t-test $p = 4.0 \times 10^{-10}$, Fig. 3.4A), suggesting that G3BP1 expression alone is insufficient to entirely explain the race-based difference in global RNA structure.

To investigate which genes most explain the difference in HSU/PSU ratio between AA and EA BRCA samples, we subtracted each isoform for which expression and 3'UTR structure information was available one at a time from the HSU/PSU ratio and calculated how much that omission affected the difference in HSU/PSU ratio between the races. We ranked all isoforms based on their effect on the difference in ratio between the races, where negative values suggest the isoform drives the HSU/PSU ratio difference between AA and EA BRCA samples. 13,903 of 25,463 isoforms tested decrease HSU/PSU ratio difference between AA and EA breast cancer samples when removed, indicating that many genes' expressions contribute to the race-based differences of HSU/PSU (Fig. 3.4B). Removing the top 22 isoforms decreases the race-based HSU/PSU ratio difference only ~10%, and even the top 598 isoforms only account for ~50% of the difference, suggesting that no single isoform or small group of isoforms is responsible for the HSU/PSU ratio difference between AA and EA BRCA samples.

HSU/PSU differences could be affected by changes in either HSU or PSU expression, so we assessed isoform structure relative to its contribution to the race-based difference in HSU/PSU ratio. We found a strong trend toward structuredness as isoforms contribute more to the race-based HSU/PSU ratio difference (linear model, $p = 0$). We divided the isoforms ordered by their effect on HSU/PSU ratio race difference into 20 equal-sized bins for visualization and observe that the isoforms that most contribute to the race-based HSU/PSU ratio difference show a bimodal split in their structure distribution (Fig. 3.4C). Separating isoforms into HSUs and PSUs, there is an enrichment of HSUs among isoforms that increase the race-based difference in HSU/PSU ratio and a depletion of HSUs among isoforms that reduce it (Fig. 3.4D).

We performed gene ontology (GO) analysis on the isoforms ranked by their contribution to the race-based HSU/PSU ratio discrepancy to determine which differentially-expressed pathways drive the differences in global RNA structure. Ranked GO analysis suggests that isoforms that contribute to the race-based difference in HSU/PSU ratio are enriched for genes involved in actin filament-based processes, cell adhesion, cell migration, mitotic cell cycle, and extracellular matrix organization (Fig. 3.4E) (97,98). Cumulatively, these data suggest that the difference in HSU/PSU ratio between AA and EA BRCA samples is attributable to HSU expression, plausibly due to disparities in SRD activity, and that these differences may contribute to altered cancer proliferation and migration in AA patients versus EA patients.

Verification of SRD-derived phenotype in cultured breast cancer cells

We wanted to identify breast cancer cell lines cultured from patients that could be good models for testing the effects of SRD and associated candidate factors on breast cancer phenotypes such as migration and proliferation that we had identified from publicly-available data. We used unpublished RNA-seq count data for 55 cultured breast cancer samples to calculate HSU/PSU ratio and G3BP1 expression and observe a negative correlation (linear model, $p = 0.056$, Fig. 3.5A). We selected five cell lines that fell within the 95% confidence

interval of the linear model of the relationship (Table 3.4, Fig. 3.5A) for which to test expression of G3BP1, seven HSUs known to undergo SRD, and seven PSUs as negative controls. We ordered the samples by G3BP1 expression, measured by qPCR and normalized to TBP. No obvious correlation was observed across the diverse breast cancer types between G3BP1 expression and HSU expression (Fig. 3.5B). Given the sample size is small and lacks sufficient controls to make robust claims, further analysis with better controls would be required to assess whether the selected cell lines are good candidates for functional testing.

We wanted to assess how faithfully our qPCR results agreed with the RNA-seq data to confirm the reliability of the RNA-seq data relative to what we observe within the more recently-passaged cells. An R^2 close to 1 would suggest good reproducibility between RNA-seq and qPCR results, but we found poor correlation between RNA-seq- and qPCR-derived expression for the fifteen genes tested across all five cultured breast cancer samples (Fig. 3.5C). This may be due to an erroneous data analysis pipeline, excessive noise in the data, or large natural variation across biological replicates. Regardless, both the RNA-seq and qPCR ought to be repeated for more robust analysis. Due to this poor reproducibility, we did not look more deeply into global RNA structure based on the RNA-seq data.

Searching for additional factors associated with SRD in publicly-available datasets

miRNAs are potential SRD regulators

To look for factors that explain some of the HSU/PSU ratio difference between AA and EA cancer samples, we searched for miRNAs that target G3BP1. We hypothesized that a microRNA (miRNA) could affect SRD by the following proposed mechanism: a miRNA targets G3BP1 mRNA and decays or silences it; reduced G3BP1 levels lead to less SRD, which causes an increase in the relative abundance of structured mRNA; the shift in structured RNA could

then be measured as an increase in the HSU/PSU ratio (Fig. 3.6A). PSUs are expected to remain largely unaffected by G3BP1-targeting miRNA levels.

We used miRDB (99) to identify plausible G3BP1-targeting miRNAs and found 88 miRNAs for which there is available TCGA BRCA data; 50 of these had a median expression of 0 across samples and were excluded from further analyses. To identify which miRNAs could potentially affect SRD by targeting G3BP1, we filtered for miRNA who expression demonstrated a negative correlation with G3BP1 expression and a positive correlation with HSU/PSU ratio across 849 BRCA samples. Sixteen miRNAs fulfilled both of these criteria (Fig 3.6B). Further, the levels of thirteen of these miRNAs were higher in AA patients than in EA patients (Fig 3.6C), which is consistent with the directionality of the G3BP1 expression and HSU/PSU ratio differences between AA and EA breast cancer patients.

To identify whether the identified miRNAs, all of which are predicted to target G3BP1, may have a role in SRD and race-based differences in breast cancer prognosis, we interrogated whether they affect patient survival. Survival curves for TCGA breast cancer patients between high and low levels of these thirteen differentially expressed miRNAs indicate that patients with high levels of miR-19a, miR-548j, miR-138, miR-98, and miR-548o have a significantly lower survival probability (logrank P, respectively: 0.01, 5.5×10^{-4}, $p=5.0 \times 10^{-4}$, 0.0096, and 5.8×10^{-6}, Fig. 3.6D) (100), acting with opposite directionality to G3BP1 as expected if they were inhibiting G3BP1 expression and making them candidates of particular interest for future study. miR-589, miR-362, miR-19b, miR-889, miR-520a, miR-223, miR-3613, and let-7f levels do not negatively affect survival among TCGA breast cancer patients (Fig. 3.6E) (100).

Correlation with HSU/PSU ratio is a poor method to discover other SRD factor candidates

In addition to G3BP1-targeting miRNAs, there are likely proteins other than G3BP1 and UPF1 that regulate SRD. Since the HSU/PSU metric reliably correlates with G3BP1 expression

in a manner that is consistent with SRD, we tried to leverage it to discover new SRD candidate genes. We pooled expression across isoforms for each gene and performed linear models between gene expression and HSU/PSU ratio for 17,220 genes within the 1,212 BRCA samples. We filtered for genes whose expression had an inverse relationship with global RNA structure and a Bonferroni-corrected $p < 0.05$ ($p < 2.9 \times 10^{6}$). This yielded 5,721 significant hits, including G3BP1 and G3BP2 (Table 3.2). SRD factors that are epistatic to G3BP1 are possibly also contained among the significant hits. However, the large number of hits suggests that the vast majority of the correlated genes are false positives. Since G3BP1 is epistatic to UPF1 for SRD, it is not surprising that G3BP1 but not UPF1 was identified by this list, and we can expect that other non-rate-limiting SRD factors are also absent.

Since we hypothesized that G3BP1 expression might be masking other proteins' relationship with HSU/PSU ratio, we used a multivariate model to regress out the effect of G3BP1 expression from the correlation of other proteins' expression and HSU/PSU ratio. Filtering for Bonferroni-corrected $p < 0.05$ resulted in 7,870 significant hits, including UPF1 and G3BP2 (Table 3.3). All ten of the most significantly correlated genes and UPF1 have a positive relationship with HSU/PSU ratio after controlling for G3BP1 expression. Since this method resulted in so many positive hits, we concluded that HSU/PSU ratio is a poor method on its own to discover new SRD factor candidates and that gene expression correlations in multiple cell types should be compared to confirm candidates.

Spatial correlation via nearBynding software of protein binding from eCLIP data provides insight into SRD protein colocalization

miRNAs serve as upstream modulators of SRD by interacting with G3BP1 mRNA and modulating protein levels with high specificity, but we also wanted to search for SRD factors are directly involved in the SRD protein complex. Based on protein:protein interaction data in the BioGRID repository, there are 129 proteins that directly interact with both G3BP1 and UPF1

(101). Of these, 7 are associated with GO terms involving dsRNA (RC3H1, RC3H2, AGO2, CELF1, CSDE1, DDX3X, and FMR1) and only POP1 has endoribonucleolytic activity (102), other than G3BP1 (89,103). However, it is possible that additional SRD factors only directly bind UPF1 (319 candidates) or G3BP1 (433 candidates). Further, it is also possible that additional factors interact via protein:RNA:protein colocalization rather than direct protein:protein interfaces. Since both UPF1 and G3BP1 have multiple roles in RNA regulation, many of these interactions likely do not directly relate to their functions in SRD.

It is established that UPF1 and G3BP1 must co-localize to perform SRD and that SRD occurs within the 3'UTR of target mRNAs (32). There are no available breast cancer G3BP1 CLIP-seq datasets, so we used data from ENCODE, where there is eCLIP data for G3BP1, UPF1, and an additional 101 proteins in HepG2 cells, to directly study sites of protein binding to RNA (49). Normalized peak occupancy along all mRNAs calculated in (50) demonstrates that G3BP1 binds almost exclusively in the CDS, whereas UPF1 binds 3'UTRs (Fig. 3.7A). Filtering for only 3'UTR binding sites indicates that, even within the 3'UTR, there is little overlap between G3BP1-occupied and UPF1-occupied regions (Fig. 3.7B). This method of visualization pools all protein binding events, but we wanted to look exclusively at mRNAs that both G3BP1 and UPF1 occupied. We used nearBynding (82) to spatially correlate G3BP1 and UPF1 binding across the HepG2 transcriptome and observed that, when G3BP1 and UPF1 bind in proximity to one another, G3BP1 rarely colocalizes with UPF1 and instead binds approximately 400-600 nt upstream of UPF1 (Fig. 3.7C), consistent with G3BP1's preference for binding the CDS and UPF1's preference for 3'UTRs. However, when we exclusively interrogate G3BP1 and UPF1 binding spatial correlation within 3'UTRs, we see that G3BP1 and UPF1 mostly colocalize (Fig. 3.7D). These nearBynding results suggest that even though G3BP1 and UPF1 rarely both bind the 3'UTR of an mRNA, they colocalize when they do, and prior experiments using tethering

systems to compel UPF1 and G3BP1 co-binding indicate that we are likely observing SRD targets (32).

To search for other proteins that colocalize with G3BP1 within 3'UTRs, we performed spatial correlation using nearBynding across the other 101 proteins for which HepG2 eCLIP data is available in ENCODE. Sixteen proteins were significantly correlated with G3BP1 ($p < 0.0001$, Table 3.5, Fig. 3.7E). Of these, five bind G3BP1 (FXR2, IGF2BP3, LARP4, TIA1, and TIAL1) and three bind UPF1 (GRDWD1, PCBP1, and RPS3) based on BioGRID (101). Six are inversely correlated with HSU/PSU ratio in TCGA BRCA samples (BCLAF1, LARP4, PPIG, PRPF8, TIA1, and TIAL1), and ten are correlated with HSU/PSU ratio after controlling for G3BP1 expression (BCLAF1, FXR2, GRWD1, IGF2BP3, LARP4, PABPN1, PPIG, RPS3, TIA1, and TIAL1). Some of these proteins, such as TIA1, TIAL1, and AKAP1, show periodicity in their binding relative to G3BP1 whereas others, such as LARP4, BCLAF1, and FXR2, demonstrate a single binding peak at the same site as G3BP1. Additionally, the expression of twelve of the sixteen proteins that co-localize with G3BP1 also improve breast cancer survival, similar to G3BP1 ($p< 0.01$, AKAP1, BCLAF1, DDX55, GRWD1, IGF2BP3, LIN28B, PABPN1, PCBP1, RPS3, SUB1, TIA1, and TIAL1) (Fig. 3.7F).

Discussion

G3BP1 has been repeatedly shown to promote the migration and proliferation of many cancer types, including breast cancer [reviewed (92)]. Since G3BP1 has diverse roles in the cell, researchers are searching for anti-tumor drugs that target select domains or interactions of G3BP1 such as with USP10 (104) or, debatably, Ras-GAP (105,106). We show that the HSU/PSU ratio in BRCA samples is inversely associated with G3BP1 expression and survival, suggesting that targeting G3BP1's SRD activity may exacerbate breast cancer. But, G3BP1 has been shown to promote breast cancer proliferation by targeting specific anti-cancer mRNAs (91), suggesting that, although high G3BP1 expression is associated with better breast cancer

survival, selective inhibition of G3BP1's mRNA binding or decay activity may be a valuable therapeutic.

It is currently unclear how high G3BP1 expression is associated with better breast cancer survival but also with increased tumor proliferation and metastasis. One possibility is that aberrant splicing or polyadenylation, both of which are common in cancer and differ between metastatic and non-metastatic tumors (107–110), affect 3'UTRs and therefore SRD targets in a manner that we do not currently have the modeling to predict. Perhaps an increase in G3BP1 expression and SRD removes intratumor heterogeneity and thus drives immune recognition (111,112). There may also be an improved immune or chemotherapy response to the metastasizing tumors that improves survival (113–115). All of these hypotheses are highly speculative.

G3BP1 regulation by a miRNA to affect cancer progression has been previously shown (116), but not within breast cancer and not for any of the miRNAs identified in this study. miR-19a is part of miR-17–92, also called oncomir-1, one of the best-known oncogenic miRNA clusters (117). miR-19a has been shown to drive breast cancer proliferation and metastasis (118,119) and to be upregulated in AA breast cancer samples relative to EA samples in an independent study using laser capture microdissection (LCM) (120), though a prior proteomics study of miR-19a targets in breast cancer did not yield G3BP1 as a candidate (121). Increased expression of both miR-548j and miR-548o has been shown to be associated with breast cancer invasion and proliferation, respectively (122,123), but their role in these processes has not been associated with G3BP1. In contrast, literature evidence suggests that miR-98 and miR-138 negatively regulate breast cancer (124–128); it is possible that G3BP1 is not a true target of miR-98 and miR-138 in BRCA or that suppression of other targets mask the effect of G3BP1 depletion. Although their expression did not affect breast cancer survival within the tested population, miR-362 and let-7f have been shown to negatively regulate G3BP1 expression in

renal cell carcinomas and neuroendocrine tumors, respectively (129,130). Future work will repeat the analyses herein but separate by breast cancer subtype to determine whether there are subtype-specific miRNA effects.

There are many reasons that a protein's expression may correlate with HSU/PSU ratio, beyond having a role in SRD. For example, the protein may be involved in translation and preferentially stabilize or destabilize RNAs based on their structures, causing global RNA structure shifts upstream of the effects of SRD. Alternatively, the transcript of the correlated protein may be subject to SRD, so that the HSU/PSU ratio and protein expression correlation is confounded by SRD. Similarly, factors involved in SRD may be missed by correlating HSU/PSU ratio with protein expression, even after controlling for the expression of other known SRD factors. A protein's expression alone would not necessarily correlate with the HSU/PSU ratio if its involvement in SRD is dependent on a post-translational modification. If a protein has context-specific SRD effects or is redundant with other proteins that can perform the same role, it may not reproducibly appear correlated with HSU/PSU ratio across cell types or treatments. A protein's expression may also not correlate with HSU/PSU ratio if it is not rate-limiting for SRD; that is, if the protein is in large excess for what is needed within a cell or performs its role rapidly, then large fluctuations in its expression may not impact SRD and global RNA structure.

Although there are too many proteins that directly bind or have RNA-dependent interactions with G3BP1 and/or UPF1 to molecularly assess all of them for their role in SRD, many have known properties in literature that make them enticing candidates. For example, POP1, which binds both UPF1 and G3BP1, is a known endonuclease in the RNase P/MRP complex and has been shown to be involved in circRNA decay (102,131). In addition to binding G3BP1, FXR2 has been shown to drive mRNA decay of Noggin (132). A logistic regression model constructed based on RNA-seq, proteome, and CLIP-seq data identified G3BP1, FXR2, TIA1, TIAL1, LARP4, LIN28B, and PCBP1 as top informative RBPs for mRNA-protein

expression level discrepancies (133); nearBynding showed that all of these proteins colocalize with G3BP1 in 3'UTRs. These data will need to be confirmed in breast cancer cell lines, since spatial correlation was performed on eCLIP data from HepG2 cells. There may be alternative reasons for some of the protein binding and colocalization. For example, TIA1, TIAL1, PCBP1, IGF2BP3, FXR2, LIN28B, UPF1, and G3BP1 all localize to stress granules, so they may bind to the same transcript and one another to sequester rather than decay their targets (134).

A limitation of using nearBynding to identify other potential SRD factors is that eCLIP-seq– and therefore nearBynding– may not be able to capture some proteins, such as those that have highly transient binding or such as POP1, which cleaves bound RNA. For these reasons, we have identified and filtered candidate genes via multiple methods– literature evidence, effect on global RNA structure in multiple cell types, protein- or RNA-based interactions with known SRD factors, and colocalization with known SRD factors– to create a manageable list of factors to test for their potential role in SRD. Next steps to test whether the candidate genes identified do play a role in SRD would be to compare the expression of transcripts known to be SRD targets between WT and single or compound candidate gene knock-out– or, in the case of essential genes, candidate gene knock-down– cell lines, similar to (32).

Methods

Code availability

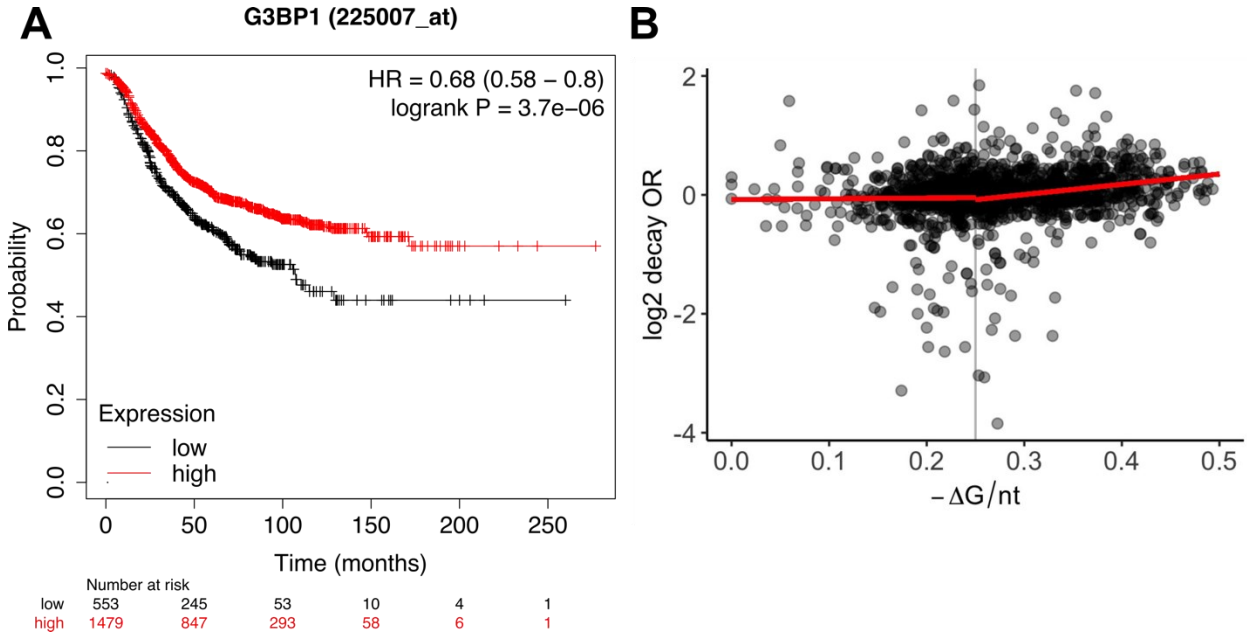
Code for all analyses and visualization is available in GitHub at https://github.com/vbusa1/RNA_structure_metric and was performed in R version 4.1.0 on a x86_64-apple-darwin17.0 (64-bit) platform. Table 3.6 contains a complete list of deposited data used in these analyses. Unpublished cultured breast cancer RNA-seq data was contributed by the lab of Dipali Sharma. Quantile normalized RSEM TCGA isoform counts of mRNA-seq and miR-seq were pulled from Firebrowse (http://firebrowse.org/).

Real-time quantitative PCR

Cultured breast cancer cells were pelleted and stored in TRIzol, and RNA was extracted via phenol-chloroform extraction. Samples were reverse transcribed with SuperScript VILO cDNA synthesis kit and 10 μL RT-qPCR reactions were performed in triplicate with 1:200 dilution cDNA, 1 μM primer mix, and 1x PowerUp SYBR Green master mix on a 7500 Fast Real-Time PCR System (Applied Biosystems). All primer sequences are derived from (32).

Figures and tables

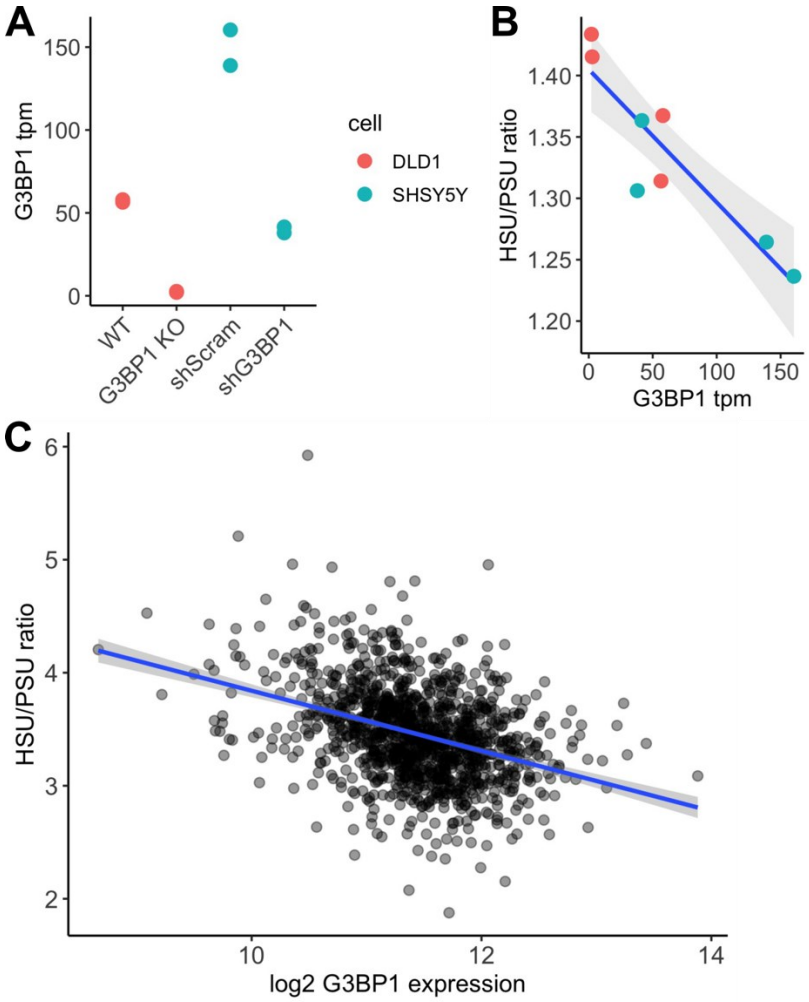
Figure 3.1 G3BP1 regulates much of the transcriptome



A. A survival curve comparing low (black) and high (red) mRNA expression of G3BP1, the epistatic protein of SRD, for breast cancer patients. G3BP1 protein expression demonstrates a similar survival trend (not shown). Figure and statistics were generated by Kaplan-Meier Plotter and the optimal G3BP1 microarray probe was selected by Jetset through Kaplan-Meier Plotter (95,135).

B. A scatter plot of transcripts' 3'UTR structure and how much the transcript's decay is affected by G3BP1, as measured by the $\log_2$ of the odds ratio between transcript levels at 0 and 16 hours post-actD treatment in WT versus G3BP1 KO DLD1 cells. The structure of transcripts' 3'UTRs based on -**Δ**G/nt is separated into poorly structured (< 0.25) and highly structured (> 0.25) by the vertical grey line. The red line is splined at -**Δ**G/nt = 0.25 and represents the linear models for PSUs and HSUs separately for 3'UTR structure versus the effect of G3BP1 on decay.

Figure 3.2 G3BP1 expression is a predictor of global RNA structure

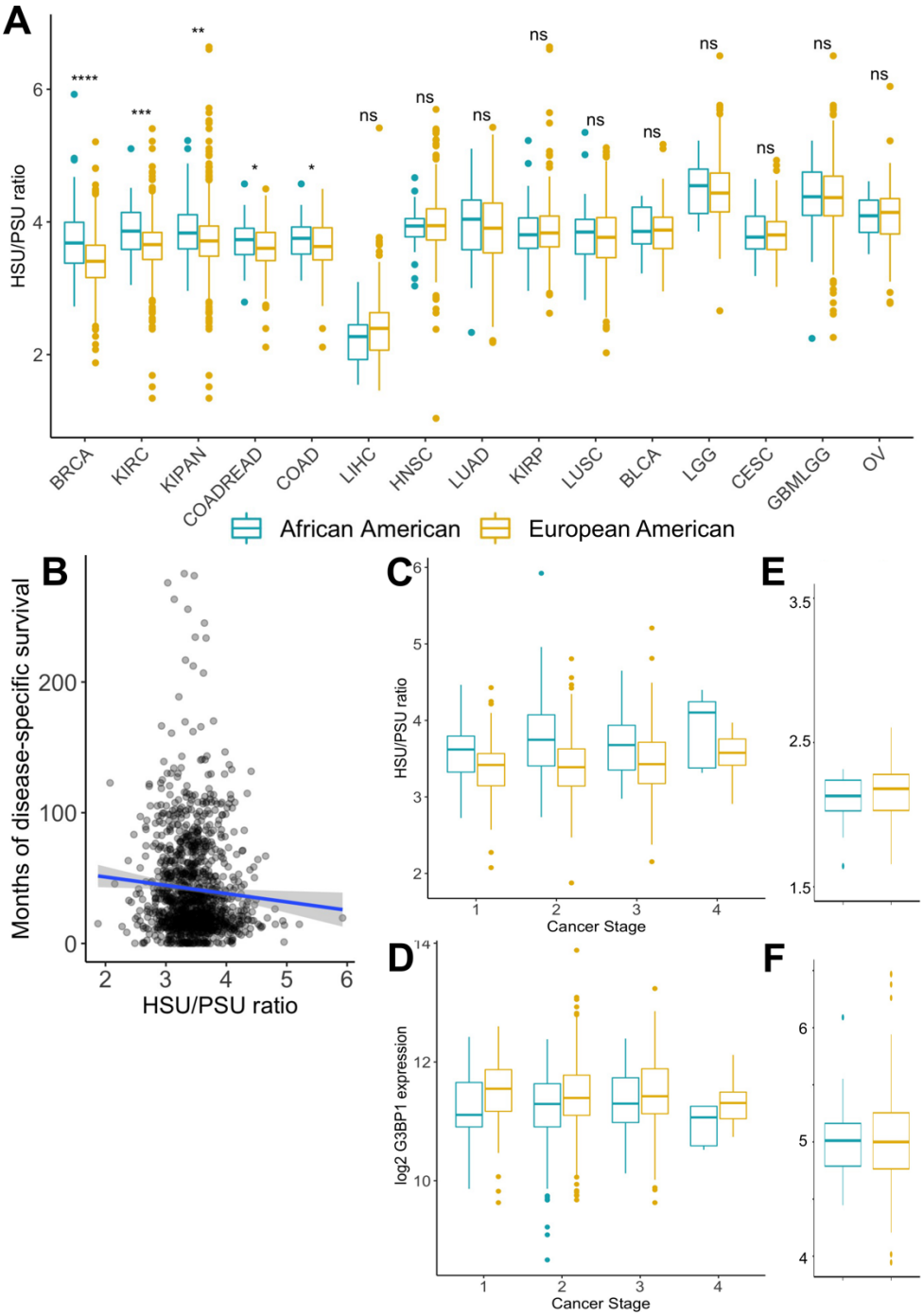


A. G3BP1 expression across cell culture lines assessed.

B. HSU/PSU ratio versus G3BP1 expression across cell culture lines assessed. The linear model is shown in blue with a 95% confidence interval.

C. HSU/PSU ratio versus G3BP1 expression for TCGA BRCA samples. The linear model is shown in blue with a 95% confidence interval.

Figure 3.3 Global RNA structure is often greater in African American cancer patients



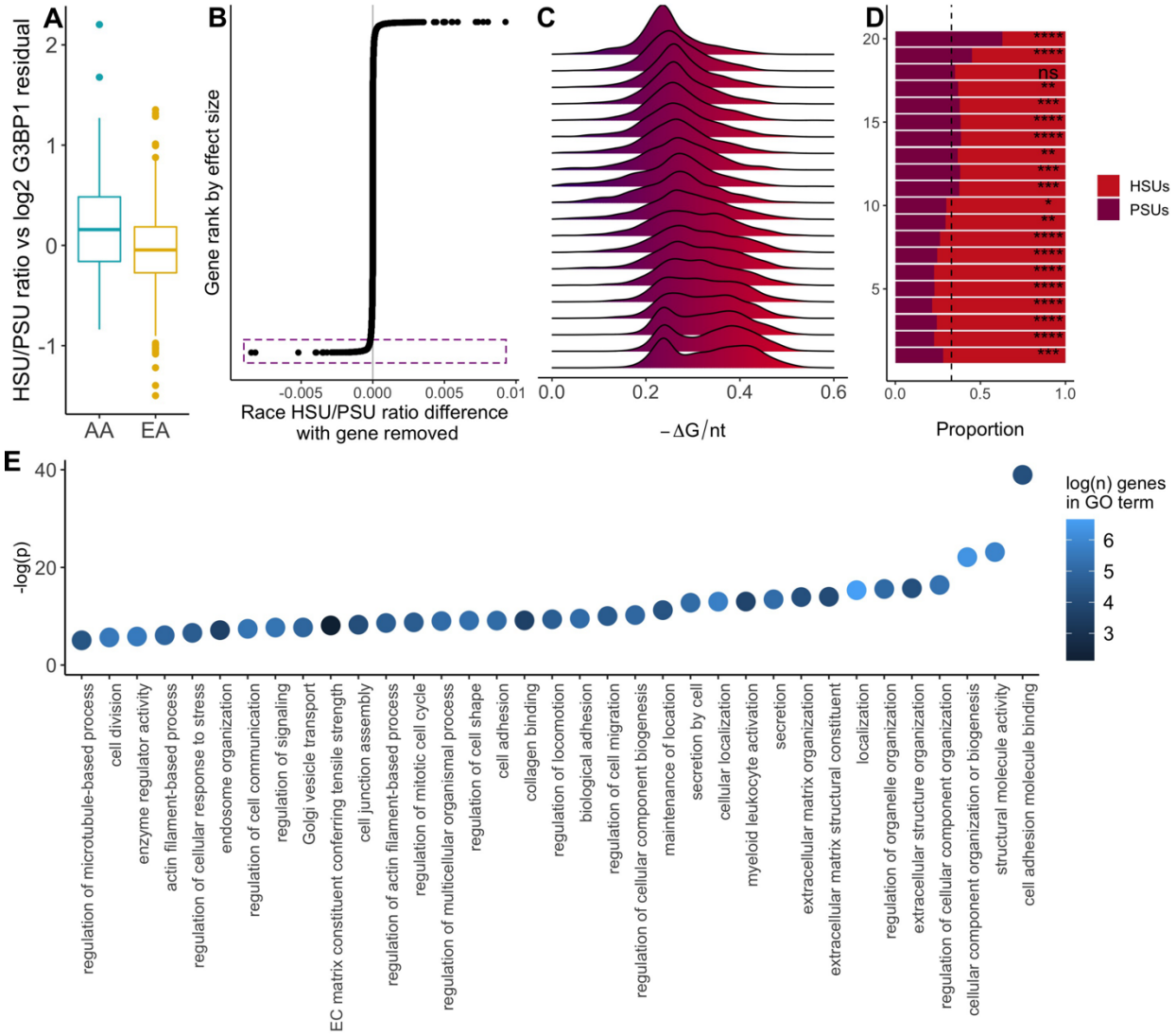
A. HSU/PSU ratios of TGCA samples from African American and European American patients across cancer types with at least 20 patients of each race. T-test p-values are shown as follows: ns p>0.05; * 0.05>p>0.01; ** 0.01>p>0.001; *** 0.001>p>0.0001; **** 0.0001>p.

BRCA breast invasive carcinoma; KIRC kidney renal cell carcinoma; KIPAN pan-kidney cancer; COADREAD colon adenocarcinoma and rectum adenocarcinoma; COAD colon adenocarcinoma; LIHC liver hepatocellular carcinoma; HNSC head and neck squamous cell carcinoma; KIRP kidney renal papillary cell carcinoma; LUSC lung squamous cell carcinoma; BLCA bladder urothelial carcinoma; LGG brain lower grade glioma; CESC cervical squamous cell carcinoma and endocervical adenocarcinoma; GBMLGG glioblastoma multiforme; OV ovarian carcinoma.

B and C. HSU/PSU ratio (B) and G3BP1 expression (C) of BRCA samples across cancer stages separated by race.

D and E. HSU/PSU ratio (D) and G3BP1 expression (E) of GTEx breast data separated by race.

Figure 3.4 Highly structured isoforms contribute disproportionately to race-based global RNA structure disparity

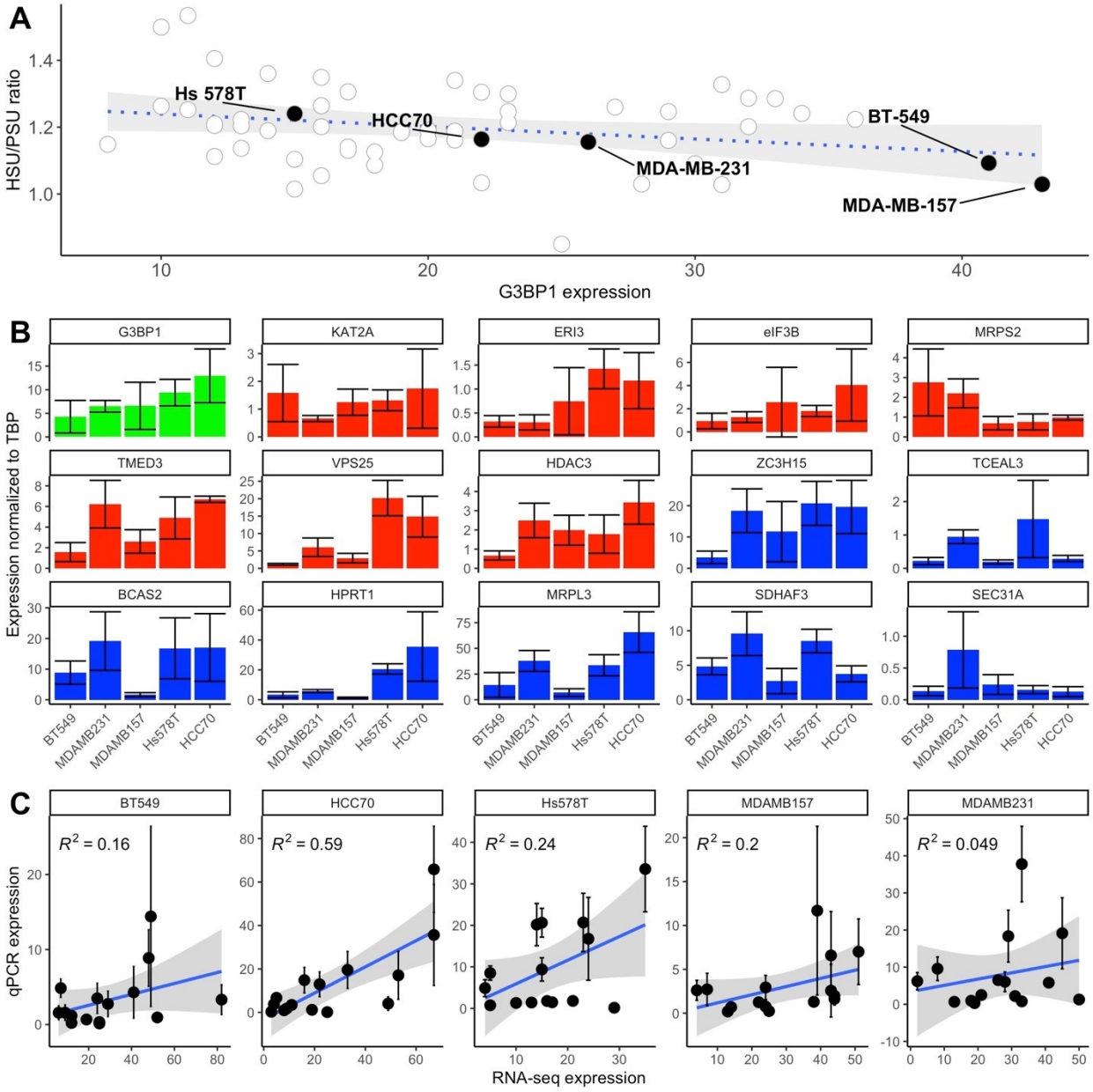


A. Ranked dot plot showing the effect on the difference in HSU/PSU ratio between AA and EA BRCA samples of removing each isoform. Isoforms that most contribute to the HSU/PSU ratio difference are boxed.

B. Density plots of isoform structure, separated into 20 bins based on their effect size in (A). Structured isoforms are shown in red, while isoforms with less structure are purple.

C. Bar plot of isoforms binned as in (B), separated into HSUs and PSUs. The dotted line shows 0.33, the proportion of PSUs across all isoforms.

Figure 3.5 Cultured breast cancer samples demonstrate poor reproducibility

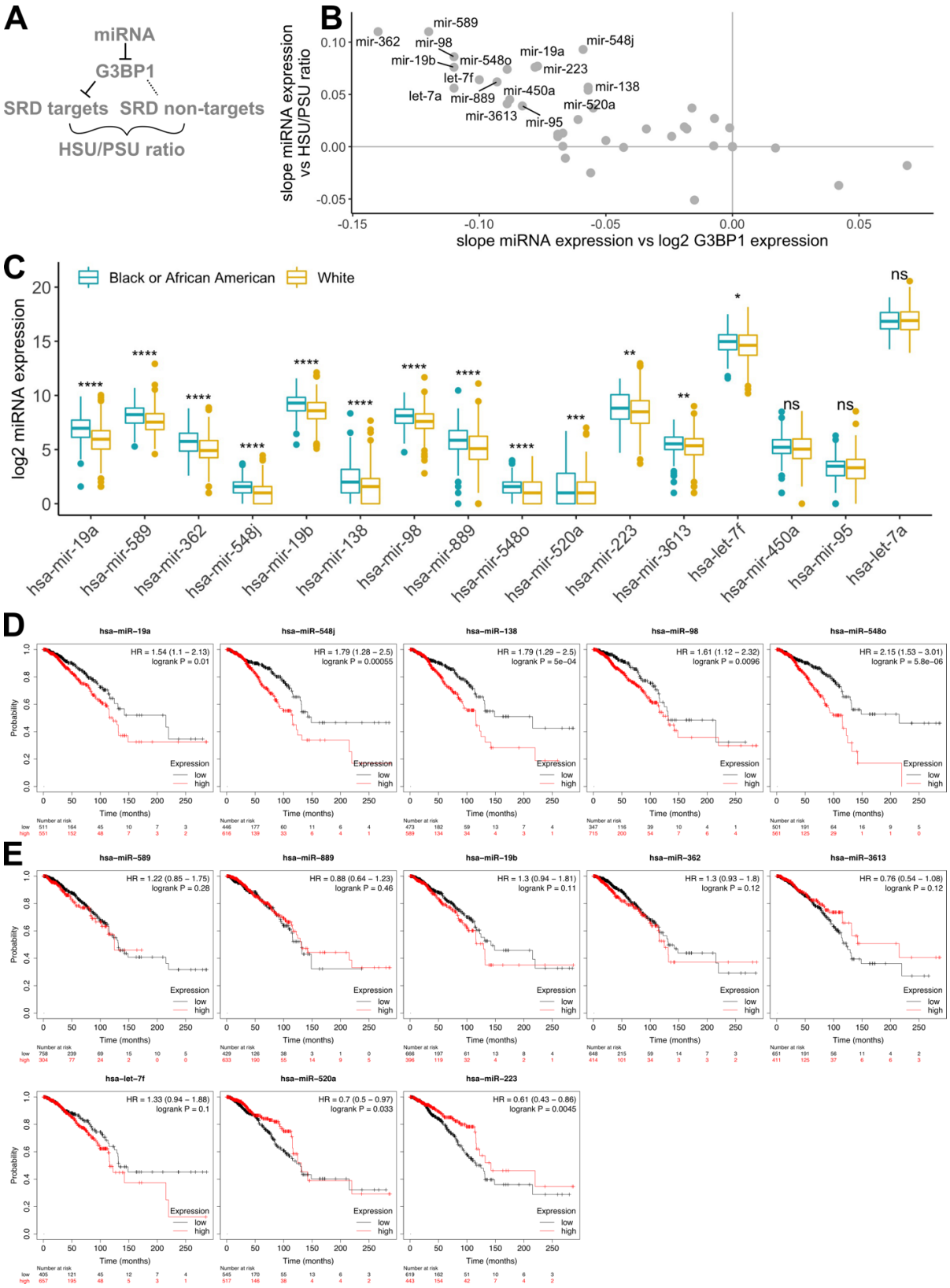


A. G3BP1 expression and HSU/PSU ratio derived from RNA-seq for 55 cultured breast cancer cell lines. The linear model is in blue, with a 95% confidence interval in grey.

B. Gene expression for G3BP1 (green), seven HSUs that are known SRD targets (red), and seven PSUs (blue) assessed by qPCR and normalized to TBP expression for five cultured breast cancer samples. Error bars represent the standard error of three replicates.

C. Gene expression for the fifteen genes shown in (B) measured via RNA-seq (x-axis, FPKM) and via qPCR (y-axis, normalized to TBP expression in triplicate). The linear model is in blue, with a 95% confidence interval in grey.

Figure 3.6 Race-based differences in miRNA expression may drive global RNA structure and cancer prognosis differences



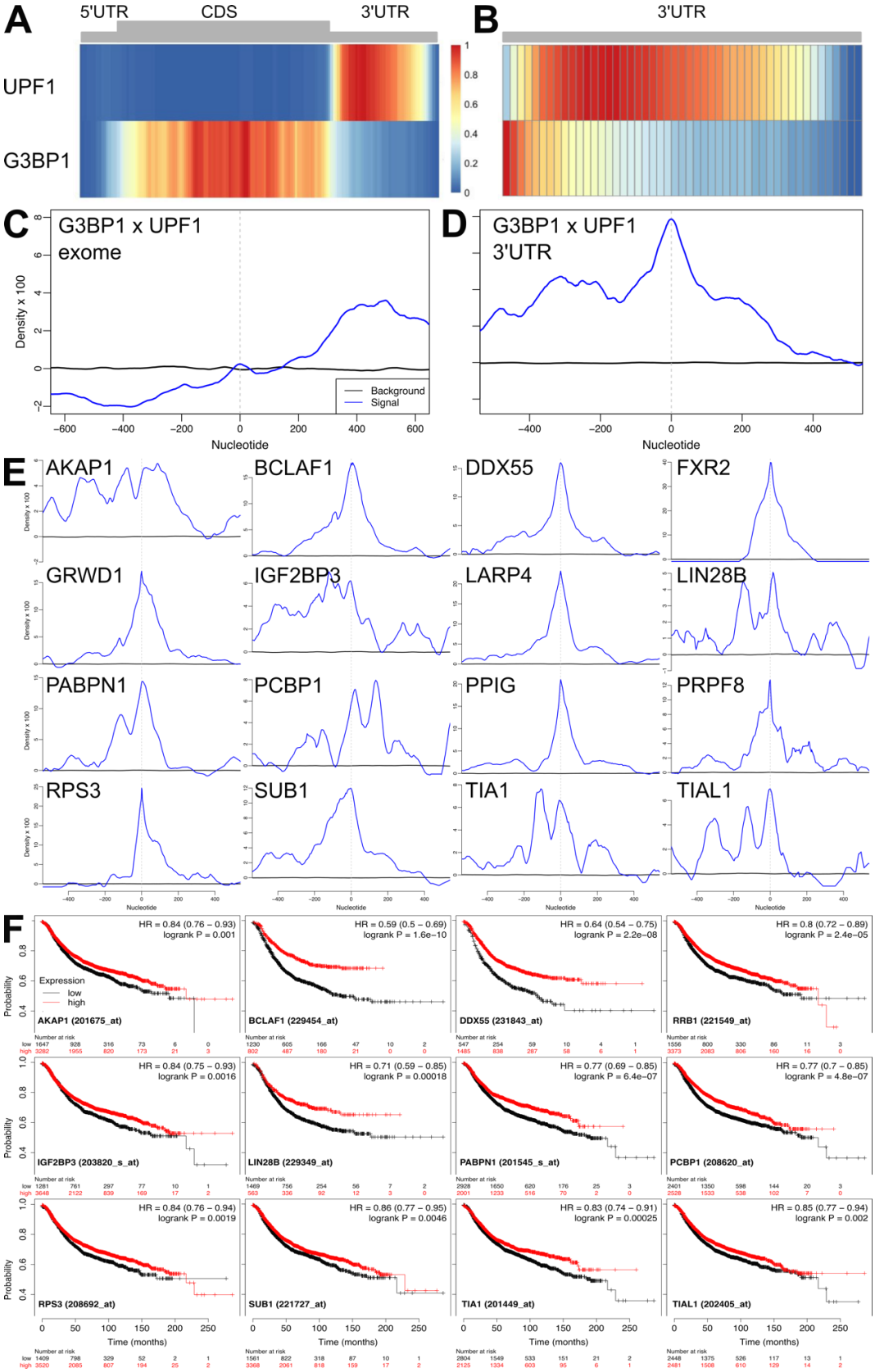
A. Model for how miRNA expression may affect SRD and how this affect might be measured by the HSU/PSU ratio.

B. X-axis values reflect the estimated slope from linear models of the $\log_2$ expression of each miRNA predicted to target G3BP1 for which there was sufficient data versus $\log_2$ G3BP1 expression across TCGA BRCA samples. Y-axis values reflect the estimated slope from linear models of the $\log_2$ expression of each miRNA predicted to target G3BP1 for which there was sufficient data versus the HSU/PSU ratio across TCGA BRCA samples. miRNAs for which both linear models are significant ($p<0.01$) are colored blue and labelled.

C. miRNA expression in TCGA BRCA samples separated by race. T-test p-values are shown as follows: ns $p>0.05$; * $0.05>p>0.01$; ** $0.01>p>0.001$; *** $0.001>p>0.0001$; **** $0.0001>p$.

D. Survival curves comparing low (black) and high (red) expression of miRNAs that correlate with a significant decrease in survivability among TCGA breast cancer patients. From left to right, these miRNAs are: miR-19a, miR-548j, miR-138, miR-98, and miR-548o. Figures and statistics were generated by Kaplan-Meier Plotter (100).

Figure 3.7 nearBynding identifies proteins that colocalize with G3BP1



A. Normalized UPF1 and G3BP1 peak density along binned lengths of all mRNAs from (50).

B. Normalized UPF1 and G3BP1 peak density along binned lengths of 3'UTRs from (50).

C. nearBynding spatial correlation of G3BP1 and UPF1 binding within the HepG2 exome.

D. nearBynding spatial correlation of G3BP1 and UPF1 binding within HepG2 3'UTRs.

E. nearBynding spatial correlation of G3BP1 and sixteen proteins that significantly colocalize to G3BP1 (Table 3.5).

Table 3.1 Coefficients of linear models assessing HSU/PSU ratio versus $\log_2$ G3BP1 expression for all TCGA cancer types

Cancer	# samples	p-value	R^2	slope
BRCA	1212	1.20E-41	0.14	-0.27
KIPAN	1020	3.70E-39	0.16	-0.19
GBMLGG	701	8.50E-36	0.2	-0.54
KIRC	606	1.10E-25	0.17	-0.31
LUAD	576	2.50E-22	0.15	-0.38
HNSC	566	1.60E-12	0.085	-0.23
LUSC	552	1.50E-26	0.19	-0.41
LGG	530	1.70E-26	0.19	-0.45
COADREAD	433	3.50E-07	0.058	-0.19
BLCA	427	9.20E-12	0.1	-0.15
LIHC	423	0.00016	0.033	-0.16
COAD	328	1.10E-07	0.083	-0.24
KIRP	323	1.30E-11	0.13	-0.24
CESC	309	0.01	0.021	-0.1
OV	307	0.00085	0.036	-0.17
ESCA	196	0.43	0.0032	0.043
PCPG	187	3.90E-07	0.13	-0.32
PAAD	183	0.058	0.02	-0.17
LAML	173	1.80E-15	0.31	-0.59
GBM	171	7.70E-07	0.13	-0.49
KICH	91	3.40E-06	0.22	-0.45
MESO	87	0.14	0.025	-0.13
ACC	79	0.81	0.00078	0.011

DLBC	48	0.72	0.0029	0.028
CHOL	45	0.74	0.0025	0.089

BRCA breast invasive carcinoma; KIPAN pan-kidney cancer; GBMLGG glioblastoma multiforme; KIRC kidney renal cell carcinoma; LUAD lung adenocarcinoma; HNSC head and neck squamous cell carcinoma; LUSC lung squamous cell carcinoma; LGG brain lower grade glioma; COADREAD colon adenocarcinoma and rectum adenocarcinoma; BLCA bladder urothelial carcinoma; LIHC liver hepatocellular carcinoma; KIRP kidney renal papillary cell carcinoma; COAD colon adenocarcinoma; KIRP kidney renal papillary cell carcinoma; CESC cervical squamous cell carcinoma and endocervical adenocarcinoma; OV ovarian carcinoma; ESCA esophageal carcinoma; PCPG pheochromocytoma and paraganglioma (adrenal gland); PAAD pancreatic adenoma; LAML acute myeloid leukemia; GBM glioblastoma multiforme; KICH kidney chromophobe; MESO mesothelioma; ACC adrenocortical carcinoma; DLBC lymphoid neoplasm diffuse large B-cell lymphoma; CHOL cholangiocarcinoma (bile duct).

Table 3.2 Top 10 hits HSU/PSU ~ $\log_2$(gene expression)

Gene	p-value	Slope
GTF2A1	4.50E-107	-0.47
DDX6	5.30E-107	-0.45
BCLAF1	2.00E-106	-0.45
TMEM167A	6.70E-106	-0.5
TNPO1	1.40E-105	-0.45
APC	3.50E-105	-0.38
ARL5A	5.80E-105	-0.45
PAPOLG	8.30E-103	-0.42
COPS2	1.10E-102	-0.48
EXOC5	3.90E-102	-0.4
G3BP1	3.10E-51	-0.31
G3BP2	2.10E-58	-0.3

Table 3.3 Top 10 hits HSU/PSU ~ $\log_2$(gene expression) + $\log_2$(G3BP1)

Gene	p-value	Slope
ZNF628	1.30E-132	0.35

TELO2	6.10E-132	0.31
TSSC4	1.00E-122	0.34
ZGPAT	1.20E-120	0.3
BRAT1	1.70E-119	0.38
SCAF1	8.50E-116	0.39
EPN1	1.00E-114	0.39
SPHK2	3.30E-114	0.36
POLRMT	3.60E-114	0.4
CACTIN	6.60E-114	0.38
UPF1	6.90E-52	0.39
G3BP2	7.20E-24	-0.21

Table 3.4 Breast cancer cell lines tested

Line	Race	Type
HCC 70	African American	Ductal carcinoma
MDA-MB-157	African American	Carcinoma
BT-549	European American	Invasive ductal carcinoma
Hs 578T	European American	Invasive ductal carcinoma
MDA-MB-231	European American	Adenocarcinoma

Table 3.5 Proteins that co-localize with G3BP1 in HepG2 3'UTRs

Protein	nWindows	p-value
AKAP1	80	3.64E-11
BCLAF1	47	3.16E-07
DDX55	71	6.04E-12
FXR2	14	1.72E-05
GRWD1	53	1.62E-10
IGF2BP3	43	3.39E-07
LARP4	91	1.96E-23
LIN28B	49	1.42E-05
PABPN1	45	3.70E-05
PCBP1	38	7.62E-05
PPIG	39	1.86E-05
PRPF8	39	4.25E-06
RPS3	35	1.22E-09
SUB1	92	3.74E-18

TIA1	54	5.25E-07
TIAL1	50	8.85E-05
UPF1	104	1.85E-12

Table 3.6 Key Resources

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
The Cancer Genome Atlas (TCGA)		https://www.cancer.gov/tcga
mirDB	(99)	http://mirdb.org/
Genotype Tissue Expression Project (GTEx)	(136)	http://www.gtexportal.org/home/ dbGaP accession number phs000424.v8.p2
Global Analysis of All 3′ UTRs	(32)	Table S1
RNA-seq of cell culture	(32)	NCBI BioProject: PRJNA591294
ENCODE eCLIP	(49)	https://www.encodeproject.org/
BioGRID	(101)	https://thebiogrid.org/
Software and algorithms		
nearBynding	This work (82)	https://bioconductor.org/packages/nearBynding/ ; https://github.com/vbusa1/nearBynding
Kaplan-Meier Plotter	(100)	https://kmplot.com/analysis/
Experimental models: Cell lines		
BT-549		
HCC70		
Hs 578T		
MDA-MB-157		
MDA-MB-231		
Critical commercial assays and chemicals		
PowerUp SYBR Green master mix	Thermo Fisher	cat #A25742
SuperScriptTM IV VILO Master Mix	Invitrogen	cat #11754050
TRIzol reagent	Invitrogen	cat # 15596026

Chapter 4: Transcriptome regulation by PARP13 in basal and antiviral states

Highlights

- PARP13 regulates the transcriptome in basal and antiviral states.
- Antiviral response shifts PARP13 target localization but not its binding preference.
- PARP13 supports the expression of ISGylation-related genes, including TRIM25.
- PARP13 and TRIM25 associate in part via RNA and share common targets.

Introduction

The innate immune response is the first-line defense against viral pathogens. To confer a broad defense, components of the innate immune response target conserved aspects of pathogen structure, invasion, and proliferation. PARP13, also known as ZAP and coded by the ZC3HAV1 gene, contributes to host defense against a plethora of RNA and DNA viruses, including members of *Togaviridae* (e.g. sinbus virus), *Coronaviridae* (e.g. SARS-CoV-2), *Herpesviridae* (e.g. herpes simplex), and *Retroviridae* (e.g. human immunodeficiency virus) [reviewed (137)]. Upon viral infection, PARP13 activates innate immune pathways through interaction with the viral RNA-sensing protein RIG-I and improving its association with downstream effectors (138). PARP13 directly binds target viral RNA (139,140) and can mediate both translation repression of bound RNAs by interfering with initiation (141) and mRNA decay by recruiting the exosome and associated mRNA degradation machinery (142–145). PARP13-mediated decay requires translational repression, but translational repression can occur without degradation of target mRNAs (141).

Humans have four PARP13 isoforms, with two being most abundant: the longer common isoform, PARP13.1, contains a catalytically inactive but antivirally important ADP-ribosyltransferase domain and membrane-localization CaaX motif at the C-terminus, both of

which are absent in the shorter, cytoplasmic isoform, PARP13.2 (146–149). PARP13.1 is constitutively expressed and localized to endoplasmic reticulum (ER), and both isoforms are induced by viral infection or type I interferons (IFN), with PARP13.2 more upregulated than PARP13.1 (138,150). Both isoforms share an N-terminus that contains four zinc finger domains; these domains preferentially bind CG-rich viral RNA (151,152) and are necessary for PARP13 antiviral activity (153,154). PARP13 also directly regulates the decay of human TRAILR4 mRNA through binding to its 3'UTR (155,156). Although PARP13 binds host RNAs, as shown by cross-linking experiments (155,156), the identities of genome-wide PARP13 host targets are not known.

PARP13 interacts with many proteins encoded by IFN-stimulated genes (ISGs) to perform its antiviral activities, such as activation of the pathway mediated by ISGylation—the conjugation of the ubiquitin-like protein, interferon-stimulating gene-15 (ISG15), onto proteins post-translationally (157–160). In particular, TRIM25, an E3 ligase of ISG15, is the primary cofactor of PARP13 and is necessary for PARP13's role in the immune response (160–162). TRIM25 binding to PARP13 requires the SPRY domain and N-terminus, respectively (160,161,163). Antiviral activity requires TRIM25 and PARP13 colocalization at membranes RNA-binding domains of both proteins (147,153,154,160,163,164). Ablation of TRIM25's putative RNA-binding domain (RBD) in its SPRY domain decreases PARP13 interaction (163), suggesting the interaction may be mediated via protein-RNA interaction and indicating the possibility of co-regulation of shared targets.

Here we integrated RNA-seq and CLIP-seq data to assess the role of PARP13 expression and RNA binding in transcriptome regulation in human cells. We explored transcriptomic signatures that correspond to basal PARP13 activity as well as shifts that occur exclusively in the presence of an antiviral response. PARP13 autoregulates its own long-isoform in basal states and contributes to the maintenance of the expression of transcripts essential for

the ISGylation response. We also demonstrate that TRIM25 and PARP13 bind proximally on shared transcripts using spatial correlation analysis and that the TRIM25:PARP13 complex is stabilized by protein-RNA interactions.

Results

PARP13 regulates cellular homeostasis and the innate antiviral response

To study the role of PARP13 in basal and antiviral states, we performed RNA-seq on wild-type (WT) and PARP13 knock-out (KO) HEK293T cells treated with triphosphorylated single-stranded RNA (3p-RNA) as an RNA virus mimic, or a single-stranded (ss)RNA of the same length that elicits no observable immune response (138). WT cells demonstrated a strong, unidirectional transcriptomic shift upon 3p-RNA treatment versus ssRNA treatment, with many of the most differentially-expressed genes associated with the innate immune response ($\log_2$ fold-change > 1 and $p < 0.01$, Fig. 4.1A). In contrast, PARP13 KO cells treated with 3p-RNA do not demonstrate a strong transcriptomic difference compared to ssRNA treatment, indicating the important of PARP13 in immune responses (Fig. 4.1B). To interrogate the role of PARP13 in the absence of an immune response, we compared ssRNA-treated WT and PARP13 KO cell expression and observed 530 differentially-expressed genes (Fig. 4.1C). Upon 3p-RNA treatment, 806 differentially-expressed genes were observed in WT versus KO cells (Fig. 4.1D). Consistent with the published result that TRAILR4 is upregulated in PARP13 KO cells (155), we also observed a 5-fold increase in TRAILR4 expression in PARP13 KO cells by qPCR (Fig. 4.1E). Similarly, our RNA-seq data showed a 3.5-fold TRAILR4 increase in ssRNA-treated cells and a 7-fold increase in 3p-RNA-treated cells (Fig. 4.1C).

138 differentially-expressed genes of 3p-RNA-treated WT versus PARP13 KO cells (Fig 4.1D) are also differentially-expressed genes in 3p-RNA versus ssRNA treatment in WT cells (Fig.4.1A), and 388 differentially-expressed genes of 3p-RNA-treated WT versus PARP13 KO cells are also differentially-expressed genes in ssRNA-treated WT versus PARP13 KO (Fig.

4.1C); the large number of overlapping differentially-expressed genes reveals that there are multiple simultaneous biologically-important signals in our data that make simple pairwise comparisons of differentially-expressed genes insufficient. We performed independent component analysis (ICA) across all RNA-seq samples to deconvolute simultaneous transcriptomic programs. Three independent components were adequate to separate samples into groups of anticipated biological relevance, suggesting that the biological signals of these grouping are robust and that replicates had comparatively minor technical variation (Fig. 4.1F). We were able to identify genes that have consistent expression across all samples (IC1), genes that are differentially expressed between WT and PARP13 KO samples, regardless of treatment (IC2), and genes that are only differentially expressed in WT cells treated with 3p-RNA (IC3).

To investigate the functions of genes that drive IC2, we ordered genes by their weight in this individual component and performed gene ontology (GO) analysis. Transcripts persistently differentially expressed between WT and PARP13 KO samples are generally associated with the endoplasmic reticulum (ER) and involved in stress responses (Fig. 4.1G), which is consistent with the membrane localization and constitutive expression of PARP13.1 and recapitulates transcriptomic differences previously shown between untreated WT and PARP13 KO cells (147,148,155). GO analysis of ranked IC3 gene weights showed that genes only upregulated in WT cells treated with 3p-RNA are generally involved in the innate antiviral response (Fig. 4.1H), which is consistent with the well-documented role of PARP13, where 3p-RNA-treated samples phenocopy an antiviral state [reviewed (137)].

PARP13 suppresses PARP13.1 but not PARP13.2 translation in the absence of an antiviral response

To study the role of PARP13 RNA-binding, we performed PARP13 CLIP-seq (165) on WT and PARP13 KO HEK293T cells treated with 3p-RNA or ssRNA. To globally characterize PARP13's RNA binding preferences across treatments, we identified peaks in the PARP13

CLIP-seq data for ssRNA- and 3p-RNA-treated samples. ssRNA-treated samples showed 1,405 peaks across 730 genes, and 3p-RNA-treated samples showed 1,950 peaks across 1,013 genes. The majority of PARP13 peaks mapped to coding regions and the 3'UTR of mRNAs, with little difference between treatments (χ^2 p = 0.03, Fig. 4.2A).

To verify CLIP-seq identified targets, we performed qPCR on the RNA of the immunoprecipitates of GFP-PARP13.1 (Fig. 4.2B). The negative control target transcript, HDAC3, showed no enrichment in the GFP-PARP13.1 pull-down over input, whereas the PARP13 target transcripts—PARP13, ATP6AP1, EIF1, SLC3A2, STT3A, TMED3, and XBP1—that were identified in both ssRNA and 3p-RNA treatments by the CLIP data did. Interestingly, PARP13.1 mRNA was enriched 40-fold more than PARP13.2 mRNA, indicating that the PARP13 protein does not bind both PARP13 isoform transcripts to the same extent.

We further characterized how PARP13 regulates its own transcripts. The RNA-seq data demonstrated that both isoforms were upregulated upon mock-viral infection, but PARP13.1 showed 4.4-fold increase in expression based on aligned reads within its specific 3'UTR while PARP13.2 showed a 31.2-fold increase in reads based on its 3'UTR, which is consistent with prior publications (Fig. 4.2C, top) (138). In contrast, CLIP-seq shows an enrichment for PARP13 binding within the PARP13.1 3'UTR but not within the PARP13.2 3'UTR over background in both treatments (Fig. 4.2C, middle), which corresponds to the differential enrichment of PARP13 mRNAs seen in our RIP-qPCR results (Fig. 4.2B).

Since these two isoforms contain different 3'UTRs, we used a dual-reporter luciferase construct to assess the effect of PARP13 binding on these UTRs. We cloned the PARP13.1 or PARP13.2 3'UTR downstream of the *Renilla* luciferase gene, with the firefly luciferase under an independent promoter as a control (Fig. 4.2D) (32). By using WT and PARP13 KO cells, our dual luciferase system allowed us to assess the effect of PARP13 expression on the respective 3’UTRs at the mRNA and protein levels. We co-transfected the PARP13.1 or PARP13.2 3'UTR

luciferase constructs into WT and PARP13 KO HEK293T cells along with either ssRNA or 3p-RNA; 24 hours later, we assessed the transcript expression by qPCR and luminescence of the *Renilla* luciferase and normalized with firefly luciferase expression. We observed that mRNA levels for luciferase constructs with either the PARP13.1 or PARP13.2 3'UTRs were the same in WT and PARP13 KO cells for both treatments (Fig. 4.2E), suggesting that PARP13 binding to the PARP13.1 3'UTR does not drive its mRNA decay. Luminescence—and therefore protein levels—for the PARP13.2 3'UTR luciferase construct are the same in WT and PARP13 KO cells for both treatments (Fig. 4.2F), suggesting the PARP13.2 isoform is not subject to PARP13 regulation. The lack of PARP13-dependent shift in protein expression of the PARP13.2 3'UTR construct is consistent with our CLIP data which showed that PARP13 does not bind the PARP13.2 3'UTR above background (Fig. 4.2C). In contrast, protein levels for the PARP13.1 3'UTR construct are higher in PARP13 KO cells with ssRNA treatment but the same in WT and PARP13 KO cells with the 3p-RNA treatment (Fig. 4.2F). The treatment-specific protein inhibition suggests that PARP13 binds the 3'UTR and inhibits PARP13.1 translation at basal state, but such inhibition is not observed upon antiviral response.

The antiviral response shifts PARP13 target localization but not its binding preferences

Based on our CLIP-seq data, 518 PARP13 targets were common across samples, but 212 PARP13 targets were unique to ssRNA-treated cells and 495 targets were unique to 3p-RNA-treated cells (Fig. 4.3A). To determine whether the shift in PARP13 targets is attributable to differences in binding sequence preference, we calculated the enrichment of dinucleotides within PARP13 binding peaks, normalized to the dinucleotide frequency of all expressed transcripts. There was no shift in dinucleotide preference between treatments, with CG substantially enriched within PARP13-bound regions (Fig. 4.3B). This CG binding preference is congruent with prior evidence from viral RNA studies (151,152). Identification of hexameric motifs within peaks also demonstrated CG enrichment and consistent binding preference across

conditions (Fig. 4.3C). The invariability of the binding motif across treatments can be explained by the fact that PARP13.1 and PARP13.2 share all RNA binding domains.

PARP13.1 is membrane-associated and available to bind transcripts in all states, whereas PARP13.2 is cytoplasmic and only constitutes a sizeable proportion of the PARP13 population upon +3p-RNA treatment (138,147,148). Therefore, we expect that there may be difference in the localization of the transcription. Given that PARP13.1 is associated with ER-derived membranes and the RNA-seq data analyses implicated PARP13 in ER regulation (Fig. 4.1G) (155), we explored the relationship with transcript localization to the ER using available data from APEX-RIP, which used engineered ascorbate peroxidase-catalyzed proximity biotinylation of endogenous proteins and RNA immunoprecipitation to isolate transcripts localized to a variety of subcellular compartments including the ER (166). Transcripts that are bound by PARP13 in both treatments are enriched in the ER (Fisher's exact test $p= 2.96 \times 10^{-132}$). In contrast, transcripts that are only PARP13-bound in 3p-RNA treatment are enriched in the cytoplasm ($p=1.04 \times 10^{-69}$, Fig. 4.3D). This shift in localization may be explained by the constitutive expression of PARP13.1 versus the upregulation of PARP13.2 upon viral response.

PARP13 supports expression of transcripts related to the ISGylation and RIG-I pathway

Beside focusing on directly bound PARP13 targets, we are also interested which genes are regulated in a PARP13-dependent manner and upregulated for antiviral responses. Based on our RNA-seq data, twelve genes were downregulated in PARP13 KO cells compared to WT cells at basal states and up-regulated upon 3p-RNA treatment in WT cells. Of these, five are involved in the process for the post-translational addition of ISG15—a ubiquitin-like protein that has broad antiviral effects and whose expression is initiated by RIG-I-activated IRFs and NFκB [reviewed (167)]. Notably, ISG15, its E2 ligase UBE2L6, E3 ligase TRIM25, and its downstream activator RIG-I have all been shown to enhance the ability of PARP13 to mediate antiviral activities (159). GO analysis of IC3, a signal corresponding to genes that are only differentially

expressed in WT cells treated with 3p-RNA, further demonstrated an enrichment for genes associated with ISG15 protein conjugation (Fig. 4.1F and H). Given that PARP13 also binds the TRIM25, RIG-I, and unconjugated ISG15 proteins (Fig. 4.4A) (138,160,161,163,168,169), we aimed to directly assess the effect of PARP13 expression on the ISGylation pathway.

Amongst genes involved in ISGylation, UBA7 and HERC5 had the same expression between WT and PARP13 KO cells in control treatment. qPCR of untreated WT and PARP13 KO HEK293T cells recapitulated the RNA-seq finding, with UBE2L6, RIG-I, ISG15, USP18, USP41 and TRIM25, but not HERC5, down-regulated in PARP13 KO cells (Fig. 4.4B). CLIP analyses identified that PARP13 binds to an intron of TRIM25 and the CDS of ISG15 upon 3p-RNA treatment. Corresponding to the reduced transcript expression observed, TRIM25 protein expression was reduced ~50% in PARP13 KO cells (Fig. 4.4C).

To exclude the possibility of clonal selection during the generation of PARP13 KO cells causing the observed ISGylation factor downregulation, we performed a PARP13 knockdown time course using siRNAs in WT HEK293T cells. RIG-I did not demonstrate PARP13 expression-dependent downregulation within the 60-hour time course, but PARP13 expression decreased within 12 hours post-siPARP13 treatment, along with downregulation of ISG15, TRIM25, UBE2L6, USP18, and USP41 by 36 hours post-treatment (Fig. 4.4D). These data suggest that PARP13 expression is responsible for the steady-state reduction of ISGylation gene expression observed in PARP13 KO cells.

Binding to common transcripts reinforces TRIM25:PARP13 interactions

The ISG15 E3 ligase TRIM25 is also a primary co-factor of PARP13 for immune regulation. Both proteins require their RNA-binding domains to perform their roles in the innate antiviral response (153,154,164). To interrogate whether the requirement of these proteins for RNA binding is related to their interaction with one another, we used CLIP-seq data to identify

their shared targets of TRIM25 and PARP13. Although no datasets of TRIM25 are published for 293T cells, comparison with CLIP-seq data from HeLa cells transiently expressing T7-tagged TRIM25 (163) revealed that 87% of all identified PARP13 targets were also bound by TRIM25, which is far greater than could be expected by chance (Fig. 4.5A, Fisher's exact test, $p = 1.11 \times 10^{-168}$). There is not a significant shift in the proportion of PARP13 targets across treatment conditions that are bound by TRIM25 (Fig. 4.5B).

Although TRIM25 shares many targets with PARP13, it was unclear whether the proteins bind in proximity to each other. Using the transcriptome-wide cross-correlation tool nearBynding (82), we spatially compared the binding peaks of PARP13 and TRIM25 to determine the location of PARP13 binding relative to TRIM25. TRIM25-bound regions positively correlated with PARP13-bound regions identified for both ssRNA and 3p-RNA treatments at relative position 0, which shows that when PARP13 and TRIM25 bind the same transcript, they bind at approximately the same location (Fig. 4.5C).

In addition, there were periodic intervals of positive spatial correlation both upstream and downstream of this primary site of colocalization, especially for correlation of TRIM25 with ssRNA PARP13 samples (Fig. 4.5C, left). This phenomenon could be caused by 1) having multiple TRIM25 binding sites located upstream and/or downstream of a single PARP13 binding site at regular intervals, 2) having multiple PARP13 sites upstream and downstream of a single TRIM25 site, or 3) a mixture of these two scenarios (Fig. 4.5D). To distinguish between these binding geometry possibilities, we auto-correlated the TRIM25 and PARP13 peak data. Auto-correlation of TRIM25-bound regions shows a single peak, suggesting that TRIM25 does not regularly bind a set distance away from itself upstream or downstream (Fig. 4.5E). In contrast, auto-correlation of PARP13-bound regions for both ssRNA and 3p-RNA treatment shows small peaks upstream and downstream of the central signal, signifying that PARP13 tends to bind near itself at set intervals (Fig. 4.5F and G). Previously-published PARP13 CLIP data also show

that PARP13 binds various viral transcripts at ~200 nt intervals (152,156), which suggests that PARP13 often occupies multiple, regularly spaced sites. Therefore, the periodicity observed in the TRIM25 and PARP13 correlation is generally due to a single TRIM25 molecule binding near multiple PARP13 molecules (model 2 of Fig. 4.5D).

To assess the degree to which PARP13 RNA binding contributes to its well-documented association with TRIM25 (160,161,163), we performed co-immunoprecipitation experiments. We transfected GFP-tagged PARP13.1, PARP13.2, and PARP13.2(Q) mutant, which is RNA-binding-deficient due to quintuple mutations (155), into WT HEK293T cells and immunoprecipitated using GFP antibodies. Consistent with prior literature, PARP13.1 bound TRIM25 better than PARP13.2 (Fig. 4.5H) (147,160). The RNA-binding-deficient PARP13.2(Q) construct bound TRIM25 less effectively than PARP13.2, suggesting that RNA binding does contribute to the interaction of PARP13 and TRIM25 (Fig. 4.5H). However, PARP13.2(Q) pulled down still more TRIM25 than the vector control, demonstrating that RNA co-binding alone cannot explain the entirety of PARP13 and TRIM25's interaction (Fig. 4.5H). Therefore, the stability of the TRIM25:PARP13 complex depends on both direct protein binding and protein-RNA-protein interactions.

Discussion

Integrative genomics provides a robust framework for analyzing cellular shifts, such as during antiviral response. Although recent studies have used diverse genomics methods to study the effect of PARP13, none have captured PARP13 RNA binding during both basal state and antiviral response, and the emphasis of the prior studies have been on the effect of PARP13 binding to viral RNA (151,152,156). Despite PARP13's well-known role in RNA decay, there are few transcriptome-wide RNA-seq experiments comparing WT and PARP13 KD or KO cells (155,156)– most employ qPCR (140,151)– and none have deconvoluted transcriptomic signals to isolate the components of PARP13's activity attributable to its role in the antiviral

response versus cellular homeostasis. Here we leverage CLIP-seq and RNA-seq data to provide evidence that PARP13 may serve to create and maintain a cellular environment poised for an antiviral response through limiting PARP13 translation, regulating access to distinct mRNA pools, and elevating ISGylation machinery expression in the absence of infection.

We observe that PARP13 sequence-based binding specificity remains constant, regardless of antiviral response. However, the cellular localization of PARP13-bound transcripts shifts, likely due to a disproportionate upregulation of the cytosolic PARP13.2 isoform relative to the ER-derived membrane-associated PARP13.1 isoform. A context-specific shift in PARP13 localization would provide the opportunity for the cell to create a poised cytoplasmic PARP13-targeted RNA population, sequestered from regulation in healthy circumstances. Transcript localization may regulate which targets PARP13 has access to, so that certain transcripts are only bound and regulated by PARP13 in circumstances of antiviral response while other PARP13 targets are constitutively available for binding and regulation.

Autoregulation via binding to their own mRNA is common among RBPs, especially in response to stress. For example, some proteins bind their own pre-mRNA to affect splicing and drive transcript decay (171,172), some bind their own mRNA's 3'UTR to improve stability (28), and others alter the structure of their mRNA's 5'UTR to tune translation (173). Approximately one-third of PARP13 binding events are within 3'UTRs. The fact that PARP13.1 and PARP13.2 isoforms do not share 3'UTRs provides an opportunity for isoform-specific PARP13 regulation. Our observation that the 3' UTR of PARP13.1 is only regulated by PARP13 in the absence of an antiviral response further supports a shift in PARP13 targeting and regulation between cells in a basal state and cells staging an antiviral response.

This is the first study to assess the role of PARP13 on a host gene pathway. PARP13 not only directly binds multiple proteins of the ISGylation pathway but also affects the expression of their transcripts. Depletion of PARP13 globally down-regulated expression of

genes implicated in the innate antiviral response upon treatment with a viral mimic, in agreement with PARP13's studied role as a stimulator or antiviral response signaling (138). But PARP13 depletion also depressed expression of many factors of the ISGylation pathway in basal state cells, suggesting that PARP13 plays an additional role in keeping the ISGylation response primed for stress responses. Since our data only identified significant PARP13 binding sites within ISG15 and TRIM25 transcripts, we infer that depletion of PARP13 affects ISGylation machinery expression indirectly. PARP13 may directly affect expression of a single protein such as a transcription factor or decay factor that regulates the expression of multiple proteins in the ISGylation pathway. Further studies should deconvolute the role of PARP13 in the primary activation of the ISGylation pathway in the basal state from its secondary effect supporting the ISGylation-dependent antiviral response.

A limitation of applying spatial correlation methods to assess TRIM25 and PARP13 co-localization is that it is impossible to discern whether all PARP13 sites are bound simultaneously, only that they are regularly spaced relative to TRIM25 binding sites and each other. Our data cannot inform us whether TRIM25 binds different target transcripts post-viral infection, so it is possible that TRIM25 binding in 3p-RNA-treated cells may have a different relationship with PARP13 binding than what we are observing. The reason for PARP13's binding periodicity is unclear and ought to be more deeply explored via low-throughput methods to determine whether binding events are simultaneous and cooperative and to explore the role of periodic PARP13 binding on transcript translation and decay. Since this periodicity is apparent across CLIP experiments from multiple labs, it is unlikely a technical artifact of any single CLIP-seq protocol or peak-calling algorithm (152,156). Both PARP13 and TRIM25 dimerize and bind with many proteins to form large complexes (139,174), but no modelling so far has predicted a nucleoprotein complex sufficiently large to explain the spatial correlation signal observed that spans hundreds of nucleotides. The effect of multiple binding sites along a

single transcript has not been studied for many RBPs besides those that bind viral genomes for virion assembly (175,176), but kinetics studies of DAZL show its periodicity is cooperative and affects mRNA levels and ribosome association of bound proteins (177). Since PARP13 binding has also been implicated in mRNA decay and translational repression, multiple proximal binding sites may similarly modulate the effects of PARP13.

Methods

Cell culture

Human embryonic kidney (HEK) 293T cells were maintained in 1X Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS) at 37°C in a 5% CO_2 incubator. HEK293T PARP13 (GenBank accession number NM_024625.3) KO cells were gifted from (138). The PARP13 KO cells were treated for mycoplasma infection before experimental use.

ssRNA and 3p-RNA synthesis and transfection

A synthetic ssRNA (unphosphorylated RNA) oligo was used as a transfection control. A 3p-RNA (5'-triphosphorylated RNA viral mimic) oligo was produced by first annealing two synthetic oligos and transcribing the RNA fragment using T7 transcriptase with RNase inhibitor, followed by DNase I digestion. For transfection of HEK293T cells, 2.5 µL Lipofectamine 2000 and 125 µL Opti-MEM medium were combined with a mixture of 1 µg RNA (either ssRNA or 3p-RNA) and 125 µL Opti-MEM per well of a 6-well plate. After 5 minutes at room temperature, the solution was added drop-wise to cells grown to 70-90% confluence in 2 mL DMEM.

Dual luciferase assay

1000 HEK293T WT or PARP13 KO cells per well were seeded in a 96-well plate in 150 µL DMEM. 24 h later, 100 ng dual luciferase plasmid and 80 ng RNA treatment were combined with 120 µL Opti-MEM and 1 µL lipofectamine 2000 per well and sat for five min at room

temperature, and then were added to the cells (Fig. 4.2C). 24 h later, media was removed from the wells and cells were washed with 1x PBS. Samples to be used for parallel RT-qPCR were stored in 100 µL TRIzol and stored at -20˚C until RNA extraction. Dual luciferase reporter assay buffers were defrosted 2 h prior to experiment and diluted to 1x as relevant immediately before use. 20 µL per well of 1x lysis solution was added to cells and shaken at room temperature 15 min. To measure luminescence, 100 µL LAR II reagent was added to the lysed cells in each well and immediately measured (detects ctrl firefly). 100 µL 1x Stop and Glo reagent was then added to each well and immediately measured (detect *Renilla*). Preliminary tests of untransfected cells demonstrated < 1% luminescence compared to the lowest-luminescence transfected cells and so background is considered negligible in normalization calculations. To calculate luminescence, we normalized *Renilla* expression to firefly expression and compared it to normalized *Renilla* expression without a 3'UTR insert.

siRNA transfection

2.5 µL Lipofectamine 2000 and 125 µL Opti-MEM medium were combined with a mixture of 100 pM siPARP13 and 125 µL Opti-MEM per well of a 6-well plate. After 5 minutes at room temperature, the solution was added drop-wise to cells grown to ~70% confluence in 2 mL DMEM.

GFP construct transfection and immunoprecipitation

12 µL Lipofectamine 2000 and 750 µL Opti-MEM Medium were combined with a mixture of 18 µg EGFP plasmid and 750 µL Opti-MEM per 10-cm plate. After 5 minutes at room temperature, the solution was added drop-wise to cells grown to 70-90% confluence. 24 hours after transfection, cells were washed once with 1x PBS, scraped from the plate, and pelleted. Cells were lysed at 4˚C for 15 minutes in 500 µL mRIPA buffer (50 mM Tris-HCl pH 7.4, 150 mM NaCl, 0.1% sodium deoxycholate, 1% NP-40, 1 mM EDTA pH 8.0, 1x Halt protease

inhibitor cocktail, and 0.2 U/μL SUPERase-In RNase inhibitor). Samples were centrifuged at 20,000 x g for 10 minutes to separate cell debris and the supernatant was transferred to a new tube. 30 μL GFP-Trap bead slurry per sample was washed twice in 500 μL mRIPA buffer, resuspended in 500 μL mRIPA buffer, and added too cell supernatant. GFP-Trap beads were rotated with cell supernatant at 4˚C for 2 hours. Beads were spun down and washed twice. Beads were resuspended in 20 μL mRIPA buffer with either 6 μL NuPAGE LDS sample buffer for loading on a western blot or with 500 μL TRIzol for qRT-PCR.

Western blot

Samples were run on a denaturing polyacrylamide gel using NuPAGE MOPS SDS running buffer and transferred onto a nitrocellulose membrane at 20 V for 16 h using NuPAGE transfer buffer. Membrane was blocked with 6% milk in TBS for 45 min and then incubated with 1:1,000 primary antibody for 60 min in 6% milk in TBST. Membrane was washed three times with 6% milk in TBST and incubated with 1:10,000 secondary antibody for 90 min in 6% milk in TBST. Membrane was washed three times with TBS and imaged.

RT-PCR

Cells were harvested by suspension in TRIzol and RNA was isolated by phenol-chloroform extraction. Samples were reverse transcribed with SuperScript VILO cDNA synthesis kit and 10 μL RT-qPCR reactions were performed in triplicate with 1:200 dilution cDNA, 1 μM primer mix (IDT), and 1x PowerUp SYBR Green master mix on a 7500 Fast Real-Time PCR System (Applied Biosystems).

RNA-seq analysis

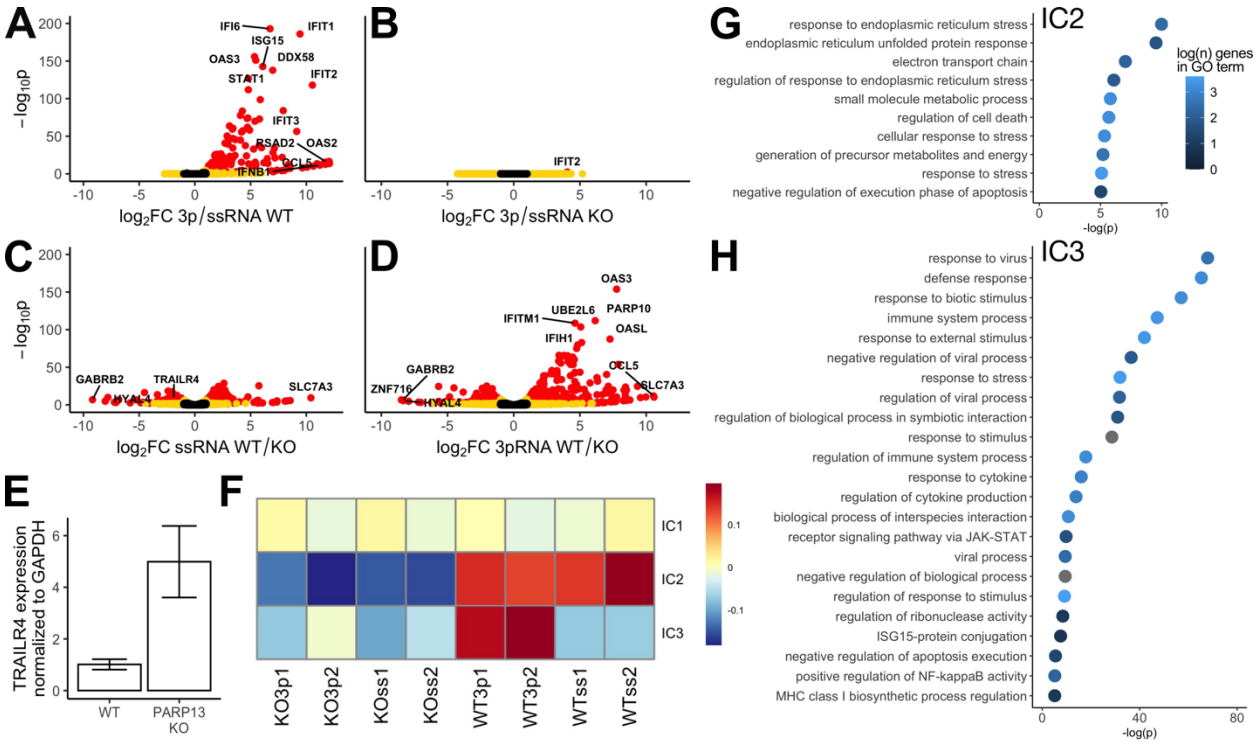
RNA-seq was performed on WT and PARP13 KO HEK293T cells 24 hours after transfection with either an ssRNA control or a 3p-RNA viral mimic in duplicate. RNA was purified via phenol-chloroform extraction and RNA quality was measured via RT-qPCR. All data

manipulation and visualization was performed in R. Adaptor-trimmed FASTQ RNA-seq reads were aligned using Bowtie2 (178). Rsubread (179) was used to identify genes corresponding to the peak loci, and DESeq2 (180) was used to determine which transcripts were differentially expressed (FDR < .05, fold-change > 2). Gene set enrichment analysis was performed by GOrilla and simplified by Revigo (97,98).

CLIP-seq analysis

Enhanced CLIP-seq (eCLIP) was performed on WT and PARP13 KO HEK293T cells 24 hours after transfection with either an ssRNA control or a 3p-RNA viral mimic in duplicate. Cells were washed with 1X PBS and UV cross-linked at 400 mJ/cm^2. Cells were scraped from the plate, washed in 1X PBS, pelleted at 500 x g for 5 minutes at 4˚C, and snap-frozen on liquid nitrogen. eCLIP was performed as previously published (165) using anti-PARP13 antibody. Reproducible peaks between replicates were determined by irreproducible discovery rate (IDR, https://www.encodeproject.org/software/idr/, https://github.com/YeoLab/merge_peaks). K-mer analysis was performed using tools from the Yeo Lab (https://github.com/YeoLab /clip_analysis_legacy/kmerdiff.py). Dinucleotide frequency of RNA-seq and CLIP-seq data was calculated by *compseq* from the EMBOSS suite (181). Hexamereric motif of CLIP-seq data was calculated using *findMotifs.pl* from HOMER (182).

Figure 4.1 PARP13 expression affects expression of many constitutive transcripts and is required for transcriptomic upregulation of innate immune response



A. Volcano plot of RNA-seq data 24 hours after treatment with 3p-RNA vs ssRNA in HEK293T WT cells (n=2). Black points have $\log_2$ fold-change < 1 and p > 0.01; yellow have $\log_2$ fold-change > 1 or p < 0.01; red points have $\log_2$ fold-change > 1 and p < 0.01.

B. Volcano plot of RNA-seq data 24 hours after treatment with +3p-RNA vs +ssRNA in HEK293T PARP13 KO cells (n=2).

C. Volcano plot of RNA-seq data in WT vs PARP13 KO cells 24 hours after treatment with +ssRNA (n=2).

D. Volcano plot of RNA-seq data in WT vs PARP13 KO cells 24 hours after treatment with +3p-RNA (n=2).

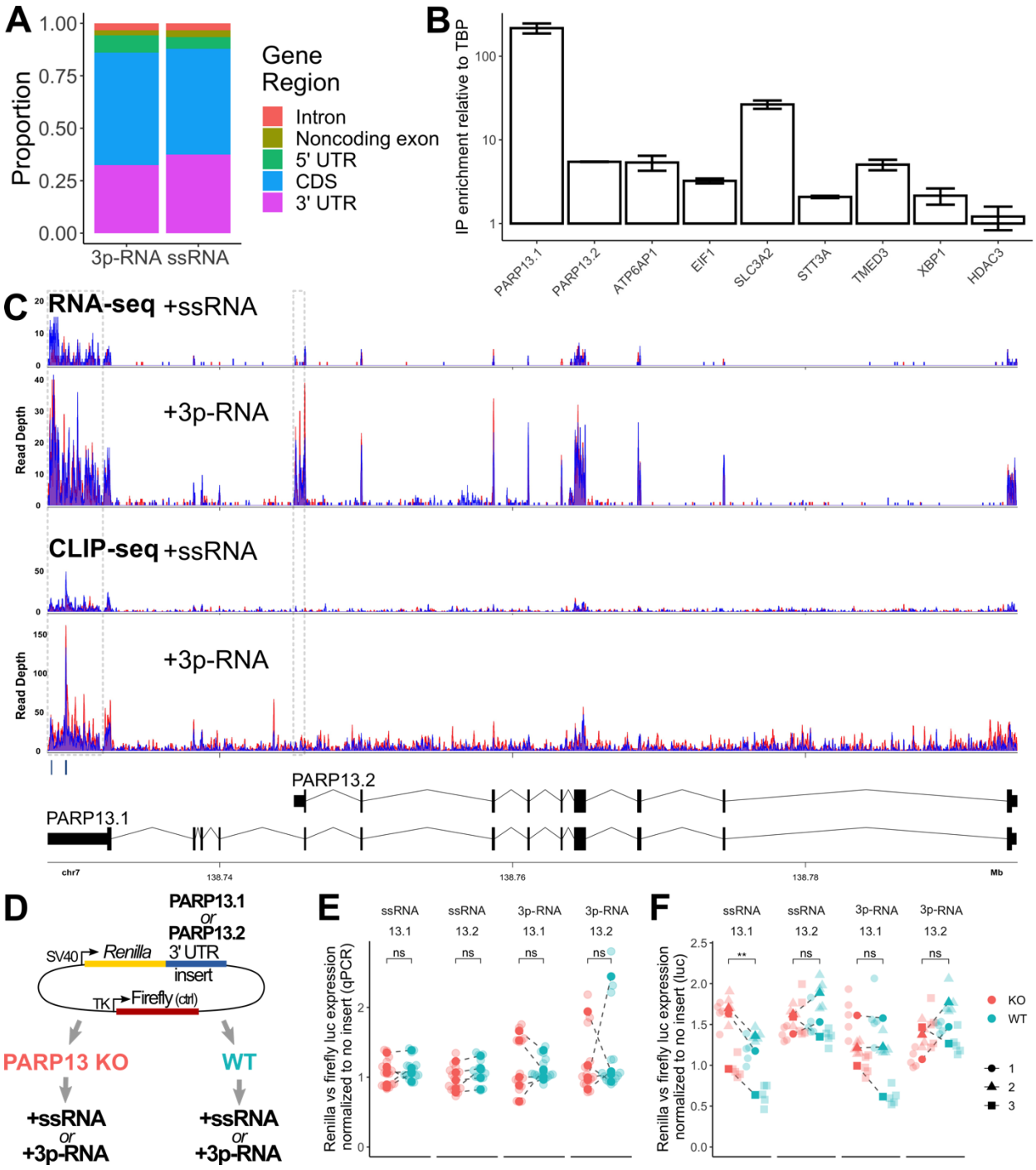
E. TRAILR4 expression measured by qPCR in WT HEK293T and PARP13 KO cells (n=3).

F. IC pattern weights across RNA-seq samples. The color scale represent values of the linear mixing matrix centered at 0 where columns of S contain the independent components, A is a linear mixing matrix, and data matrix X is considered to be a linear combination of non-Gaussian components such that X = SA.

G. Top GO terms that are enriched among genes that contribute to IC2, shown in (F).

H. Top GO terms that are enriched among genes that contribute to IC3, shown in (F).

Figure 4.2 PARP13 selectively binds PARP13.1 and regulates its translation via the 3' UTR



A. Proportions of PARP13 binding sites that fall within the indicated regions of mRNAs.

B. RNA immunoprecipitation using GFP-trap beads against exogenous GFP-PARP13.1 in HEK293T cells.

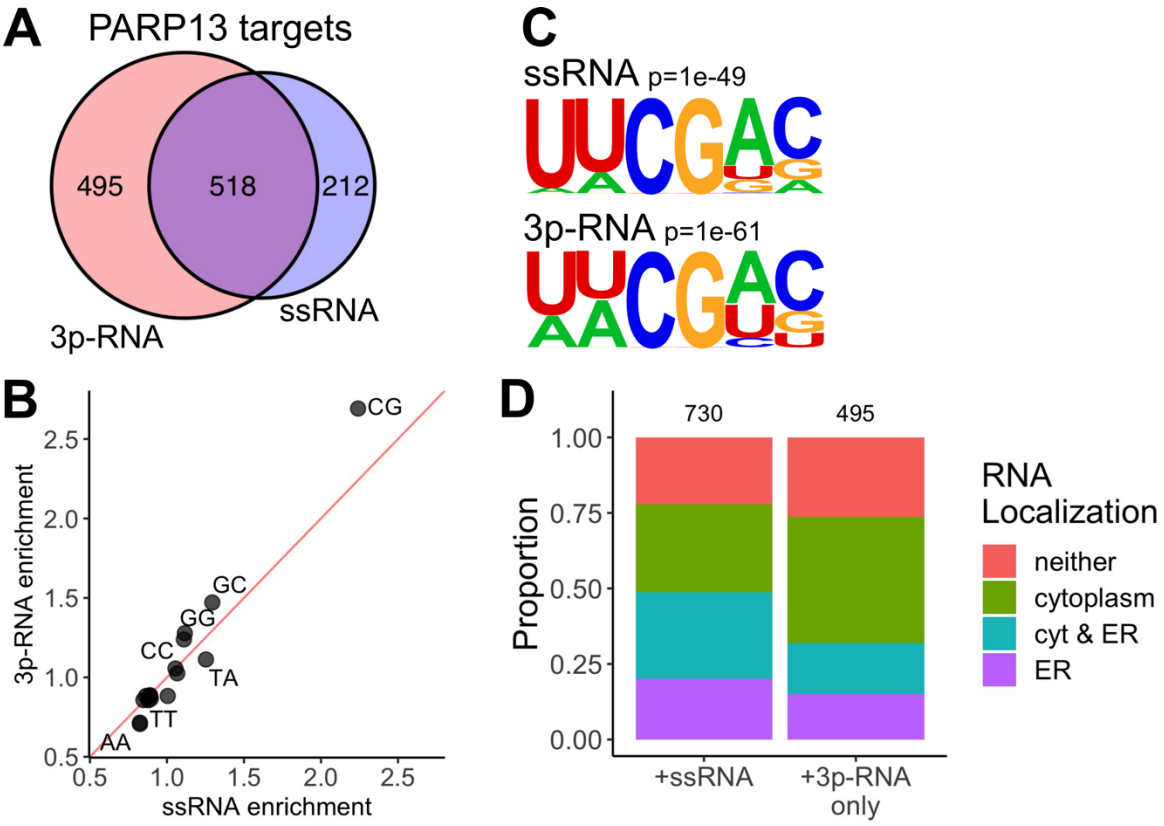
C. Aligned RNA-seq (top) and CLIP-seq (middle) reads within PARP13 gene region for +ssRNA and +3p-RNA treatments. Red and blue alignments correspond to duplicate experiments. Two PARP13 binding peaks located within the PARP13.1 3'UTR that were significant above background are indicated below the aligned reads. Bottom: a splicing map of the two PARP13 isoforms with grey boxes surround the PARP13.1-specific (left) and PARP13.2-specific (right) 3'UTRs.

D. Luciferase construct map and design of luciferase expression experiments.

E. Transcript levels of *Renilla* luciferase with either PARP13.1 or PARP13.2 3'UTR treated with ssRNA or 3p-RNA (n = 5). Within each experiment, *Renilla* expression was normalized to firefly expression and compared to normalized *Renilla* expression for a construct without a 3'UTR insert.

F. Luminescence of *Renilla* luciferase (n = 3) with the same normalization as (D).

Figure 4.3 Antiviral response shifts PARP13 target localization but not binding preferences



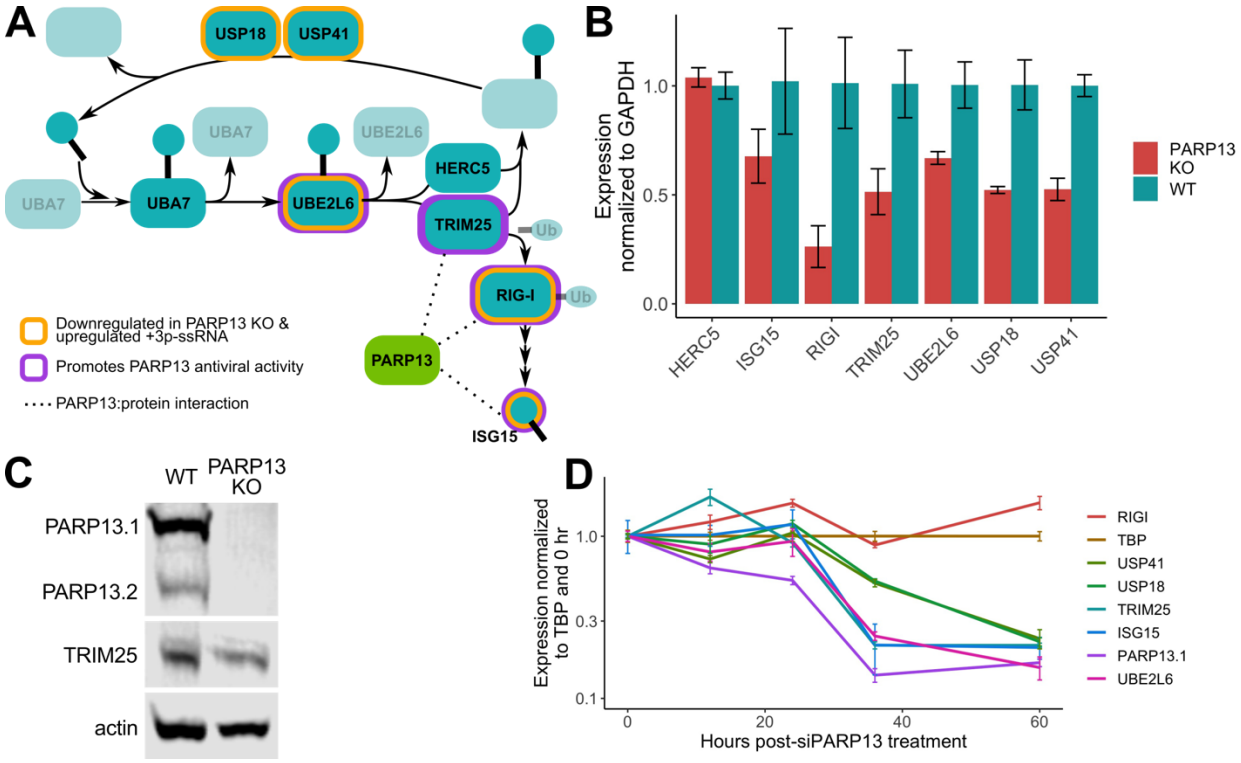
A. Venn diagram of transcripts that have at least one significantly bound region across both treatments.

B. Dinucleotide enrichment within PARP13-bound regions, normalized to dinucleotide expression of all transcribed RNAs.

C. A hexameric motif calculated by HOMER (182) that is enriched in samples with either +ssRNA (top) and +3p-RNA (bottom) treatments.

D. Proportions of previously-shown cellular localizations (166) for PARP13-bound transcripts across both treatments.

Figure 4.4 Genes associated with ISGylation are dysregulated in PARP13 KO cells



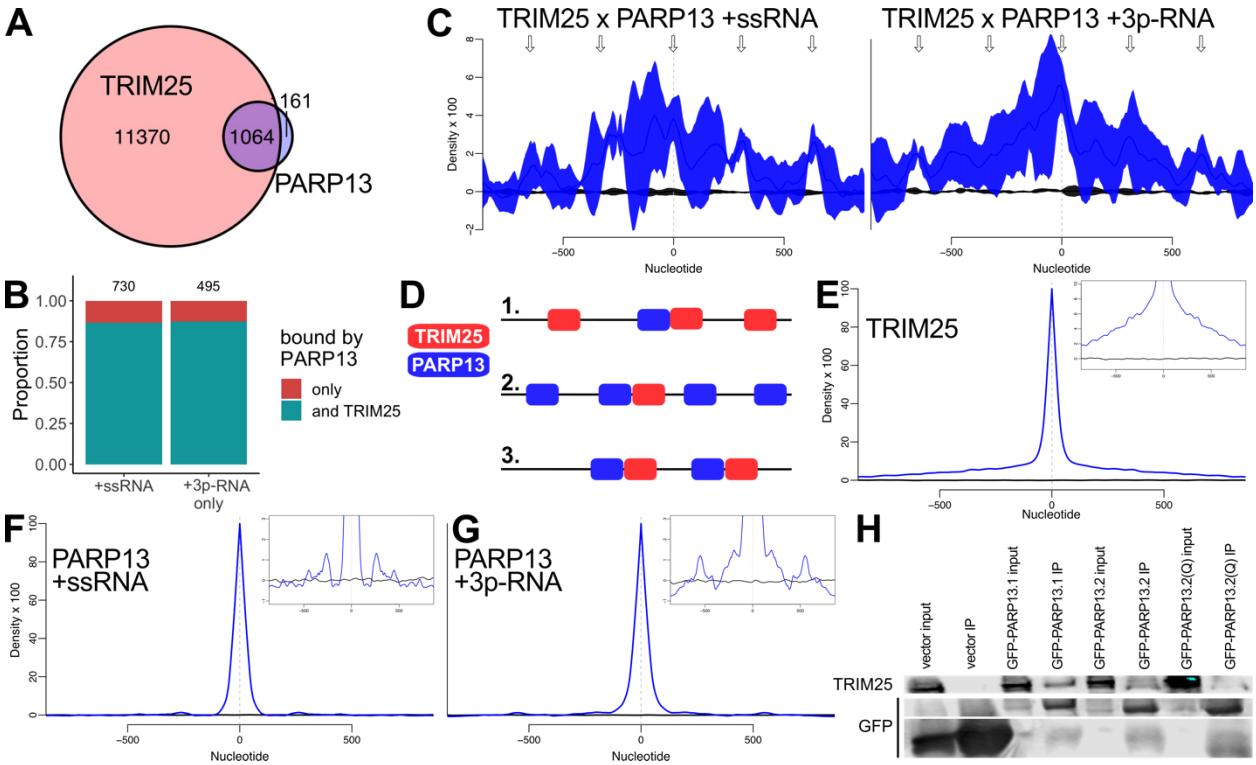
A. A schematic of the ISGylation pathway includes ISG15, a ubiquitin-like protein; UBA7, the E1; UBE2L6, the E2; HERC5 and TRIM25, E3 ligases; and USP18 and USP41, which remove ISG15 from conjugated proteins. RIG-I, which is ubiquitinated by TRIM25, activates expression of ISG15. Genes that have been shown to promote PARP13 antiviral activity are highlighted in purple. Genes identified by the RNA-seq data to be down-regulated in PARP13 KO cells in control condition and upregulated upon mock viral infection in WT cells are highlighted in gold. Proteins shown to bind PARP13 are connected by dotted lines.

B. qPCR of ISGylation genes in untreated WT and PARP13 KO HEK293T cells (n=3). UBA7 was too lowly expressed to quantify.

C. Western blot of PARP13 and TRIM25 expression in WT and PARP13 KO HEK293T cell lysate, with actin as a loading control.

D. PARP13 KD time course in WT HEK293T cells assessing the expression of ISGylation genes shown in (B) to have reduced expression in PARP13 KO cells.

Figure 4.5 TRIM25 and PARP13 interact via both RNA-dependent and RNA-independent contacts



A. Venn diagram of previously-shown TRIM25 targets from HeLa cells transiently expressing T7-tagged TRIM25 (163) and all PARP13 targets from HEK293T cells.

B. Proportions of PARP13-bound transcripts across both treatments that are also targets of TRIM25 (163).

C. Spatial correlation of PARP13- and TRIM25-bound regions across the HEK293T transcriptome calculated by nearBynding (82). Black line indicates mean shuffled background signal with standard error (n = 10,000) and blue line represents mean spatial correlation signal with standard error (n=3). Arrows are the same on left and right plot and represent periodic regions of significant spatial correlation between TRIM25 and PARP13. Left:

Correlation of TRIM25 binding peaks from HeLa cells and PARP13 binding peaks from +ssRNA-treated HEK293T cells. Right: Correlation of TRIM25 binding peaks and PARP13 binding peaks from +3p-RNA-treated cells.

D. Three possible models of TRIM25 and PARP13 binding proximity to explain the periodic signal observed in the spatial correlation in (C). 1. Multiple TRIM25 proteins bind near, upstream, and downstream a single PARP13 protein. 2. Multiple PARP13 proteins bind near, upstream, and downstream of a single TRIM25 protein. Or 3. a combination of 1. and 2. where multiple binding events of both proteins are proximal.

E. Auto-correlation of TRIM25 binding. Spatial correlation of a binding site to itself equals 1 at position 0. The same auto-correlation with a magnified y axis is shown in the top right.

F and G. Auto-correlation of PARP13 binding with ssRNA treatment (F) and 3p-RNA treatment (G). Spatial correlation of a binding site to itself equals 1 at position 0, but peaks upstream and downstream show additional PARP13 binding sites nearby. The same auto-correlation with a magnified y axis is shown in the top right.

H. Western blot for GFP and TRIM25 expression from co-immunoprecipitation using GFP-trap beads against exogenous GFP-PARP13 constructs in HEK293T cells.

Table 4.1 Key resources

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
TRIM25 polyclonal rabbit	Thermo Scientific	cat # 12573-1-AP
ISG15 monoclonal rabbit (7H29L24)	Thermo Scientific	cat # 703131
PARP13 (ZC3HAV1) polyclonal rabbit	ProteinTech	cat # 16820-1-AP
β-Actin monoclonal mouse (clone AC-15)	Sigma Aldrich	cat # A5441
GFP mouse IgG1κ (clones 7.1 and 13.1)	Roche	cat # 11814460001
OAS2 polyclonal rabbit	ProteinTech	cat # 19279-1-AP
Critical commercial assays and chemicals		
Dual-Luciferase Reporter Assay System	Promega	cat # E1910
PowerUp SYBR Green master mix	Thermo Fisher	cat # A25742
SuperScriptTM IV VILO Master Mix	Invitrogen	cat # 11754050

TRIzol reagent	Invitrogen	cat # 15596026
100x Halt protease inhibitor cocktail	Thermo Fisher	cat # 78429
SUPERase-In RNase inhibitor	Thermo Fisher	cat # AM2694
4x NuPAGE LDS sample buffer	Thermo Fisher	cat # NP0007
20x NuPAGE MOPS SDS running buffer	Thermo Fisher	cat # NP0001
20x NuPAGE transfer buffer	Thermo Fisher	cat # NP0006
AmpliScribeTM T7 High Yield Transcription Kit	Lucigen	cat # AS3107
Lipofectamine 2000 Transfection Reagent	Thermo Fisher	cat # 11668500
1X Dulbecco's Modified Eagle Medium	Gibco	cat # 11995073
fetal bovine serum	Sigma Aldrich	cat # F6765
Opti-MEM(R) I Reduced-Serum Medium (1X)	Gibco	cat # 31985062
Deposited data		
TRIM25 CLIP-seq	(163)	GEO: GSE104949
APEX-RIP	(166)	GEO: GSE106493
PARP13 CLIP and RNA-seq data		supplied by lab of Gene Yeo
Experimental models: Cell lines		
HEK 293T	ATCC	
PARP13 KO HEK 293T	(138)	clone 32, c.1023_1047del25insGTT, c.1027_1046del20, c.1031_1032insATCCA and c.1033_1034insCAATC
Oligonucleotides		
ssRNA	IDT	CACUUUCACUUCUCCCUUUUAGUUUCC
3p-RNA	IDT	TAATACGACTCACTATAGGAAACTAAAAGGGAGAAGTGAAAGTG and CACTTTCACTTCTCCCTTTTAGTTTCCTATAGTGAGTCGTATTA
siPARP13	Thermo Fisher	siRNA ID # HSS125638, cat # 1299001
Recombinant DNA		
EGFP-PARP13.1		
EGFP-PARP13.2		
EGFP-PARP13.2(Q)		
dual luciferase PARP13.1 3'UTR		
dual luciferase PARP13.2 3'UTR		
Software and algorithms		
nearBynding	This work (82)	https://bioconductor.org/packages/nearBynding/ ; https://github.com/vbusa1/nearBynding
HOMER	(182)	http://homer.ucsd.edu/homer/motif/
GOrilla	(97)	http://cbl-gorilla.cs.technion.ac.il/
Revigo	(98)	http://revigo.irb.hr/
EMBOSS: compseq	(181)	http://emboss.sourceforge.net/apps/cvs/emboss/apps/compseq.html

Chapter 5: Discussions

Our new algorithm, nearBynding, enables us to spatially resolve transcriptome-wide protein binding to inform RNA regulation. For example, we demonstrate STAU1's preference for binding stem-hairpin-stem structures with single-nucleotide resolution (Fig. 2.10D) and IGF2BP1's preference for binding m6A-modified RNA (Fig. 2.15). We illustrate the differences in RNA structure binding between WT UPF1 and binding mutants (Fig. 2.10E-G) and the spatial relationships of RBPs known to bind within introns (Fig. 2.12). We leveraged nearBynding to identify RBPs that co-localize with G3BP1 within 3'UTRs in order to generate a candidate list of potential SRD factors. We further applied nearBynding to query the spatial relationship of PARP13 and its cofactor TRIM25 and found that TRIM25 and PARP13 colocalize on common targets, though PARP13 also shows periodic binding surrounding the co-bound site.

Limitations of nearBynding

nearBynding is only able to provide aggregate information about RBP binding preferences, in contrast to other available machine-learning tools that predict individual binding events across the transcriptome (38) or RBP binding to alternative RNA structures (183). Future work will use machine-learning outputs of RBP binding preferences as an input track to nearBynding for correlation with variables that are not currently accommodated by those other programs for a more complete picture of protein binding. Another complexity of transcriptional data is that a single gene may have multiple transcribed variants that are overlapping on a traditional genomic scale; since nearBynding does not allow redundant mapping of data (e.g., one RBP binding event in an overlapping region of two variants cannot be duplicated in the generation of the transcriptome), it is unable to accommodate two or more variants from the same gene if the queried regions of the transcripts overlap.

The current software for nearBynding does not accommodate data from in vitro experiments of proteins bound to RNA oligos such as from SELEX or RNA Bind-N-Seq, since these methods use RNA sequences that do not correspond to transcripts that are mappable to the genome (184,185). To correlate in vitro data via nearBynding, the user would need to create a novel annotated genome containing sequences for every oligonucleotide probe in the queried experiment; though the current pipeline could process a novel annotated genome without further customization, there are no functions within nearBynding to assist in that task. While nearBynding accepts two transcriptome-wide feature inputs for correlation, the software currently only supports the consideration of replicates and background signal for one of these inputs, such as for an input control in a CLIP experiment. Future work will support the possibility of accommodating replicates and removing background signal for both input features, such as for analyzing RBP binding to RNA data derived from an RNA immunoprecipitation (RIP) experiment that use antibodies targeting RNA structures or modifications (74,131).

The resolution of the nearBynding output is limited by the lower resolution input data track, so an output may imprecisely depict the correlation between one narrow data type (e.g. nucleotide sequence, which has functional peak widths of one nt) and one broader data type (e.g. CLIP peaks). However, in cases of similar-width data types (e.g. single-nucleotide RBP binding information such as iCLIP or hiCLIP (55,186) [see Fig. 2.10D]), it is possible to, for example, assess RNA sequence preferences relative to RBP binding with single-nucleotide resolution. Current work has not probed RNA sequence correlations, but future work will integrate RNA sequence as a feature by separating each of the four nucleotides into individual RNA tracks, similar to how CapR separates six RNA structures into different tracks.

An additional modification to nearBynding that would extend its functionality would be to save the bedGraph file of local kernelled correlations that are calculated during the course of nearBynding processing and allow for correlation of a third dataset to those local kernelled

correlations. This would, for example, enable the user to correlate RNA structure and the correlation between UPF1 and G3BP1 binding within 3' UTRs (Fig. 3.7D) to observe what RNA structures are enriched at sites with both UPF1 and G3BP1 binding. Further, a function to access the saved local kernelled correlation information could help identify transcriptomic regions with a strong correlation in a non-hypothesis-driven manner or, inversely, provide correlation information about specific transcripts of interest.

Limitations of data comparisons

Although spatial transcriptomic correlation can be a very sensitive method (see Fig. 2.15), it is important that all data is properly controlled and contextualized to prevent the introduction of confounders or the over-interpretation of noise (see Fig. 2.2). Due to the difficulty and expense of producing genomic data, researchers often have to rely on publicly-available datasets, which may or may not contain data from the same cell lines or treatment conditions (e.g. Fig. 4.5C: PARP13 CLIP is from RNA-transfected HEK293T cells and TRIM25 CLIP is from untreated HeLa cells). Ideally, biologically meaningful signal will be reproducible across cell lines and robust to minor differences in growth conditions, but closely matching data context if possible or reproducing correlations across diverse contexts when not possible is important to minimize the effect of unknown variables and increase confidence in results (see Fig. 2.8E: TIA1 binding preferences are consistent across cell lines and therefore high-confidence, whereas LARP7 binding preferences are not reproduceable).

Wild-type conditions are often considered a gold standard for data collection, but there are constraints that may make WT transcriptomic data collection and interpretation impossible or of limited informativeness. For example, a circRNA of interest may be too lowly expressed endogenously relative to its linear counterpart, so it may be preferable to instead introduce an exogenous construct for genomics analysis (Appendix A). A protein of interest may not have an antibody available to perform CLIP-seq; or, even if there were a sufficiently-reliable antibody for

a given protein, the protein may have two isoforms that an antibody cannot distinguish between (e.g. PARP13, Fig. 4.2B), so it may be necessary to knock-in a protein tag or transfect in an exogenous, tagged construct (e.g. Fig. 4.5E). In all circumstances, it is important to ensure the models and systems relevant to the experimental question have appropriate controls and the transcriptomic correlation results are sufficiently contextualized.

New areas of interrogation

Single-cell transcriptomic data types such as Surveying Targets by APOBEC Mediated Profiling (STAMP), Ribo-STAMP, and Ribo-seq are only now emerging, but they show promise in providing single-cell-resolution information about RBP binding, RNA modification, and ribosomal occupancy (187,188). Future work could expand the functionality of nearBynding to analyze single-cell transcriptomic data, allowing for the interrogation of shifting transcriptomic relationships among different cell types.

nearBynding is unique in that it allows for efficient correlation of diverse data types for specific sections of mRNA transcripts or subcategories of the transcriptome. This contextualization of transcriptomic correlation can be powerful when assessing signals that may be smothered by looking across the entire transcriptome. For example, G3BP1 strongly prefers to bind the CDS of transcripts (Fig. 3.7A) and shows very little colocalization with UPF1 across the exome (Fig. 3.7C). However, when correlation is limited to only within 3'UTRs, we are able to observe the colocalization of G3BP1 and UPF1, likely reflecting SRD events (Fig. 3.7D). There are many other transcriptome localization relationships that are specific to certain regions and types of RNA for which delimited correlation may be informative (21,50).

Comparing correlations across different RBP mutants or isoforms can provide insight into how those differences affect protein localization and binding preferences (e.g. Fig. 2.10E-G). For example, it is known that mutation of the S149A phosphorylation site on G3BP1 ablates SRD, but it is unclear whether this mutation alters G3BP1's colocalization with UPF1 at SRD

targets or whether it affects interactions with other SRD factors (32). A S149A G3BP1 CLIP dataset could deconvolute these two possibilities. An alternative UPF1 isoform was recently identified that modifies NMD specificity in response to cellular stress, and a CLIP dataset of the isoform could show whether it has different G3BP1 spatial correlation and SRD activity (1).

Similarly, it may benefit our understanding of the innate immune response if we were able to deconvolute RNA binding of the two primary PARP13 isoforms. Currently-available PARP13 CLIP datasets do not distinguish between PARP13.1 and PARP13.2 binding. Although we infer that shifts in target transcript cellular localization are due to differential expression of PARP13 isoforms (Fig.4.3E), isoform-specific CLIP data would allow us to evaluate these claims more directly. Analogously, the currently-available TRIM25 CLIP dataset does not test TRIM25 binding in the context of an antiviral response (163). TRIM25 and PARP13 may bind RNA and interact differently between healthy and viral-infected cells, so it would be beneficial to test TRIM25 binding in HEK293T cells treated with ssRNA and 3p-RNA.

Protein binding correlation via nearBynding, partnered with literature evidence and computational methods, can provide intriguing hypotheses that must be tested in a manner appropriate to the specific phenomenon observed. For example, nearBynding demonstrated that many proteins colocalize with G3BP1 within 3'UTRs (Fig. 3.7E). In order to assess whether these proteins are involved in SRD with G3BP1, we ought to test their effect on known SRD target expression levels in KO or KD cells. However, different methods would be appropriate to test the biological relevance of the TRIM25 and PARP13 transcriptomic colocalization (Fig 4.5C). For example, we may perform *in vitro* gel-shift assays to assess the cooperativity of TRIM25 and PARP13 binding and whether multiple PARP13 proteins bind a single transcript simultaneously. In this way, the dialogue between computational and molecular experimentation can contribute to a more complete understanding of post-transcriptional RNA regulation and function.

References

1. Mehta SL, Dempsey RJ, Vemuganti R. Role of Circular RNAs in Brain Development and CNS Diseases. Prog Neurobiol. 2020 Mar;186:101746.

2. Pozzo E, Chai YC, Sampaolesi M. Comprehensive Overview of Non-coding RNAs in Cardiac Development. In: Xiao J, editor. Non-coding RNAs in Cardiovascular Diseases [Internet]. Singapore: Springer Singapore; 2020. p. 197–211. Available from: https://doi.org/10.1007/978-981-15-1671-9_11

3. Goodall GJ, Wickramasinghe VO. RNA in cancer. Nat Rev Cancer. 2021 Jan;21(1):22–36.

4. Barbieri I, Kouzarides T. Role of RNA modifications in cancer. Nat Rev Cancer. 2020 Jun;20(6):303–22.

5. Nation GK, Saffold CE, Pua HH. Secret messengers: Extracellular RNA communication in the immune system. Immunological Reviews. 2021;304(1):62–76.

6. Chen YG, Satpathy AT, Chang HY. Gene regulation in the immune system by long noncoding RNAs. Nat Immunol. 2017 Sep;18(9):962–72.

7. Shulman Z, Stern-Ginossar N. The RNA modification N6-methyladenosine as a novel regulator of the immune system. Nat Immunol. 2020 May;21(5):501–12.

8. Tan J, Yang L, Ong AAL, Shi J, Zhong Z, Lye ML, et al. A Disease-Causing Intronic Point Mutation C19G Alters Tau Exon 10 Splicing via RNA Secondary Structure Rearrangement. Biochemistry. 2019 Mar 26;58(12):1565–78.

9. Moqtaderi Z, Geisberg JV, Struhl K. Extensive Structural Differences of Closely Related 3′ mRNA Isoforms: Links to Pab1 Binding and mRNA Stability. Molecular Cell [Internet]. 2018 Oct 11 [cited 2018 Oct 30]; Available from: http://www.sciencedirect.com/science/article/pii/S1097276518307366

10. Edupuganti RR, Geiger S, Lindeboom RGH, Shi H, Hsu PJ, Lu Z, et al. N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis. Nat Struct Mol Biol. 2017 Oct;24(10):870–8.

11. Sugimoto Y, König J, Hussain S, Zupan B, Curk T, Frye M, et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. Genome Biology. 2012 Aug 3;13(8):R67.

12. Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009 Apr 10;324(5924):218–23.

13. Okada S, Ueda H, Noda Y, Suzuki T. Transcriptome-wide identification of A-to-I RNA editing sites using ICE-seq. Methods. 2019 Mar 1;156:66–78.

14. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, Schroth GP, et al. Multiplexed RNA structure characterization with selective 2′-hydroxyl acylation analyzed by primer

extension sequencing (SHAPE-Seq). Proc Natl Acad Sci U S A. 2011 Jul 5;108(27):11063–8.

15. Stavrovskaya ED, Niranjan T, Fertig EJ, Wheelan SJ, Favorov AV, Mironov AA. StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data. Bioinformatics. 2017 Oct 15;33(20):3158–65.

16. Favorov A, Mularoni L, Cope LM, Medvedeva Y, Mironov AA, Makeev VJ, et al. Exploring Massive, Genome Scale Datasets with the GenometriCorr Package. PLoS Comput Biol [Internet]. 2012 May 31 [cited 2020 Dec 8];8(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3364938/

17. Zhang Y, Liu H, Lv J, Xiao X, Zhu J, Liu X, et al. QDMR: a quantitative method for identification of differentially methylated regions by entropy. Nucleic Acids Res. 2011 May;39(9):e58.

18. Kravatsky YV, Chechetkin VR, Tchurikov NA, Kravatskaya GI. Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression. DNA Res. 2015 Feb;22(1):109–19.

19. Carlile TM, Martinez NM, Schaening C, Su A, Bell TA, Zinshteyn B, et al. mRNA structure determines modification by pseudouridine synthase 1. Nat Chem Biol. 2019 Sep 2;1–9.

20. Beltran M, Tavares M, Justin N, Khandelwal G, Ambrose J, Foster BM, et al. G-tract RNA removes Polycomb repressive complex 2 from genes. Nature Structural & Molecular Biology. 2019 Oct;26(10):899–909.

21. Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, et al. A large-scale binding and functional map of human RNA-binding proteins. Nature. 2020 Jul;583(7818):711–9.

22. Waldron JA, Tack DC, Ritchey LE, Gillen SL, Wilczynska A, Turro E, et al. mRNA structural elements immediately upstream of the start codon dictate dependence upon eIF4A helicase activity. Genome Biology. 2019 Dec 30;20(1):300.

23. Briese M, Haberman N, Sibley CR, Faraway R, Elser AS, Chakrabarti AM, et al. A systems view of spliceosomal assembly and branchpoints with iCLIP. Nature Structural & Molecular Biology. 2019 Oct;26(10):930–40.

24. Heller D, Krestel R, Ohler U, Vingron M, Marsico A. ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. Nucleic Acids Res. 2017 Nov 2;45(19):11004–18.

25. Blaszczyk J, Gan J, Tropea JE, Court DL, Waugh DS, Ji X. Noncatalytic Assembly of Ribonuclease III with Double-Stranded RNA. Structure. 2004 Mar 1;12(3):457–66.

26. Li X, Quon G, Lipshitz HD, Morris Q. Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. RNA. 2010 Jun;16(6):1096–107.

27. Bahrami-Samani E, Penalva LOF, Smith AD, Uren PJ. Leveraging cross-link modification events in CLIP-seq for motif discovery. Nucleic Acids Res. 2015 Jan 9;43(1):95–103.

28. Wilbert ML, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, et al. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. Mol Cell. 2012 Oct 26;48(2):195–206.

29. Taliaferro JM, Lambert NJ, Sudmant PH, Dominguez D, Merkin JJ, Alexis MS, et al. RNA sequence context effects measured in vitro predict in vivo protein binding and regulation. Mol Cell. 2016 Oct 20;64(2):294–306.

30. Li X, Kazan H, Lipshitz HD, Morris QD. Finding the target sites of RNA-binding proteins. Wiley Interdiscip Rev RNA. 2014 Jan;5(1):111–30.

31. Ge Y, Jin J, Li J, Ye M, Jin X. The roles of G3BP1 in human diseases (review). Gene. 2022 May 5;821:146294.

32. Fischer JW, Busa VF, Shao Y, Leung AKL. Structure-Mediated RNA Decay by UPF1 and G3BP1. Molecular Cell [Internet]. 2020 Feb 3 [cited 2020 Feb 3]; Available from: http://www.sciencedirect.com/science/article/pii/S1097276520300423

33. Maticzka D, Lange SJ, Costa F, Backofen R. GraphProt: modeling binding preferences of RNA-binding proteins. Genome Biol. 2014;15(1):R17.

34. Pan X, Shen HB. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. Bioinformatics. 2018 Oct 15;34(20):3427–36.

35. Yan Z, Hamilton WL, Blanchette M. Graph neural representational learning of RNA secondary structures for predicting RNA-protein interactions. Bioinformatics. 2020 Jul;36(Suppl 1):i276–84.

36. Chen X, Castro SA, Liu Q, Hu W, Zhang S. Practical considerations on performing and analyzing CLIP-seq experiments to identify transcriptomic-wide RNA-protein interactions. Methods. 2019 Feb 15;155:49–57.

37. Sasse A, Laverty KU, Hughes TR, Morris QD. Motif models for RNA-binding proteins. Current Opinion in Structural Biology. 2018 Dec 1;53:115–23.

38. Sun L, Xu K, Huang W, Yang YT, Li P, Tang L, et al. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. Cell Res. 2021 Feb 23;1–22.

39. Jarmoskaite I, Denny SK, Vaidyanathan PP, Becker WR, Andreasson JOL, Layton CJ, et al. A Quantitative and Predictive Model for RNA Binding by Human Pumilio Proteins. Molecular Cell. 2019 Jun 6;74(5):966-981.e18.

40. Fukunaga T, Ozaki H, Terai G, Asai K, Iwasaki W, Kiryu H. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. 2014;15(1):R16.

41. Schuhmacher D, diagrams) BB (aha and power, Bonneel (networkflow) N, shortlist) CG (simplex and, Hartmann (semidiscrete1) V, integration) FH (transport_track and networkflow, et al. transport: Computation of Optimal Transport Plans and Wasserstein

Distances [Internet]. 2020 [cited 2020 Apr 21]. Available from: https://CRAN.R-project.org/package=transport

42. Drewe-Boss P, Wessels HH, Ohler U. omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data. Genome Biology. 2018 Nov 1;19(1):183.

43. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat Struct Mol Biol. 2013 Dec;20(12):1434–42.

44. Uren PJ, Bahrami-Samani E, Burns SC, Qiao M, Karginov FV, Hodges E, et al. Site identification in high-throughput RNA–protein interaction data. Bioinformatics. 2012 Dec 1;28(23):3013–20.

45. Zhang Z, Xing Y. CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome. Nucleic Acids Research. 2017 Sep 19;45(16):9260–71.

46. Matunis MJ, Michael WM, Dreyfuss G. Characterization and primary structure of the poly(C)-binding heterogeneous nuclear ribonucleoprotein complex K protein. Mol Cell Biol. 1992 Jan;12(1):164–71.

47. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research. 2016 Jul 8;44(W1):W160–5.

48. Fukunaga T, Ozaki H, Terai G, Asai K, Iwasaki W, Kiryu H. CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. Genome Biol. 2014;15(1):R16.

49. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. 2018 04;46(D1):D794–801.

50. Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, et al. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. Genome Biology. 2020 Apr 6;21(1):90.

51. Beaudoin JD, Jodoin R, Perreault JP. New scoring system to identify RNA G-quadruplex folding. Nucleic Acids Res. 2014 Jan;42(2):1209–23.

52. Wang X, McLachlan J, Zamore PD, Hall TMT. Modular Recognition of RNA by a Human Pumilio-Homology Domain. Cell. 2002 Aug 23;110(4):501–12.

53. Nam Y, Chen C, Gregory RI, Chou JJ, Sliz P. Molecular Basis for Interaction of let-7 MicroRNAs with Lin28. Cell. 2011 Nov 23;147(5):1080–91.

54. Ramos A, Grünert S, Adams J, Micklem DR, Proctor MR, Freund S, et al. RNA recognition by a Staufen double-stranded RNA-binding domain. EMBO J. 2000 Mar 1;19(5):997–1009.

55. Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, et al. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature. 2015 Mar 26;519(7544):491–4.

56. Kim YK, Maquat LE. UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. RNA. 2019 Apr;25(4):407–22.

57. Lee SR, Pratt G, Martinez F, Yeo GW, Lykke-Andersen J. Target discrimination in nonsense-mediated mRNA decay requires Upf1 ATPase activity. Mol Cell. 2015 Aug 6;59(3):413–25.

58. JiaYan W, JingFa X, LingPing W, Jun Z, HongYan Y, ShuangXiu W, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. Sci China Life Sci. 2013 Oct;56(10):968–74.

59. Brázda V, Hároníková L, Liao JCC, Fojta M. DNA and RNA Quadruplex-Binding Proteins. Int J Mol Sci. 2014 Sep 29;15(10):17493–517.

60. Puig Lombardi E, Londoño-Vallejo A. A guide to computational methods for G-quadruplex prediction. Nucleic Acids Res. 2020 Jan 10;48(1):1–15.

61. Kwok CK, Marsico G, Sahakyan AB, Chambers VS, Balasubramanian S. rG4-seq reveals widespread formation of G-quadruplex structures in the human transcriptome. Nature Methods. 2016 Oct;13(10):841–4.

62. Simko EAJ, Liu H, Zhang T, Velasquez A, Teli S, Haeusler AR, et al. G-quadruplexes offer a conserved structural motif for NONO recruitment to NEAT1 architectural lncRNA. Nucleic Acids Res. 2020 Jun 4;

63. Yagi R, Miyazaki T, Oyoshi T. G-quadruplex binding ability of TLS/FUS depends on the β-spiral structure of the RGG domain. Nucleic Acids Res. 2018 Jul 6;46(12):5894–901.

64. Pietras Z, Wojcik MA, Borowski LS, Szewczyk M, Kulinski TM, Cysewski D, et al. Dedicated surveillance mechanism controls G-quadruplex forming non-coding RNAs in human mitochondria. Nat Commun [Internet]. 2018 Jul 2 [cited 2020 Jul 22];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6028389/

65. Rouleau SG, Garant JM, Bolduc F, Bisaillon M, Perreault JP. G-Quadruplexes influence pri-microRNA processing. RNA Biol. 2017 Dec 11;15(2):198–206.

66. Zyner KG, Mulhearn DS, Adhikari S, Martínez Cuesta S, Di Antonio M, Erard N, et al. Genetic interactions of G-quadruplexes in humans. eLife [Internet]. [cited 2020 Jul 22];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6615864/

67. Lee DSM, Ghanem LR, Barash Y. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. Nature Communications. 2020 Jan 27;11(1):1–12.

68. Kamura T, Katsuda Y, Kitamura Y, Ihara T. G-quadruplexes in mRNA: A key structure for biological function. Biochemical and Biophysical Research Communications. 2020 May 21;526(1):261–6.

69. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019 Jan 8;47(D1):D427–32.

70. Brázda V, Červeň J, Bartas M, Mikysková N, Coufal J, Pečinka P. The Amino Acid Composition of Quadruplex Binding Proteins Reveals a Shared Motif and Predicts New Potential Quadruplex Interactors. Molecules [Internet]. 2018 Sep 13 [cited 2020 May 5];23(9). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6225207/

71. Roost C, Lynch SR, Batista PJ, Qu K, Chang HY, Kool ET. Structure and Thermodynamics of N6-Methyladenosine in RNA: A Spring-Loaded Base Modification. J Am Chem Soc. 2015 Feb 11;137(5):2107–15.

72. Yu J, Chen M, Huang H, Zhu J, Song H, Zhu J, et al. Dynamic m6A modification regulates local translation of mRNA in axons. Nucleic Acids Res. 2018 Feb 16;46(3):1412–23.

73. Patil DP, Chen CK, Pickering BF, Chow A, Jackson C, Guttman M, et al. m6A RNA methylation promotes XIST-mediated transcriptional repression. Nature. 2016 Sep 15;537(7620):369–73.

74. Huang H, Weng H, Sun W, Qin X, Shi H, Wu H, et al. Recognition of RNA N6-methyladenosine by IGF2BP Proteins Enhances mRNA Stability and Translation. Nat Cell Biol. 2018 Mar;20(3):285–95.

75. Huang H, Weng H, Zhou K, Wu T, Zhao BS, Sun M, et al. Histone H3 trimethylation at lysine 36 guides m6A RNA modification co-transcriptionally. Nature. 2019 Mar;567(7748):414–9.

76. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Curr Protoc Bioinformatics. 2014 Sep 8;47:11.12.1-11.12.34.

77. Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, et al. Ensembl Genomes 2020—enabling non-vertebrate genomic research. Nucleic Acids Res. 2020 Jan 8;48(D1):D689–95.

78. Sun Z, Xue S, Xu H, Hu X, Chen S, Yang Z, et al. Effects of NSUN2 deficiency on the mRNA 5-methylcytosine modification and gene expression profile in HEK293 cells. Epigenomics. 2018 Dec 11;11(4):439–53.

79. Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, et al. UCSC Genome Browser enters 20th year. Nucleic Acids Res. 2020 Jan 8;48(D1):D756–61.

80. Ozdilek BA, Thompson VF, Ahmed NS, White CI, Batey RT, Schwartz JC. Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. Nucleic Acids Res. 2017 Jul 27;45(13):7984–96.

81. Navarro D. Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.5) [Internet]. Adelaide, Australia: University of Adelaide; 2015. Available from: http://ua.edu.au/ccs/teaching/lsr

82. Busa VF, Favorov AV, Fertig EJ, Leung AKL. Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding. Cell Reports Methods [Internet]. 2021 Oct 25 [cited 2021 Dec 7];1(6). Available from: https://www.cell.com/cell-reports-methods/abstract/S2667-2375(21)00143-0

83. Beaudoin JD, Novoa EM, Vejnar CE, Yartseva V, Takacs CM, Kellis M, et al. Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. Nat Struct Mol Biol. 2018 Aug;25(8):677–86.

84. Felekkis K, Touvana E, Stefanou C, Deltas C. microRNAs: a newly described class of encoded molecules that play a role in health and disease. Hippokratia. 2010;14(4):236–40.

85. Binder R, Horowitz JA, Basilion JP, Koeller DM, Klausner RD, Harford JB. Evidence that the pathway of transferrin receptor mRNA degradation involves an endonucleolytic cleavage within the 3' UTR and does not involve poly(A) tail shortening. EMBO J. 1994 Apr 15;13(8):1969–80.

86. Kim YK, Furic L, DesGroseillers L, Maquat LE. Mammalian Staufen1 Recruits Upf1 to Specific mRNA 3′UTRs so as to Elicit mRNA Decay. Cell. 2005 Jan 28;120(2):195–208.

87. Mino T, Murakawa Y, Fukao A, Vandenbon A, Wessels HH, Ori D, et al. Regnase-1 and Roquin Regulate a Common Element in Inflammatory mRNAs by Spatiotemporally Distinct Mechanisms. Cell. 2015 May 21;161(5):1058–73.

88. Fiorini F, Bagchi D, Le Hir H, Croquette V. Human Upf1 is a highly processive RNA helicase and translocase with RNP remodelling activities. Nat Commun [Internet]. 2015 Jul 3 [cited 2019 Nov 12];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4506499/

89. Tourrière H, Gallouzi I eddine, Chebli K, Capony JP, Mouaikel J, van der Geer P, et al. RasGAP-Associated Endoribonuclease G3BP: Selective RNA Degradation and Phosphorylation-Dependent Localization. Mol Cell Biol. 2001 Nov;21(22):7747–60.

90. Dou N, Chen J, Yu S, Gao Y, Li Y. G3BP1 contributes to tumor metastasis via upregulation of Slug expression in hepatocellular carcinoma. Am J Cancer Res. 2016 Nov 1;6(11):2641–50.

91. Winslow S, Leandersson K, Larsson C. Regulation of PMP22 mRNA by G3BP1 affects cell proliferation in breast cancer cells. Molecular Cancer. 2013 Dec 9;12(1):156.

92. Zhang CH, Wang JX, Cai ML, Shao R, Liu H, Zhao WL. The roles and mechanisms of G3BP1 in tumour promotion. Journal of Drug Targeting. 2019 Mar 16;27(3):300–5.

93. Zheng H, Zhan Y, Zhang Y, Liu S, Lu J, Yang Y, et al. Elevated expression of G3BP1 associates with YB1 and p-AKT and predicts poor prognosis in nonsmall cell lung cancer patients after surgical resection. Cancer Medicine [Internet]. [cited 2019 Sep 30];0(0). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/cam4.2579

94. Wang Y, Fu D, Chen Y, Su J, Wang Y, Li X, et al. G3BP1 promotes tumor progression and metastasis through IL-6/G3BP1/STAT3 signaling axis in renal cell carcinomas. Cell Death & Disease. 2018 May 2;9(5):501.

95. Györffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q, et al. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res Treat. 2010 Oct 1;123(3):725–31.

96. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. CA: A Cancer Journal for Clinicians. 2021;71(1):7–33.

97. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009 Feb 3;10:48.

98. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS One. 2011 Jul 18;6(7):e21800.

99. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020 Jan 8;48(D1):D127–31.

100. Lánczky A, Nagy Á, Bottai G, Munkácsy G, Szabó A, Santarpia L, et al. miRpower: a web-tool to validate survival-associated miRNAs utilizing expression data from 2178 breast cancer patients. Breast Cancer Res Treat. 2016 Dec 1;160(3):439–46.

101. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D535–9.

102. Lygerou Z, Mitchell P, Petfalski E, Séraphin B, Tollervey D. The POP1 gene encodes a protein component common to the RNase MRP and RNase P ribonucleoproteins. Genes Dev. 1994 Jun 15;8(12):1423–33.

103. Gallouzi I eddine, Parker F, Chebli K, Maurier F, Labourier E, Barlat I, et al. A Novel Phosphorylation-Dependent RNase Activity of GAP-SH3 Binding Protein: a Potential Link between Signal Transduction and RNA Stability. Mol Cell Biol. 1998 Jul;18(7):3956–65.

104. Oi N, Yuan J, Malakhova M, Luo K, Li Y, Ryu J, et al. Resveratrol induces apoptosis by directly targeting Ras-GTPase activating protein SH3 domain binding protein 1 (G3BP1). Oncogene. 2015 May 14;34(20):2660–71.

105. Kennedy D, French J, Guitard E, Ru K, Tocque B, Mattick J. Characterization of G3BPs: Tissue specific expression, chromosomal localisation and rasGAP120 binding studies. Journal of Cellular Biochemistry. 2002;84(1):173–87.

106. Annibaldi A, Dousse A, Martin S, Tazi J, Widmann C. Revisiting G3BP1 as a RasGAP Binding Protein: Sensitization of Tumor Cells to Chemotherapy by the RasGAP 317–326 Sequence Does Not Involve G3BP1. PLoS One. 2011 Dec 19;6(12):e29024.

107. Feng X, Li L, Wagner EJ, Li W. TC3A: The Cancer 3′ UTR Atlas. Nucleic Acids Res. 2018 Jan 4;46(Database issue):D1027–30.

108. Climente-González H, Porta-Pardo E, Godzik A, Eyras E. The Functional Impact of Alternative Splicing in Cancer. Cell Reports. 2017 Aug 29;20(9):2215–26.

109. Trincado JL, Sebestyén E, Pagés A, Eyras E. The prognostic potential of alternative transcript isoforms across human tumors. Genome Medicine. 2016 Aug 17;8(1):85.

110. Lu Z xiang, Huang Q, Park JW, Shen S, Lin L, Tokheim CJ, et al. Transcriptome-wide Landscape of Pre-mRNA Alternative Splicing Associated with Metastatic Colonization. Molecular Cancer Research. 2015 Feb 19;13(2):305–18.

111. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018 Feb;15(2):81–94.

112. Vitale I, Shema E, Loi S, Galluzzi L. Intratumoral heterogeneity in cancer progression and response to immunotherapy. Nat Med. 2021 Feb;27(2):212–24.

113. Ho WJ, Erbe R, Danilova L, Phyo Z, Bigelow E, Stein-O'Brien G, et al. Multi-omic profiling of lung and liver tumor microenvironments of metastatic pancreatic cancer reveals site-specific immune regulatory pathways. Genome Biology. 2021 May 13;22(1):154.

114. Amadori D, Volpi A, Maltoni R, Nanni O, Amaducci L, Amadori A, et al. Cell proliferation as a predictor of response to chemotherapy in metastatic breast cancer: A prospective study. Breast Cancer Res Treat. 1997 Mar 1;43(1):7–14.

115. Janssen LME, Ramsay EE, Logsdon CD, Overwijk WW. The immune system in cancer metastasis: friend or foe? Journal for ImmunoTherapy of Cancer. 2017 Oct 17;5(1):79.

116. Deng W, Yan M, Yu T, Ge H, Lin H, Li J, et al. Quantitative Proteomic Analysis of the Metastasis-Inhibitory Mechanism of miR-193a-3p in Non-Small Cell Lung Cancer. CPB. 2015;35(5):1677–88.

117. He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, et al. A microRNA polycistron as a potential human oncogene. Nature. 2005 Jun;435(7043):828–33.

118. Wu K, Feng J, Lyu F, Xing F, Sharma S, Liu Y, et al. Exosomal miR-19a and IBSP cooperate to induce osteolytic bone metastasis of estrogen receptor-positive breast cancer. Nat Commun. 2021 Aug 31;12:5196.

119. Lee S, Lee H, Bae H, Choi EH, Kim SJ. Epigenetic silencing of miR-19a-3p by cold atmospheric plasma contributes to proliferation inhibition of the MCF-7 breast cancer cell. Sci Rep. 2016 Jul 21;6:30005.

120. MacCuaig WM, Thomas A, Claros-Sorto JC, Gomez-Gutierrez JG, Alexander AC, Wellberg EA, et al. Differential expression of microRNA between triple negative breast cancer patients of African American and European American descent. Biotechnic & Histochemistry. 2022 Jan 2;97(1):1–10.

121. Ouchida M, Kanzaki H, Ito S, Hanafusa H, Jitsumori Y, Tamaru S, et al. Novel Direct Targets of miR-19a Identified in Breast Cancer Cells by a Quantitative Proteomic Approach. PLOS ONE. 2012 Aug 30;7(8):e44095.

122. Zhan Y, Liang X, Li L, Wang B, Ding F, Li Y, et al. MicroRNA-548j functions as a metastasis promoter in human breast cancer by targeting Tensin1. Mol Oncol. 2016 Jun;10(6):838–49.

123. Zhao B, Zhou R, Ji C, Liu D, Wu T, Xu H, et al. The Function of circRNA-0047604 in Regulating the Tumor Suppressor Gene DACH1 in Breast Cancer. Biomed Res Int. 2022 Jan 20;2022:6589651.

124. Cai C, Huo Q, Wang X, Chen B, Yang Q. SNHG16 contributes to breast cancer cell migration by competitively binding miR-98 with E2F5. Biochemical and Biophysical Research Communications. 2017 Apr 1;485(2):272–8.

125. Siragam V, Rutnam ZJ, Yang W, Fang L, Luo L, Yang X, et al. MicroRNA miR-98 inhibits tumor angiogenesis and invasion by targeting activin receptor-like kinase-4 and matrix metalloproteinase-11. Oncotarget. 2012 Nov 6;3(11):1370–85.

126. Si H, Chen P, Li H, Wang X. Long non-coding RNA H19 regulates cell growth and metastasis via miR-138 in breast cancer. Am J Transl Res. 2019 May 15;11(5):3213–25.

127. Zhang J, Liu D, Feng Z, Mao J, Zhang C, Lu Y, et al. MicroRNA-138 modulates metastasis and EMT in breast cancer cells by targeting vimentin. Biomedicine & Pharmacotherapy. 2016 Feb 1;77:135–41.

128. Liang Z, Feng Q, Xu L, Li S, Zhou L. CREPT regulated by miR-138 promotes breast cancer progression. Biochemical and Biophysical Research Communications. 2017 Nov 4;493(1):263–9.

129. Døssing KBV, Binderup T, Kaczkowski B, Jacobsen A, Rossing M, Winther O, et al. Down-Regulation of miR-129-5p and the let-7 Family in Neuroendocrine Tumors and Metastases Leads to Up-Regulation of Their Targets Egr1, G3bp1, Hmga2 and Bach1. Genes (Basel). 2014 Dec 24;6(1):1–21.

130. Wang Z, Chang X, Zhu G, Gao X, Chang L. Depletion of lncRNA MALAT1 inhibited sunitinib resistance through regulating miR-362-3p-mediated G3BP1 in renal cell carcinoma. Cell Cycle. 19(16):2054–62.

131. Park OH, Ha H, Lee Y, Boo SH, Kwon DH, Song HK, et al. Endoribonucleolytic Cleavage of m6A-Containing RNAs by RNase P/MRP Complex. Molecular Cell. 2019 May 2;74(3):494-507.e8.

132. Guo W, Zhang L, Christopher DM, Teng ZQ, Fausett SR, Liu C, et al. RNA-Binding Protein FXR2 Regulates Adult Hippocampal Neurogenesis by Reducing Noggin Expression. Neuron. 2011 Jun 9;70(5):924–38.

133. Wu X, Zhao W, Cui Q, Zhou Y. Computational screening of potential regulators for mRNA-protein expression level discrepancy. Biochemical and Biophysical Research

Communications [Internet]. 2019 Dec 14 [cited 2019 Dec 20]; Available from: http://www.sciencedirect.com/science/article/pii/S0006291X19323721

134. Jain S, Wheeler JR, Walters RW, Agrawal A, Barsic A, Parker R. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. Cell. 2016 Jan 28;164(3):487–98.

135. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC. Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics. 2011 Dec 15;12:474.

136. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318–30.

137. Ficarelli M, Neil SJD, Swanson CM. Targeted Restriction of Viral Gene Expression and Replication by the ZAP Antiviral System. Annu Rev Virol [Internet]. 2021 Sep 29 [cited 2021 Jun 17]; Available from: https://www.annualreviews.org/doi/10.1146/annurev-virology-091919-104213

138. Hayakawa S, Shiratori S, Yamato H, Kameyama T, Kitatsuji C, Kashigi F, et al. ZAPS is a potent stimulator of signaling mediated by the RNA helicase RIG-I during antiviral responses. Nature Immunology. 2011 Jan;12(1):37–44.

139. Chen S, Xu Y, Zhang K, Wang X, Sun J, Gao G, et al. Structure of N-terminal domain of ZAP indicates how a zinc-finger protein recognizes complex RNA. Nature Structural & Molecular Biology. 2012 Apr;19(4):430–5.

140. Schwerk J, Soveg FW, Ryan AP, Thomas KR, Hatfield LD, Ozarkar S, et al. RNA-binding protein isoforms ZAP-S and ZAP-L have distinct antiviral and immune resolution functions. Nat Immunol. 2019 Dec;20(12):1610–20.

141. Zhu Y, Wang X, Goff SP, Gao G. Translational repression precedes and is required for ZAP-mediated mRNA decay. EMBO J. 2012 Nov 5;31(21):4236–46.

142. Chen G, Guo X, Lv F, Xu Y, Gao G. p72 DEAD box RNA helicase is required for optimal function of the zinc-finger antiviral protein. Proc Natl Acad Sci U S A. 2008 Mar 18;105(11):4352–7.

143. Guo X, Ma J, Sun J, Gao G. The zinc-finger antiviral protein recruits the RNA processing exosome to degrade the target mRNA. Proc Natl Acad Sci U S A. 2007 Jan 2;104(1):151–6.

144. Zhu Y, Chen G, Lv F, Wang X, Ji X, Xu Y, et al. Zinc-finger antiviral protein inhibits HIV-1 infection by selectively targeting multiply spliced viral mRNAs for degradation. Proc Natl Acad Sci U S A. 2011 Sep 20;108(38):15834–9.

145. Ye P, Liu S, Zhu Y, Chen G, Gao G. DEXH-Box protein DHX30 is required for optimal function of the zinc-finger antiviral protein. Protein Cell. 2010 Oct;1(10):956–64.

146. Gläsker S, Töller M, Kümmerer BM. The alternate triad motif of the poly(ADP-ribose) polymerase-like domain of the human zinc finger antiviral protein is essential for its antiviral activity. Journal of General Virology. 2014;95(4):816–22.

147. Kmiec D, Lista MJ, Ficarelli M, Swanson CM, Neil SJD. S-farnesylation is essential for antiviral activity of the long ZAP isoform against RNA viruses with diverse replication strategies. PLoS Pathog. 2021 Oct 25;17(10):e1009726.

148. Charron G, Li MMH, MacDonald MR, Hang HC. Prenylome profiling reveals S-farnesylation is crucial for membrane targeting and antiviral activity of ZAP long-isoform. Proc Natl Acad Sci U S A. 2013 Jul 2;110(27):11085–90.

149. Li MMH, Aguilar EG, Michailidis E, Pabon J, Park P, Wu X, et al. Characterization of Novel Splice Variants of Zinc Finger Antiviral Protein (ZAP). Journal of Virology. 2019 Aug 28;93(18):e00715-19.

150. Mao R, Nie H, Cai D, Zhang J, Liu H, Yan R, et al. Inhibition of Hepatitis B Virus Replication by the Host Zinc Finger Antiviral Protein. PLoS Pathog [Internet]. 2013 Jul 11 [cited 2018 Dec 11];9(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3708887/

151. Takata MA, Gonçalves-Carneiro D, Zang TM, Soll SJ, York A, Blanco-Melo D, et al. CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature. 2017 Oct;550(7674):124–7.

152. Meagher JL, Takata M, Gonçalves-Carneiro D, Keane SC, Rebendenne A, Ong H, et al. Structure of the zinc-finger antiviral protein in complex with RNA reveals a mechanism for selective targeting of CG-rich viral sequences. PNAS [Internet]. 2019 Nov 12 [cited 2019 Nov 18]; Available from: https://www.pnas.org/content/early/2019/11/11/1913232116

153. Gao G, Guo X, Goff SP. Inhibition of Retroviral RNA Production by ZAP, a CCCH-Type Zinc Finger Protein. Science. 2002 Sep 6;297(5587):1703–6.

154. Guo X, Carroll JWN, MacDonald MR, Goff SP, Gao G. The Zinc Finger Antiviral Protein Directly Binds to Specific Viral mRNAs through the CCCH Zinc Finger Motifs. J Virol. 2004 Dec;78(23):12781–7.

155. Todorova T, Bock FJ, Chang P. PARP13 regulates cellular mRNA post-transcriptionally and functions as a pro-apoptotic factor by destabilizing TRAILR4 transcript. Nature Communications [Internet]. 2014 Dec [cited 2018 Jul 20];5(1). Available from: http://www.nature.com/articles/ncomms6362

156. Gonzalez-Perez AC, Stempel M, Wyler E, Urban C, Piras A, Hennig T, et al. The Zinc Finger Antiviral Protein ZAP Restricts Human Cytomegalovirus and Selectively Binds and Destabilizes Viral *UL4* / *UL5* Transcripts. Horner SM, editor. mBio [Internet]. 2021 May 4 [cited 2021 Jun 9]; Available from: https://journals.asm.org/doi/10.1128/mBio.02683-20

157. Zhang Y, Burke CW, Ryman KD, Klimstra WB. Identification and Characterization of Interferon-Induced Proteins That Inhibit Alphavirus Replication. J Virol. 2007 Oct;81(20):11246–55.

158. MacDonald MR, Machlin ES, Albin OR, Levy DE. The Zinc Finger Antiviral Protein Acts Synergistically with an Interferon-Induced Factor for Maximal Activity against Alphaviruses. Journal of Virology. 2007 Dec 15;81(24):13509–18.

159. Karki S, Li MMH, Schoggins JW, Tian S, Rice CM, MacDonald MR. Multiple Interferon Stimulated Genes Synergize with the Zinc Finger Antiviral Protein to Mediate Anti-Alphavirus Activity. PLOS ONE. 2012 May 16;7(5):e37398.

160. Li MMH, Lau Z, Cheung P, Aguilar EG, Schneider WM, Bozzacco L, et al. TRIM25 Enhances the Antiviral Action of Zinc-Finger Antiviral Protein (ZAP). PLOS Pathogens. 2017 Jan 6;13(1):e1006145.

161. Zheng X, Wang X, Tu F, Wang Q, Fan Z, Gao G. TRIM25 Is Required for the Antiviral Activity of Zinc Finger Antiviral Protein. J Virol [Internet]. 2017 Apr 13 [cited 2018 Dec 10];91(9). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391446/

162. Zou W, Zhang DE. The Interferon-inducible Ubiquitin-protein Isopeptide Ligase (E3) EFP Also Functions as an ISG15 E3 Ligase. J Biol Chem. 2006 Feb 17;281(7):3989–94.

163. Choudhury NR, Heikel G, Trubitsyna M, Kubik P, Nowak JS, Webb S, et al. RNA-binding activity of TRIM25 is mediated by its PRY/SPRY domain and is required for ubiquitination. BMC Biol [Internet]. 2017 Nov 8 [cited 2018 Dec 15];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5678581/

164. Sanchez JG, Sparrer KMJ, Chiang C, Reis RA, Chiang JJ, Zurenski MA, et al. TRIM25 binds RNA to modulate cellular anti-viral defense. J Mol Biol. 2018 Dec 7;430(24):5280–93.

165. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016 Jun;13(6):508–14.

166. Kaewsapsak P, Shechner DM, Mallard W, Rinn JL, Ting AY. Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. Gavis ER, editor. eLife. 2017 Dec 14;6:e29224.

167. Villarroya-Beltri C, Guerra S, Sánchez-Madrid F. ISGylation – a key to lock the cell gates for preventing the spread of threats. J Cell Sci. 2017 Sep 15;130(18):2961–9.

168. Zhao C, Denison C, Huibregtse JM, Gygi S, Krug RM. Human ISG15 conjugation targets both IFN-induced and constitutively expressed proteins functioning in diverse cellular pathways. PNAS. 2005 Jul 19;102(29):10200–5.

169. Youn JY, Dunham WH, Hong SJ, Knight JDR, Bashkurov M, Chen GI, et al. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. Molecular Cell. 2018 Feb 1;69(3):517-532.e11.

170. Khong A, Matheny T, Jain S, Mitchell SF, Wheeler JR, Parker R. The stress granule transcriptome reveals principles of mRNA accumulation in stress granules. Mol Cell. 2017 Nov 16;68(4):808-820.e5.

171. Konieczny P, Stepniak-Konieczna E, Taylor K, Sznajder ŁJ, Sobczak K. Autoregulation of MBNL1 function by exon 1 exclusion from MBNL1 transcript. Nucleic Acids Res. 2017 Feb 28;45(4):1760–75.

172. Pervouchine D, Popov Y, Berry A, Borsari B, Frankish A, Guigó R. Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay. Nucleic Acids Res. 2019 Jun 4;47(10):5293–306.

173. Zhang Y, Burkhardt DH, Rouskin S, Li GW, Weissman JS, Gross CA. A stress response that monitors and regulates mRNA structure is central to cold-shock adaptation. Mol Cell. 2018 Apr 19;70(2):274-286.e7.

174. Sanchez JG, Chiang JJ, Sparrer KMJ, Alam SL, Chi M, Roganowicz MD, et al. Mechanism of TRIM25 Catalytic Activation in the Antiviral RIG-I Pathway. Cell Reports. 2016 Aug 2;16(5):1315–25.

175. Askjaer P, Kjems J. Mapping of Multiple RNA Binding Sites of Human T-cell Lymphotropic Virus Type I Rex Protein within 5′- and 3′-Rex Response Elements *. Journal of Biological Chemistry. 1998 May 8;273(19):11463–71.

176. Brown RS, Anastasakis DG, Hafner M, Kielian M. Multiple capsid protein binding sites mediate selective packaging of the alphavirus genomic RNA. Nat Commun. 2020 Sep 17;11(1):4693.

177. Sharma D, Zagore LL, Brister MM, Ye X, Crespo-Hernández CE, Licatalosi DD, et al. The kinetic landscape of an RNA-binding protein in cells. Nature. 2021 Feb 10;1–5.

178. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

179. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019 May 7;47(8):e47.

180. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol [Internet]. 2014 [cited 2019 Jan 19];15(12). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/

181. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000 Jun 1;16(6):276–7.

182. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010 May 28;38(4):576–89.

183. Tomezsko PJ, Corbin VDA, Gupta P, Swaminathan H, Glasgow M, Persad S, et al. Determination of RNA structural diversity and its role in HIV-1 RNA splicing. Nature. 2020 Jun;582(7812):438–42.

184. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990 Aug 3;249(4968):505–10.

185. Zykovich A, Korf I, Segal DJ. Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. Nucleic Acids Res. 2009 Dec;37(22):e151.

186. Huppertz I, Attig J, D'Ambrogio A, Easton LE, Sibley CR, Sugimoto Y, et al. iCLIP: Protein–RNA interactions at nucleotide resolution. Methods. 2014 Feb;65(3):274–87.

187. Brannan KW, Chaim IA, Marina RJ, Yee BA, Kofman ER, Lorenz DA, et al. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. Nature Methods. 2021 May 1;18(5):1–13.

188. VanInsberghe M, van den Berg J, Andersson-Rolf A, Clevers H, van Oudenaarden A. Single-cell Ribo-seq reveals cell cycle-dependent translational pausing. Nature. 2021 Sep;597(7877):561–5.

Appendix A: RNA structure impacts circular RNA regulation and function

Methods 196 (2021) 56–67

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Thrown for a (stem) loop: How RNA structure impacts circular RNA regulation and function

Veronica F. Busa [a,b], Anthony K.L. Leung [a,b,c,d,*]

[a] McKusick-Nathans Institute of the Department of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[b] Department of Biochemistry and Molecular Biology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA
[c] Department of Molecular Biology and Genetics, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA
[d] Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

ARTICLE INFO

Keywords:
Circular RNA
RNA secondary structure

ABSTRACT

Exonic circular RNAs (circRNAs) are RNA molecules that are covalently closed by back-splicing via canonical splicing machinery. Despite overlapping sequences, exon circularization generates RNA secondary structures through intramolecular base-pairing that are different from the linear transcript. Here we review factors that may affect circRNA structure and how structure affects circRNA function and regulation. We highlight considerations for RNA sequencing and expression measurement to ensure highly structured circRNAs are accurately represented by the data and discuss issues that need to be addressed in generating circRNAs to recapitulate their endogenous structures. We conclude our review by discussing experimental strategies on revealing the varied roles of RNA structure in circRNA biogenesis, function and decay.

1. Introduction

Circular RNAs (circRNAs) are generated during RNA splicing, where the 3′ end of a downstream exon splices to the 5′ end of an upstream exon to form a back-splice junction [1–3]. CircRNAs are highly stable, since they are covalently closed and therefore resistant to RNA decay pathways initiated by exoribonucleases [4]. Despite their resistance to most major RNA decay pathways, circRNA expression can dynamically change during development [5,6], suggesting the existence of programs for coordinated, selective circRNA biogenesis and decay.

RNA structure characterization and circRNA regulation are two rapidly emerging fields. However, it is becoming increasingly apparent that the two fields must intersect to thoroughly examine and understand circRNA function and regulation. For example, RNA structure plays a role in circRNA biogenesis and is regulated via base-pairing of complementary intronic sequences and RNA-binding proteins [1,2,7–12]. Once formed, the structure of a circRNA can affect its function, such as in innate immune regulation [13], or in its decay, with highly-structured circRNAs potentially evading short RNA-mediated cleavage but being targeted by other pathways such as structure-mediated RNA decay [14].

Many factors may affect circRNA structure, such as bound proteins, RNA modification, or translation status [15–17]. Although circRNA and cognate linear mRNA sequences overlap, RNA circularization affects transcript structure, presumably by bringing nucleotides into proximity for base-pairings that would not otherwise have been favorable [18]. While circRNA structure can be determined using either exclusively *in silico* prediction algorithms or augmented by RNA chemical probing information, special attention is required to tweak these protocols for establishing the structure of a circular rather than linear transcript [13,19].

Besides the challenges of determining circRNA structure, highly structured circRNAs may remain unidentified in sequencing data and undetected when measuring expression levels [20]. Noisy data can make it difficult to identify circRNAs from sequencing results, but prudent sample preparation and data processing can increase confidence in circRNA identification [21]. Similarly, when measuring expression, it is

Abbreviations: circRNA, circular RNA; EJC, exon junction complex; siRNAs, small interfering RNAs; shRNAs, short hairpin RNAs; ASOs, antisense oligonucleotides; m6A, N6-methyladenosine, dsRNA double-stranded RNA, IRES internal ribosome entry site; SHAPE-MaP, selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling; RPAD, RNase R Treatment, polyadenylation, and poly(A)+ RNA depletion; SRD, structure-mediated RNA decay; AAV, adeno-associated virus; Tornado, Twister-optimized RNA for durable overexpression; LDN, linear DNA nanostructure; tricRNAs, tRNA intron-derived circular RNAs; TSEN, tRNA splicing endonuclease.

* Corresponding author at: 615 North Wolfe Street, Johns Hopkins Bloomberg School of Public Health E3132, Baltimore, MD 21205, USA.
E-mail address: anthony.leung@jhu.edu (A.K.L. Leung).

https://doi.org/10.1016/j.ymeth.2021.02.019
Received 1 December 2020; Received in revised form 9 February 2021; Accepted 26 February 2021
Available online 2 March 2021
1046-2023/© 2021 Elsevier Inc. All rights reserved.

important to select a method that is appropriate for detecting highly structured circRNAs and consider whether circRNAs generated *in vitro* or in cells recapitulate endogenous structures.

In this review, we first survey how circRNA structures are characterized and determined, followed by how circRNAs can be generated *in vitro* and in cells. We discuss special considerations of RNA structures necessary when identifying, measuring, and generating circRNAs. We conclude with an overview of known mechanisms by which RNA structures impact circRNA biogenesis, function and decay, with considerations of how existing tools can be optimized to delineate their mechanisms.

2. Characterizing circRNA structure

2.1. circRNA and linear RNA sequences overlap

circRNA is derived from the same pre-mRNA exons as linear mRNA [22]. Because circRNAs contain sequences almost entirely identical to their linear cognate RNAs, it is difficult to study circRNA independent of linear RNA [23]. Changing the sequence of a circRNA often affects the cognate linear transcript sequence. Although recent CRISPR Cas13 RNA editing technologies have emerged that allow for circRNA-specific knock-down by targeting sequences spanning the back-splice junction [24–26], these methods are limited to studying the effect of reduced circRNA levels, localization of circRNAs, and effects of localizing proteins to the RNA sequence; they do not allow for circRNA-specific sequence modification, knockout, or upregulation. The back-splice junction region is the only sequence unique to the circRNA (BSJ in Fig. 1A), and this region is often targeted in knock-down assays by small interfering RNAs (siRNAs), short hairpin RNAs (shRNAs), and antisense oligonucleotides (ASOs) [21]. But it has been observed that even this strategy is not always effective, since if a circRNA is structured at the back-splice junction, the probe may not anneal [27,28]. The probe may therefore have to target a sequence that is entirely contained in both the linear and circRNA transcripts. Shifting the target sequence slightly upstream or downstream of the back-splice junction, such that the siRNA target sequence is contained in both the linear and circular transcripts, has been shown to sometimes reduce circRNA but not linear RNA expression levels [21,29]. This selective siRNA-mediated silencing suggests that in some circumstances there is secondary structure in the linear transcript near the splice site that prevents siRNA annealing that is absent in the circRNA. The inverse may also occur: circRNA regions near the back-splice junction may be structured whereas the linear cognate RNA is not, which may allow microRNA or siRNA binding to mRNAs but not circRNAs with the same sequence. Methods to detect the presence of circRNAs often also rely on the complementarity of a probe to the

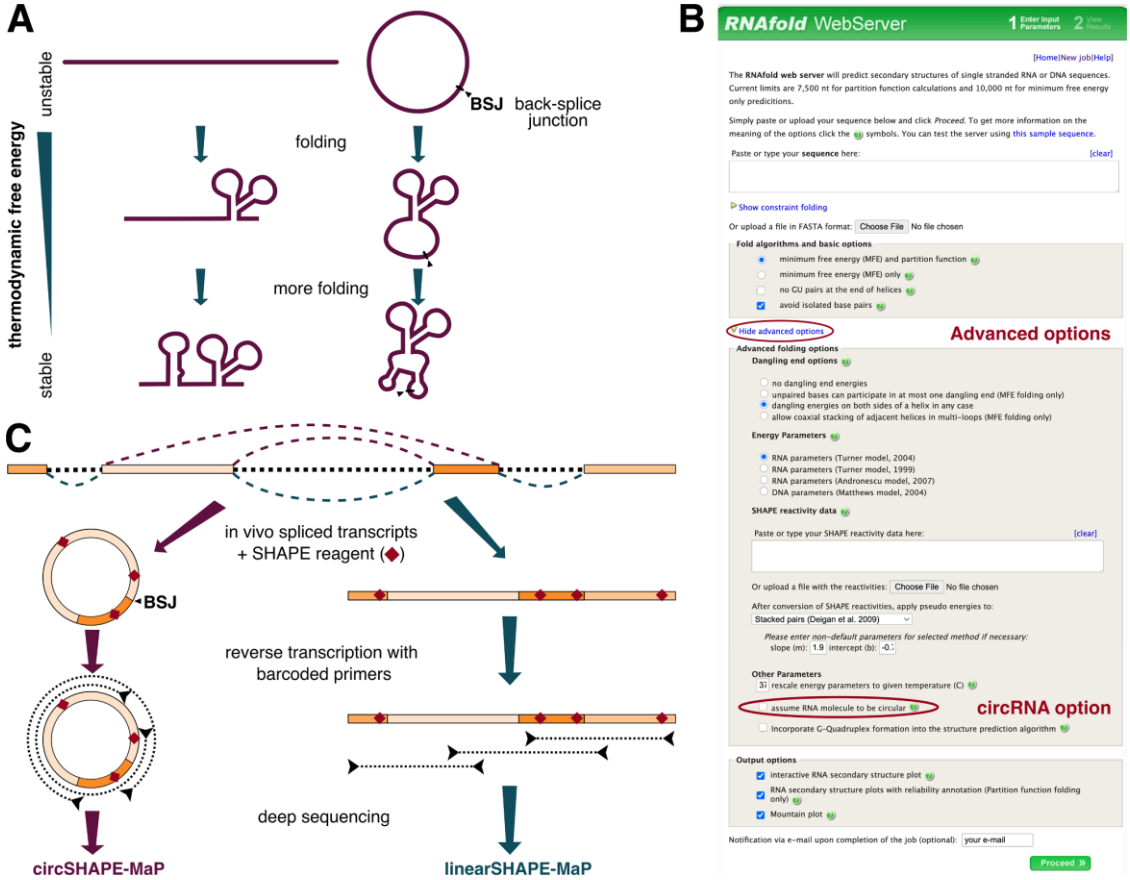


Fig. 1. CircRNA structure analyses. A. An in silico folding model of hypothetical linear and circRNA transcripts of the same sequence. Although both transcripts may adopt similar structures in some regions, the covalent linkage of the circRNA may drive it to fold into structures distinct from the linear transcript across the back-splice junction (BSJ). B. A screenshot of the RNAfold webserver, with options to specify that a sequence is for a circRNA emphasized. C. An illustration of the strategy for library construction of circSHAPE-MaP and linearSHAPE-MaP datasets.

circRNA (e.g. Northern blot [30,31] or fluorescent probe [32]), which may similarly be affected by RNA structure (see Section 5 for more discussion).

2.2. RNA circularization affects transcript structure

The covalent conjugation of circRNAs at the back-splice junction causes circRNA structure to differ from linear RNA structure, despite similarities in sequence (*cf.* Fig. 1A). For example, a comparison of folding models for the actual circular structure of the 284-nucleotide citrus viroid IV genome versus all possible linear structures formed by hypothetically cleaving the circular RNA at random locations revealed that some linear structures were predicted to fold more stably than the actual circular RNA genome (up to 7 kcal/mol difference), while other linear structures were less stable (up to 3 kcal/mol) [18]. Further, 53 of the 284 nucleotides in the citrus viroid IV genome were predicted to have different base-paired nucleotide partners when folded *in silico* as a circular RNA versus a linear RNA [18] (see Section 3.1 for more details on *in silico* folding), which indicates that the predicted structure for up to one-fifth of this circular genome could differ between linear and circular models. The differences between linear and circular RNA folding stability is likely because circRNA brings otherwise distant nucleotides into proximity and provides base-pairing opportunities that would not have been available otherwise.

In addition to *in silico* folding predictions, a recent chemical probing study revealed that some circRNAs important for immune regulation contain structured regions at least 16 base pairs in length, with 9 of them incorporating their back-splice junction into these long, structured regions [13] (see Sections 3 and 7). These data further suggest that these highly structured regions might not have been able to form if not for the covalent closure of the transcript.

2.3. Factors that affect circRNA structure

circRNA structure may differ from linear RNA structure due to their translation statuses. The majority of circRNAs do not associate with ribosomes [7,33], so sequences that are occupied by ribosomes in the linear transcript have the opportunity to form secondary structures and bind proteins in the corresponding circRNA.

Protein binding differences between circRNA and linear RNA may stabilize or alter structure. Since circRNA is formed via canonical splicing of two non-sequential splice sites [1–3], the exon junction complex (EJC) is likely also deposited at the back-splice junction [15,34]. Without translation, the EJC can remain at the junction and potentially affect circRNA folding. Or multiple proteins may bind a single circRNA at distant locations and associate with each other, causing the circRNA to form loops spanning hundreds of nucleotides [35]. Notably, some proteins preferentially bind circRNA over their linear cognate transcripts, probably because circRNA structure is more stable [13]. CircRNA stabilization has been functionally exemplified by the fact that the fluorogenic aptamer Broccoli, which must be properly folded to function, exhibits ~ 50% more fluorescence in a circular RNA than its linear counterpart [36].

CircRNA structure is intrinsically affected by nucleotide sequence. Modification, such as N^6-methyladenosine (m^6A), can also affect base-pairing, either by stabilizing RNA structure at loops or repelling the formation of dsRNA regions [37]. At least 13% of circRNAs are m^6A-modified [17]. CircRNAs have different m^6A modification patterns from their linear counterpart [16], perhaps to modify RNA structure and facilitate binding of select proteins such as in the case of regulating innate immunity [38,39].

Although most circRNAs do not associate with ribosomes, are not translated, and therefore are able to maintain secondary structures, there are two cap-independent translation initiation mechanisms of circRNAs that are driven by circRNA modification and structure: m^6A modification and internal ribosome entry site (IRES) [40]. A single m^6A modification on a circRNA is sufficient to initiate translation via interactions with the initiation factor eIF4G2 and the m^6A reader YTHDF3, and 250 circRNAs have been shown to associate with polysomes, implying that they have translation potential [17,41]. IRES-mediated translation is a widely-accepted mechanism of circRNA translation initiation, and functional proteins have been generated from engineered IRES-containing circRNAs [42–45]. Since the functionality of IRESs depend more on structure than sequence, it is unknown how many endogenous circRNAs are actively translated [43]. This knowledge gap may be resolved as methods of determining circRNA structure are expanded and improved.

3. Methods for determining circRNA structure

Although the foundations of methods for determining circRNA structure are the same as those to characterize the structure of linear transcripts, complications may arise when accounting for the back-splice junction in the data collection. For *in silico* structure prediction methods, this complication called for modification to free energy calculation algorithms. For experimental approaches, of which selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) is the most common, experimental design must differentiate linear RNA and circRNA structure data.

3.1. In silico

In silico RNA secondary structure prediction algorithms use an experimentally-defined set of RNA folding energy rules to calculate probable RNA structure [19,46]. The calculations to predict secondary structure of circRNA *in silico* are not very different from those for a linear transcript (Fig. 1A). The only difference between a linear and circular transcript is the covalent linkage between the last nucleotide and the first nucleotide in the circRNA. In a linear transcript, there is no energy contribution from the linkage that incorporates both these nucleotides, since these two terminal nucleotides are not connected. In circRNA, energy contributions of nucleotides involved in the back-splice junction must be scored like any other conformation formed from adjacent nucleotides. The energy contribution is calculated by assessing whether nucleotides to either side of the back-splice junction might form more favorable structures than the dangling-end contributions, were the nucleotides not covalently bound [18].

RNAfold from the ViennaRNA package [19] is the most commonly used program to predict circRNA structure *in silico* [13,14]. Although MFold [47] can also accommodate circRNA structure prediction, it has larger memory requirements than RNAfold [18]. RNAfold is easily downloadable and useable on all three major operating systems (Windows, MacOS, and Linux) and only requires a FASTA file containing the circRNA sequence. To fold a circRNA via RNAfold, the user must simply type:

RNAfold -c [< *circRNA_sequence.fa* >]

RNAfold is also available as a web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAfold.cgi), where users must paste the circRNA sequence and select "assume RNA molecule to be circular" under the advanced options (Fig. 1B). There is currently a 7,500 nt limit for folding calculations via the web server.

Linear versus circRNA structure may have dramatic ramifications for the predicted free energy of the structure, depending on the energy contributions of structures involving the back-splice junction [18]. Because each nucleotide can only have one base-paired nucleotide partner, a stable secondary structure formed at the junction may cause a cascade of secondary structure rearrangement relative to the linear transcript. Therefore, if structure at the back-splice junction outcompetes the structure in the linear transcript, there is a folding energy difference between the circular and linear transcripts [46].

Although this mathematical approach to circRNA structure calculations is logically sound, *in silico* algorithms assume that the RNA is naked

in solution and cannot account for other factors, such as binding proteins, that may affect RNA folding. For example, experimental analyses have demonstrated that secondary structure forms at the back-splice junction in cells [13], but no experiment has been performed to study the influence of the EJC on circRNA structure so it is unclear whether the EJC must be removed from the junction in order for the circRNA to adopt this structure or how accurate *in silico* circRNA folding predictions are, especially around the back-splice junction.

3.2. Chemical probing

A wide variety of chemical probing approaches have been used to characterize RNA structure at both large and small scales [48–52]. Of these, SHAPE-MaP has emerged as an especially useful tool since the reagents are able to probe exposed (i.e. not buried or base-paired) nucleotides and the massively parallel sequencing allows collection of more robust data [53]. Reactivity information is then input into an *in silico* folding algorithm to act as constraints to improve the accuracy of RNA folding [54,55].

When conducting targeted SHAPE-MaP on linear RNA, each nucleotide will be read once at most, since the reverse transcriptase will fall off once it reaches the end of the transcript. In contrast, since circRNAs do not have free ends, the polymerase has the opportunity to process around the transcript until it stalls and dissociates; therefore, there is the possibility that regions of a circRNA may be redundantly recorded and disproportionately represented in the data unless the experiment is properly designed.

One way to selectively characterize linear (linearSHAPE-MaP) or circRNA (circSHAPE-MaP) structure is to position primer pairs to enrich regions specific to each transcript (Fig. 1C). For example, structure information for linear RNA can be enriched by priming at exons that are exclusively present in the linear transcript for linearSHAPE-MaP [13]. Primer positioning to enrich circRNA structure information is more difficult, since all circRNA exons are often also present in the cognate linear RNA. However, primer pairs that span the back-splice junction cannot amplify linear reads, allowing for selection of circRNA reads in circSHAPE-MaP [13]. LinearSHAPE-MaP and circSHAPE-MaP libraries are prepared separately by different indexes to distinguish each other. Since the standard linearSHAPE-MaP ShapeMapper pipeline [55,56] only considers canonical splice sites when aligning reads, circSHAPE-MaP requires additional computational steps to align reads and map structure at the back-splice junction [13]. After a first round of sequencing read alignment, reads for which only the beginning or end is mapped are identified; these partially-mapped reads possibly span the back-splice junction. The junction-spanning candidates are then pooled with reads that were effectively mapped in the first iteration and re-aligned to get complete mapping of linear and circRNA transcripts via improved mapping efficiency at back-splice junctions (source code available at https://github.com/YangLab/CIRCshapemapper) [13].

Both *in silico* and chemical probing methods of determining circRNA structure rely on high-confidence circRNA annotation information. Chemical probing requires strategic primer design and *in silico* structure prediction requires a complete and accurate circRNA sequence. It is therefore vital to ensure the entire circRNA sequence, not just the location of the back-splice junction, is well-characterized when identifying circRNAs before attempting to discern circRNA structure.

4. Considerations of RNA structure when identifying circRNAs

4.1. Sequencing data pre-processing can affect available circRNA data

Multiple methods have been developed to enrich for circRNAs over other prominent RNA populations, for example, by the use of exoribonuclease RNaseR to remove linear RNAs [21]. However, some highly-structured and G-quadruplex-containing linear RNAs are also resistant to RNase R digestion [57]. Enhanced RNase R treatment, based on RPAD (RNase R Treatment, Polyadenylation, and Poly(A) + RNA Depletion), is the most stringent method for circRNA selection, since it resolves these complications and removes all categories of linear transcripts [57,58]. In addition to incorporating a step to add a poly-A tail to non-polyadenylated linear RNAs, enhanced RNase R treatment uses a lithium-containing buffer that destabilizes G-quadruplexes, allowing complete degradation of all linear RNA, even highly-structured transcripts [58]. This level of stringency is important because it minimizes false-positives and mis-characterization of circRNA annotation in downstream analysis steps, especially across highly-structured regions of the transcriptome.

The way in which sample RNA is processed for sequencing can, however, greatly affect the accuracy of circRNA identification by downstream computational analysis. For example, the incorporation of a size selection step during sequencing library preparation might filter out small circRNAs and highly structured circRNAs. Highly structured circRNAs may adopt a more globular structure, which has the dual results of making them smaller and also less prone to fragmentation; both factors can cause highly structured circRNAs to become excluded during size-selection processing steps [20].

4.2. Identifying circRNAs from sequencing data

The universal means by which circRNAs are identified in sequencing data is identification of either paired reads that map out of order or individual reads that span the back-splice junction; however, this approach is prone to errors and false positives [20,59]. Exon length should be considered when filtering computationally-predicted circRNAs for high-confidence hits. Most computationally identified circRNAs smaller than 200 nts are false positives [59], likely because circRNAs have a lower limit for exon size and sequence length. A median exon length of 353 nt is required for single-exon back-splicing [2], and circularized exons are three times longer than average expressed exons (average length 690 nt) [7]. The reason for longer exon preference in circRNA formation has not been explored in-depth, but it may be because short exons are unable to efficiently contort into a structure amenable to back-splicing. Inefficient circular structure formation may be due to the exon adopting structures that prevent the termini from coming into proximity, or due to steric hindrance from proteins bound at the splice sites on the ends of the exon, where similar phenomena were observed in linear transcript splicing [60,61].

One major problem with using reads spanning the back-splice junction to identify circRNAs and quantify circRNA levels is that it does not on its own provide information about the circRNA sequence beyond the nucleotides flanking the junction. For example, unless the junction contains the 5′ and 3′ ends of a single exon, the sequences flanking the back-splice junction are uninformative about whether any introns are retained. Further, if the back-splice junction contains flanking sequences from non-consecutive exons, junction-spanning reads are also uninformative about which exons are included in the circRNA transcript. Intron and exon inclusion information is important since a single linear transcript may produce many circRNAs, and mis-predicting the circRNA sequence may have major ramifications for analyses of its structure and function [2,14,62,63]. Although comparison of untreated versus RNase R-treated can elucidate sequence variations to some degree, it requires individual validation to confirm [2,7,64]. Long-read sequencing (e.g., by nanopore) is capable of completely covering the length of the average circRNA, thereby providing a high-throughput option to confidently determine a circRNA sequence, especially for circRNAs prone to alternative splicing [65,66].

5. Consideration of RNA structure when measuring circRNA expression

Every method for measuring circRNA expression levels has benefits and disadvantages. Here we address how precisely different methods

can measure circRNA expression and how well they can accommodate highly-structured circRNAs (Table 1).

Northern blot is a traditional method to detect and validate circRNAs, since circRNAs often migrate at a different rate than their linear counterparts [1,2,13,20,30,31,67,68]. Even if some regions of a circRNA are highly structured, probes are relatively easy and inexpensive to produce, so multiple probes can be generated to measure the expression of a single circRNA [1]. However, Northern blot may not have the sensitivity to detect circRNAs that are lowly expressed.

Similar to Northern blot, linear DNA nanostructure (LDN) uses probes to discern whether a circRNA is present [32]. Cyclical annealing and denaturing of the circRNA to the probes amplifies the signal, so that LDN is able to identify the presence of even lowly-expressed circRNAs. However, since circRNA-mediated LDN signal relies on displacement of hairpins into annealed DNA strands, the reaction temperature cannot be elevated to denature the circRNA [32]. The back-splice junction is the only region differentiating the circRNA from linear cognate RNA and structured RNA may be inhibited from binding a probe, so circRNAs with structure at the back-splice junction may not be reliably detectible via LDN.

Quantitative reverse transcription PCR (RT-qPCR) is also able to amplify even lowly-expressed circRNA fragments. Divergent primers flanking the back-splice junction provide specificity for circRNA over cognate linear RNA transcripts, and the hot denaturing step prior to primer annealing makes it largely agnostic to circRNA structure [7,14,20,31,67]. qPCR is more sensitive than Northern blot for measuring circRNA expression, but expression can only be measured relative to other transcripts' expression levels and is prone to overestimation and false positives.

A more recently-developed method, SplintQuant, relies on signal from two DNA oligonucleotides terminally tagged with universal primers that are only ligated and able to undergo qPCR when both are adjacently bound to the target RNA [69]. This arrangement allows specificity of linear versus circRNA signal and accurate quantification of even lowly abundant circRNAs through qPCR without reverse transcription [69]. However, the ligation reaction occurs at 25 °C, and the engineers of this method acknowledge that the secondary structure of circRNAs could impact the efficiency of the splint DNA oligonucleotides annealing to their RNA targets [69]. Since the reaction is stopped by heating to 65 °C, there is a limit to how hot the reaction can be to promote denaturation of RNA structure and annealing of the oligonucleotides [69].

Quantification of circRNAs by biotinylated capture probes and reporter probes [70,71] or by qPCR-based single-molecule detection in droplets via fluorescent signals (RT-ddPCR) [72] allow absolute quantification of the number of circRNA copies, unlike qPCR. Since the technologies rely on probes and incorporate secondary structure-denaturing temperatures, they have the additional benefit that circRNA structure is unlikely to interfere with the methods, so long as the circRNA probes are appropriately designed. These methods unfortunately require specialized equipment and materials that are expensive, which limits accessibility [73].

Table 1
Methods to measure circRNA expression.

Method	Quantitative	Linear vs circular	Structure-sensitive	Price
Northern Blot	No, and insensitive for lowly expressed circRNAs	Depends on probe location and linear vs circular size difference	No, but may need to test multiple probes	$
LDN	No	Depends on structure at back-splice junction	Yes	$$
RT-qPCR	Yes, normalized to other transcripts, but prone to errors	Yes	No	$
SplintQuant	Yes, normalized to other transcripts	Yes	Yes	$$
Biotinylated probes	Yes, normalized to other transcripts	Yes	No	$$$
RT-ddPCR	Yes, provides copy number	Yes	No	$$$

6. Consideration of RNA structure when generating circRNAs

Many methods have been successfully adopted to generate circRNAs [74], some of which have demonstrated preservation of functional RNA structures via fluorescent assays [36]. However, many of these methods modify structures surrounding the back-splice junction, making it impossible to study endogenous back-splice junction-containing circRNA structures (Table 2). Here we focus on methods for mammalian exonic circRNA formation and how they may affect circRNA structure.

6.1. *In vitro—T7 expression vectors*

For *in vitro* formation of circRNA, linear RNA is transcribed from a T7 expression vector via T7 RNA polymerase (Fig. 2A). The linear *in vitro*-transcribed RNA is then incubated with T4 RNA ligase [13] or T4 DNA ligase and a DNA splint [15] to form the circRNA transcript. Since this reaction results in a mixed population—uncircularized linear RNA, linear concatemers, and the desired circRNA product—the products must be run on a gel, excised, and purified [13,39]. T7 transcription from a plasmid and subsequent ligation allows for high circRNA yield, but transcript size can impair circularization efficiency, which limits circRNA expression size to only a few hundred nucleotides [75]. Additionally, it is unknown whether ligation with assistance from a DNA splint may affect circRNA structure at the back-splice junction, since the RNA:DNA complementarity may inhibit the formation of RNA:RNA base-pairing (Section 2.2).

The extensive manipulation required to form and purify circRNA *in vitro* makes it such that the final circRNA structure may not represent structure in cells. Endogenous splicing may interrupt or promote RNA structure formation differently than is evident in the post-purification *in vitro* transcript. There are no EJC or other binding proteins in an *in vitro* reaction, which may affect circRNA secondary structure and their functions [15]. In contrast to circRNAs that are heated before purification [39] or resolved on and excised from a denaturing urea polyacrylamide gel [13], endogenous RNAs are never completely unstructured during processing. Complete denaturation followed by refolding provides the opportunity for RNA to form stable but possibly

Table 2
Methods of generating circRNAs.

Method	Novel structure at back-splice junction	EJC at back-splice junction	Generates linear counterparts
In vitro—T7 expression vectors			
T4 RNA ligase	No	No	Yes, must gel-purify circRNA
T4 DNA ligase + DNA splint	Maybe	No	Yes, must gel-purify circRNA
Complementary flanking ends	Yes	No	Yes, must gel-purify circRNA
In vitro and in cells—Auto-catalytic constructs			
Anabaena group I intron	Yes	No	No
T4 group I intron	Yes	No	No
Tornado expression system	Yes	No	Yes, transiently until RtcB reaction
In cells			
plasmid	No	Yes	Yes
AAV vector	No	Yes	Yes
tricRNA vector	Yes	No	Yes, transiently until RtcB reaction

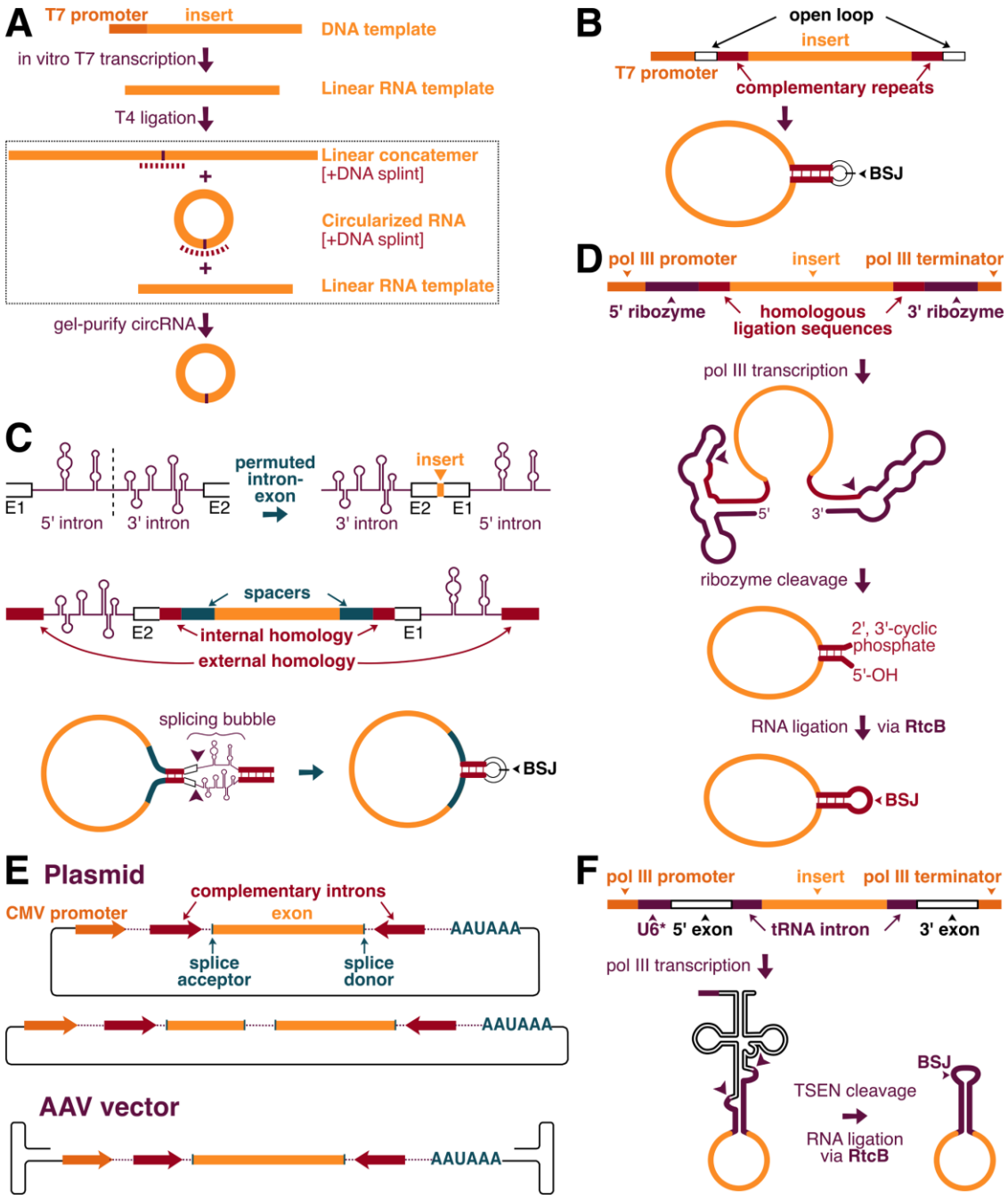


Fig. 2. Methods of generating circRNAs. A. *In vitro* transcription, circularization, and purification of circRNAs. B. Optimized *in vitro* circularization construct, with additional complementary sequences at the termini of the expressed transcript to promote circularization. C. A permuted intron–exon *Anabaena* group I catalytic intron vector. D. The Tornado expression system that can be used *in vitro* and in cells. E. Cellular plasmid circRNA expression vectors containing one (top) or two (middle) exons, and an AAV circRNA expression vector (bottom). Complementary sequences flank exons to promote circularization. F. Cellular tricRNA vector system. U6* includes the first 27 nucleotides of U6 snRNA and enhances stability of the expressed RNA. tRNA splicing endonuclease (TSEN) cleaves at the intron–exon junction and RtcB ligates the introns to form the circRNA.

biologically misrepresentative structures [76], so it is recommended that a native polyacrylamide or an agarose gel without RNA denaturation is used for *in vitro*-transcribed circRNA purification.

To promote RNA circularization, the T7 expression vector may contain complementary flanking regions in the linear T7-transcribed transcript that are able to form a stem structure and bring the 5′ and 3′ ends in close proximity [75,77] (Fig. 2B). This protocol was originally intended to create microRNA sponges, which should not contain excessive structure at microRNA binding sites by design [75,77], but it may not be prudent to create circRNAs for which structure is important for function using this type of vector since the appended artificial hairpin at the back-splice junction may affect circRNA structure. Even with this circularization-promoting modification, circRNA size is limited to a few hundred nucleotides [75].

6.2. In vitro and in cells—Auto-catalytic constructs

The catalytic ability of group I introns from the *Anabaena* pre-tRNA has been harnessed to enable circularization of RNAs up to 4.8 kb and does not require any additional enzymes [45]. Permuted intron–exon splicing constructs consist of fused partial exons that contain the sequence to be circularized flanked by half-intron sequences (Fig. 2C) [78]. When the construct undergoes the group I intron double trans-esterification reaction, the fused exon is excised as a covalently linked circle [78]. The addition of homology arms that flank the half-intron sequences and bring the 5′ and 3′ splice sites into proximity increases splicing efficiency, especially for constructs longer than 300 nt [45]. A comparison of structure predictions for phage T4—another commonly-used group I intron for permuted intron–exon constructs [15,36,78,79]—and *Anabaena* constructs suggests that the latter construct forms secondary structures that aid in the formation of an isolated splicing bubble, which improves splicing efficiency; inclusion of additional spacers improves splicing efficiency even more [45]. The permuted intron–exon *Anabaena* group I splicing construct was originally intended for potent and stable translation in eukaryotic cells, so it is inadvisable to create circRNAs for which structure is important for function using this method since the appended spacers, internal homology arms, and standardized exon termini at the back-splice junction may affect the final circRNA structure [45].

The Tornado (Twister-optimized RNA for durable overexpression) expression system is another means of achieving rapid, high-efficiency RNA circularization without the co-expression of any additional enzymes [36]. Tornado-expressed transcripts contain a sequence to be circularized flanked by Twister ribozymes (Fig. 2D), which undergo autocatalytic cleavage and leave a hydroxyl and a 2′,3′-cyclic phosphate at the 5′ and 3′ ends, respectively. These ends are then recognized and ligated by the RNA ligase RtcB to form a circular transcript [36]. RtcB recognition and circularization is promoted via homology arms between the flanking Twister ribozymes and the region to be circularized, which form a stem after the autocatalytic cleavage events and places the ends that RtcB recognizes in close proximity [36]. Although this stem is designed to be relatively short to prevent activation of endogenous double-stranded RNA-detection pathways, the appended artificial hairpin at the back-splice junction ensures this method cannot entirely reproduce endogenous circRNA structure. However, the Tornado expression system has demonstrated its ability to allow proper RNA folding into functional structures for regions that do not span the back-splice junction [36].

6.3. Cellular systems

Circularization of RNA transcripts occurs spontaneously with very low frequency unless sequences flanking exons are positioned to promote back-splicing [1,7]. Therefore, circRNA expression vectors must be specially designed to shift transcript processing in favor of circRNA formation. The first circRNA expression vector included a cytomegalovirus promoter, an upstream flanking sequence, a single exon containing a 5′ splice acceptor site and a 3′ splice donor site to be circularized, a downstream flanking sequence that was inverted from the upstream flanking sequence, and a poly-adenylation signal (Fig. 2E) [67]. Expression vectors lacking mirrored flanking sequences express effectively no circRNA transcripts, suggesting that the formation of secondary structure is essential for circRNA expression [1,2,67].

A similar circRNA expression vector containing two exons to be circularized (Fig. 2E) was used to demonstrate that the stability of the secondary structure formed by the flanking sequences affects circRNA expression, though perhaps not in the way intuitively hypothesized: flanking sequences that formed substantially more stable structures than the wild-type sequences caused reduced circRNA expression and a shift in splicing products, perhaps due to splicing machinery hindrance and intron retention [1]. The intron between the two exons is spliced out with near-perfect efficiency in the wild-type vector, suggesting that a well-engineered vector effectively recapitulates formation of endogenous circRNA [1,2]. However, endogenous pre-mRNA may be alternatively spliced into a variety of circRNAs [2,12,66,80], which has not been recapitulated from circRNA expression plasmids.

Similar to these plasmid constructs, the adeno-associated viral (AAV) circRNA constructs have endogenous exons with flanking reverse-complementary intron sequences (Fig. 2E) and therefore are likely able to undergo typical splicing and produce circRNA with endogenous structures [68,81–84]. Current AAV circRNA vectors cannot recapitulate alternative circRNA splicing like from endogenous pre-mRNA [2,12,66,80], but recent work has demonstrated that a tRNA intron-derived circRNA can be incorporated within the intronic sequence of the AAV circRNA construct without affecting back-spliced circRNA formation, which enables co-regulated expression of both circular transcripts [81]. Although overexpression of a plasmid or introduction of an AAV vector can easily express exogenous circRNA, the constructs co-express linear transcripts with the same exon sequences at some basal amount, and these linear side products cannot be removed like for *in vitro* experiments (Section 6.1) [1,2,67].

Another cellular approach to circRNA expression uses the process of tRNA splicing to produce tRNA intron-derived circular RNAs (tricRNAs) [85]. An exogenous circRNA sequence is inserted in the intron between two tRNA exons. The exon–intron junctions are recognized and cleaved by the tRNA splicing endonuclease (TSEN), and the ubiquitous RNA ligase RtcB ligates the free ends of the introns to form a stable circRNA (Fig. 2F) [85,86]. tricRNA constructs can generate circRNAs of up to 800 nt but retain sequences from the termini of the tRNA introns at the back-splice junction after splicing [85,86].

Similar to *in vitro* circRNA expression, the size of exogenous circRNA in cellular constructs is limited by vector size, though AAV constructs can introduce circRNAs over 2 kb in length [84], far larger than the ~ 700 nt average endogenous circRNA [2,7]. It is prudent to include endogenous intronic regions surrounding the circRNA exon sequences, especially from multi-exon circRNA constructs [1,68,83]; this strategy limits the maximum circRNA size in a construct but encourages endogenous splicing and the formation of correct circRNA structure.

7. RNA structure in circRNA biogenesis, function, and decay

As the efficiency of circRNA generation and the accuracy of produced circRNA structures improve, the field of circRNA biology can better simulate and therefore experimentally examine the varied biological roles of circRNA. In this section, we will review the current understanding of the role of RNA structure in circRNA biogenesis, function, and decay and conclude by suggesting potential improvements in experimental designs when delineating their mechanisms.

7.1. Biogenesis

Other than the canonical splice sites, there are no sequence motifs

biologically misrepresentative structures [76], so it is recommended that a native polyacrylamide or an agarose gel without RNA denaturation is used for *in vitro*-transcribed circRNA purification.

To promote RNA circularization, the T7 expression vector may contain complementary flanking regions in the linear T7-transcribed transcript that are able to form a stem structure and bring the 5′ and 3′ ends in close proximity [75,77] (Fig. 2B). This protocol was originally intended to create microRNA sponges, which should not contain excessive structure at microRNA binding sites by design [75,77], but it may not be prudent to create circRNAs for which structure is important for function using this type of vector since the appended artificial hairpin at the back-splice junction may affect circRNA structure. Even with this circularization-promoting modification, circRNA size is limited to a few hundred nucleotides [75].

6.2. In vitro and in cells—Auto-catalytic constructs

The catalytic ability of group I introns from the *Anabaena* pre-tRNA has been harnessed to enable circularization of RNAs up to 4.8 kb and does not require any additional enzymes [45]. Permuted intron–exon splicing constructs consist of fused partial exons that contain the sequence to be circularized flanked by half-intron sequences (Fig. 2C) [78]. When the construct undergoes the group I intron double trans-esterification reaction, the fused exon is excised as a covalently linked circle [78]. The addition of homology arms that flank the half-intron sequences and bring the 5′ and 3′ splice sites into proximity increases splicing efficiency, especially for constructs longer than 300 nt [45]. A comparison of structure predictions for phage T4—another commonly-used group I intron for permuted intron–exon constructs [15,36,78,79]—and *Anabaena* constructs suggests that the latter construct forms secondary structures that aid in the formation of an isolated splicing bubble, which improves splicing efficiency; inclusion of additional spacers improves splicing efficiency even more [45]. The permuted intron–exon *Anabaena* group I splicing construct was originally intended for potent and stable translation in eukaryotic cells, so it is inadvisable to create circRNAs for which structure is important for function using this method since the appended spacers, internal homology arms, and standardized exon termini at the back-splice junction may affect the final circRNA structure [45].

The Tornado (Twister-optimized RNA for durable overexpression) expression system is another means of achieving rapid, high-efficiency RNA circularization without the co-expression of any additional enzymes [36]. Tornado-expressed transcripts contain a sequence to be circularized flanked by Twister ribozymes (Fig. 2D), which undergo autocatalytic cleavage and leave a hydroxyl and a 2′,3′-cyclic phosphate at the 5′ and 3′ ends, respectively. These ends are then recognized and ligated by the RNA ligase RtcB to form a circular transcript [36]. RtcB recognition and circularization is promoted via homology arms between the flanking Twister ribozymes and the region to be circularized, which form a stem after the autocatalytic cleavage events and places the ends that RtcB recognizes in close proximity [36]. Although this stem is designed to be relatively short to prevent activation of endogenous double-stranded RNA-detection pathways, the appended artificial hairpin at the back-splice junction ensures this method cannot entirely reproduce endogenous circRNA structure. However, the Tornado expression system has demonstrated its ability to allow proper RNA folding into functional structures for regions that do not span the back-splice junction [36].

6.3. Cellular systems

Circularization of RNA transcripts occurs spontaneously with very low frequency unless sequences flanking exons are positioned to promote back-splicing [1,7]. Therefore, circRNA expression vectors must be specially designed to shift transcript processing in favor of circRNA formation. The first circRNA expression vector included a cytomegalovirus promoter, an upstream flanking sequence, a single exon containing a 5′ splice acceptor site and a 3′ splice donor site to be circularized, a downstream flanking sequence that was inverted from the upstream flanking sequence, and a poly-adenylation signal (Fig. 2E) [67]. Expression vectors lacking mirrored flanking sequences express effectively no circRNA transcripts, suggesting that the formation of secondary structure is essential for circRNA expression [1,2,67].

A similar circRNA expression vector containing two exons to be circularized (Fig. 2E) was used to demonstrate that the stability of the secondary structure formed by the flanking sequences affects circRNA expression, though perhaps not in the way intuitively hypothesized: flanking sequences that formed substantially more stable structures than the wild-type sequences caused reduced circRNA expression and a shift in splicing products, perhaps due to splicing machinery hindrance and intron retention [1]. The intron between the two exons is spliced out with near-perfect efficiency in the wild-type vector, suggesting that a well-engineered vector effectively recapitulates formation of endogenous circRNA [1,2]. However, endogenous pre-mRNA may be alternatively spliced into a variety of circRNAs [2,12,66,80], which has not been recapitulated from circRNA expression plasmids.

Similar to these plasmid constructs, the adeno-associated viral (AAV) circRNA constructs have endogenous exons with flanking reverse-complementary intron sequences (Fig. 2E) and therefore are likely able to undergo typical splicing and produce circRNA with endogenous structures [68,81–84]. Current AAV circRNA vectors cannot recapitulate alternative circRNA splicing like from endogenous pre-mRNA [2,12,66,80], but recent work has demonstrated that a tRNA intron-derived circRNA can be incorporated within the intronic sequence of the AAV circRNA construct without affecting back-spliced circRNA formation, which enables co-regulated expression of both circular transcripts [81]. Although overexpression of a plasmid or introduction of an AAV vector can easily express exogenous circRNA, the constructs co-express linear transcripts with the same exon sequences at some basal amount, and these linear side products cannot be removed like for *in vitro* experiments (Section 6.1) [1,2,67].

Another cellular approach to circRNA expression uses the process of tRNA splicing to produce tRNA intron-derived circular RNAs (tricRNAs) [85]. An exogenous circRNA sequence is inserted in the intron between two tRNA exons. The exon–intron junctions are recognized and cleaved by the tRNA splicing endonuclease (TSEN), and the ubiquitous RNA ligase RtcB ligates the free ends of the introns to form a stable circRNA (Fig. 2F) [85,86]. tricRNA constructs can generate circRNAs of up to 800 nt but retain sequences from the termini of the tRNA introns at the back-splice junction after splicing [85,86].

Similar to *in vitro* circRNA expression, the size of exogenous circRNA in cellular constructs is limited by vector size, though AAV constructs can introduce circRNAs over 2 kb in length [84], far larger than the ~ 700 nt average endogenous circRNA [2,7]. It is prudent to include endogenous intronic regions surrounding the circRNA exon sequences, especially from multi-exon circRNA constructs [1,68,83]; this strategy limits the maximum circRNA size in a construct but encourages endogenous splicing and the formation of correct circRNA structure.

7. RNA structure in circRNA biogenesis, function, and decay

As the efficiency of circRNA generation and the accuracy of produced circRNA structures improve, the field of circRNA biology can better simulate and therefore experimentally examine the varied biological roles of circRNA. In this section, we will review the current understanding of the role of RNA structure in circRNA biogenesis, function, and decay and conclude by suggesting potential improvements in experimental designs when delineating their mechanisms.

7.1. Biogenesis

Other than the canonical splice sites, there are no sequence motifs

specifically required for circRNA back-splicing. Instead, the unprocessed transcript must adopt a structure that brings the splice donor and splice acceptor into close proximity to form the back-splice junction of a circRNA. Both *cis* and *trans* factors facilitate the formation of this unprocessed transcript structure (Fig. 3A).

7.1.1. Intron complementarity

Back-splicing often requires intronic regulatory elements flanking circularized exons. The formation of secondary structure between introns can promote back-splicing, since the base-pairing brings otherwise distal splice sites into close proximity [1,2,7]. Intronic complementary sequences, which facilitate the formation of inter-intronic secondary structure, are often derived from repetitive elements, such as *Alu* elements [7]. Although inverted repeat *Alu* and other repetitive elements are responsible for most circRNA formation in mammals [87], inter-intronic structures can also form from non-repetitive complementary sequences [1,2], and short sequences (about 30 to 40 nt) have been shown to be sufficient to induce circRNA formation from expression vectors [1] (Section 6.3).

7.1.2. RNA-binding protein-regulated structure

The expression levels of circRNAs from the same gene differ between cell lines and tissues [11,20,80], suggesting that *trans* factors also play a substantial role in circRNA regulation. Although many RNA-binding proteins likely play a role in the promotion or disruption of structure and subsequent circRNA biogenesis in the unprocessed transcript, here we have selected a few illustrative mechanisms by which proteins mediate RNA structure and influence circRNA biogenesis.

NF90 and NF110 both contain dsRNA binding domains and facilitate circRNA production by binding to inverted repeat *Alu* elements in introns flanking circRNA-forming exons [8]. It is likely that, upon binding to paired regions, these proteins stabilize otherwise transient circRNA-promoting secondary structures. QKI is also able to augment circRNA biogenesis, possibly via its ability to dimerize [9]. QKI can directly bind a single-stranded sequence motif in introns flanking circRNA-forming exons. Rather than promoting base-pairing between introns, QKI can form dimers, and it has been hypothesized that RNA looping resulting from this protein–protein interaction may act as an alternative to RNA secondary structure formation to bring distal splice sites into close proximity and enhance circRNA formation [9].

In contrast, the RNA binding proteins can inhibit circRNA formation by destabilizing inter-intronic pairing. ADAR1 ablates the complementarity and weakens the stability of base-pairing for introns flanking circRNA exons by A-to-I editing, shifting splicing to favor linear

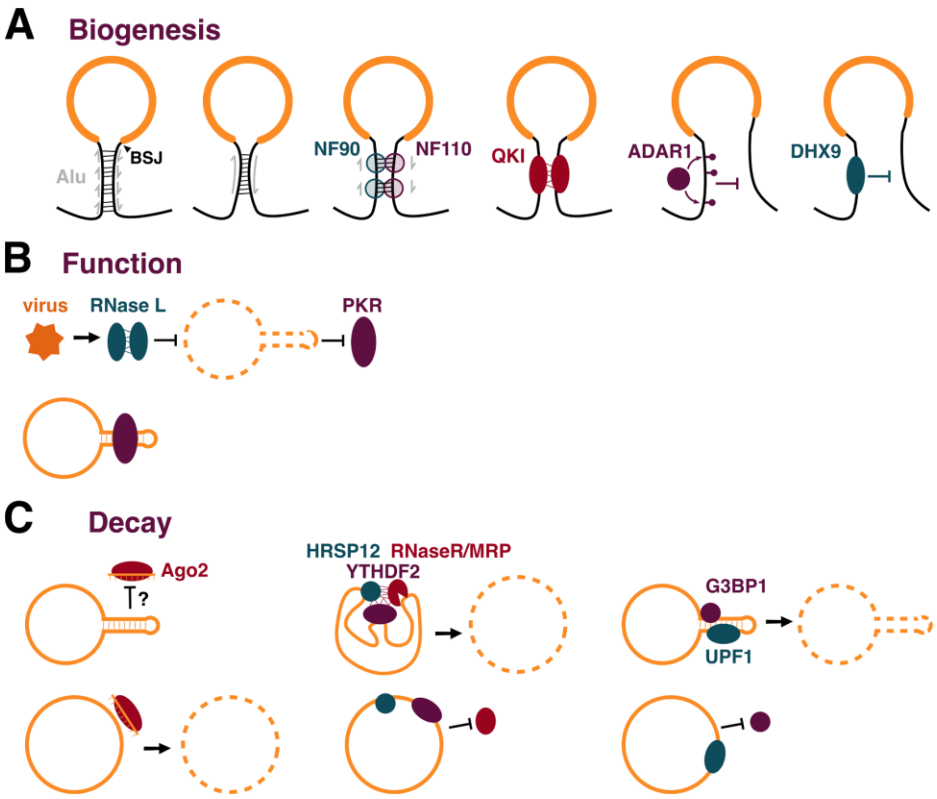


Fig. 3. CircRNA biogenesis, function, and decay affected by RNA structures. A. From left to right: Reverse-complement *Alu* repeats or intron fragments promote circRNA formation via base-pairing. Protein-protein interactions can also bring introns in proximity such as via proteins binding intronic *Alu* elements and forming protein complexes or by binding intronic sequences and dimerizing. Proteins can prevent flanking intron proximity and circRNA formation by modifying the intronic RNA to reduce base-pairing or by directly binding introns and preventing proximity of other introns. B. Structured circRNA inhibits PKR activity, but upon viral infection, RNase L dimerizes and degrades circRNA, freeing PKR. C. A comparison of RNA decay between structured (top) and unstructured (bottom) circRNAs. Upon binding to target RNAs in near-perfect complementarity, microRNAs can cleave targets in an Ago2-dependent manner. Based on precedents on linear RNA transcripts, the cleavage efficiency is affected by local RNA secondary structure, which may therefore affect where microRNA can access the binding region within circRNA structure. The RNaseP/MRP complex can degrade circRNAs able to form long-distance loops between HRSP12, YTHDF2, and RNaseP/MRP complex binding sites. Structured circRNAs bound by both G3BP1 and UPF1 undergo structure-mediated RNA decay.

transcript formation [11]. The RNA helicase DHX9 also inhibits circRNA biogenesis, though via a different hypothesized mechanism: DHX9 binds complementary intronic regions and may directly unwind any secondary structure formed between introns [10].

7.2. Function

Although few functions of circRNA structure have been definitively identified, this is likely a reflection of the degree to which the field of circRNA structure remains unexplored, and we anticipate this facet of circRNA discovery will expand in the near future.

7.2.1. PKR regulation

circRNAs are associated with innate immune regulation [8,15], but the specific mechanism by which they act remains to be clarified. Upon viral infection, RNase L, a ubiquitously-expressed cytoplasmic endoribonuclease, dimerizes and is activated, which catalyzes the degradation of viral and cellular RNAs to limit virus proliferation [88]. In contrast to a 10–30% reduction in linear RNAs, circRNA levels are reduced 80–90% in an RNase L-dependent manner following viral infection [13].

Although the RNase L decay mechanism seems to be agnostic to RNA structure, the tendency for circRNAs to form extensive imperfect RNA duplexes (16–26 bp) based on predictive and experimental analyses (see Section 3) is important for suppressing the activation of dsRNA-regulated kinase PKR. Although PKR is activated by long dsRNAs, short dsRNAs of 16–33 bp can bind PKR and block its activation [89,90] (Fig. 3B). PKR preferentially binds structured circRNAs rather than their linear cognates, and therefore structured circRNAs can act as endogenous inhibitors of PKR independent of linear RNA function. When PKR is needed to stage an antiviral response, RNase L globally decays circRNAs, freeing PKR from the structured circRNA inhibition.

7.3. Decay

circRNA decay acts as the counterpoint to biogenesis, ensuring that circRNAs are removed at the appropriate time during development or cellular response. The study of circRNA decay has recently yielded a proliferation of diverse mechanisms, many of which rely on circRNA structure (Fig. 3C).

7.3.1. microRNA-dependent Ago2 cleavage

microRNAs can direct Ago2-dependent cleavage of mRNAs via near-perfect complementary of microRNA target sites, usually in the 3′UTR of mRNAs [91,92]. It has also been demonstrated that circRNAs are subject to Ago2 endonucleolytic cleavage [7], but since only one example of circRNA regulation by an endogenous microRNA has been identified [30], this method of circRNA decay likely only drives a minor portion of circRNA regulation due to the limited number of near-perfect complementarity with microRNAs. Based on linear RNA precedents, it is possible that RNA secondary structure may limit the target site accessibility at circRNAs [27,93]. Therefore, circRNA structure should be considered when designing circRNA-targeting siRNAs or shRNAs, which take advantage of this Ago2 decay pathway.

7.3.2. m^6A-modified circRNA cleavage by RNaseP/MRP complex

m^6A is a reversible post-transcriptional modification that affects RNA folding and protein binding [37]. For example, m^6A-modified mRNA is recognized by the human YTH domain family 2 (YTHDF2) protein in the cytoplasm, and bound transcripts are selectively degraded [94]. Many circRNAs undergo m^6A modification [16], and these circRNAs can also be bound by YTHDF2 and degraded [35]. This process requires synchronization of three factors: 1) YTHDF2 binds the m^6A modification on the circRNA; 2) HRSP12 binds to a consensus sequence hundreds of nucleotides upstream of YTHDF2 binding (median = 862 nucleotides); and 3) the RNaseP/MRP complex cleaves the circRNA hundreds of nucleotides downstream of YTHDF2 binding (median = 421 nucleotides) in a sequence-independent manner [35]. The extensive linear distances between these three necessary decay components suggests that the circRNA undergoes extensive structural rearrangement to be subject to cleavage. Only three of eleven m^6A-positive circRNAs tested demonstrated increased expression after the downregulation of circRNA decay components (YTHDF2, HRSP12, or POP1, which is part of the RNaseP/MRP complex) [35]. Though the low proportion of m^6A-positive circRNAs subject to this decay mechanism may be because they lack requisite protein binding sites, it may also be because the circRNAs are unable to fold into the long-distance loops necessary to activate the decay mechanism. It is unknown whether the loops are formed by proteins in a complex drawing distal bound RNA together or by the RNA forming long-distance base-pairing interactions (*cf.* section 7.1 for similar proposed mechanisms of long-distance RNA looping).

7.3.3. Structure-mediated RNA decay

In addition to encoding information in its defined nucleotide sequences and structures, RNA can be regulated by its overall base-pairing density via structure-mediated RNA decay (SRD) [14]. Overall base-pairing density can be calculated by normalizing the change in free energy that results from folding an RNA transcript (see Section 3) by its nucleotide length. SRD requires two RNA-binding proteins, UPF1 and G3BP1, and depletion of either protein increases expression levels of highly-structured circRNAs as well as mRNAs with highly-structured 3′UTRs [14]. UPF1 binds to circRNAs irrespective of their structuredness, whereas G3BP1 preferentially binds highly-structured circRNAs [14]. UPF1 is a helicase that unwinds dsRNA regions [95,96], and G3BP1 has been implicated for endoribonuclease activity [97]. The currently proposed model for SRD is that the unwinding of dsRNA regions by UPF1 facilitates the enrichment of G3BP1 in proximity to UPF1, which would occur more frequently on highly-structured circRNAs, and that the proximity of these two RNA-binding proteins could be the signal required to trigger SRD [14]. SRD is predicted to regulate up to one-third of circRNAs in humans based on base-pairing density. Despite the broad regulatory potential for this pathway, there are no currently-identified biological roles of SRD, and additional factors required for this pathway remain elusive.

7.4. Concluding remarks

circRNA structure is central to its function and regulation, so better understanding and recapitulating it will improve experimental design and the translatability of circRNA research. For example, subtle differences in the methods used to generate circRNAs have resulted in drastically different experimental outcomes and conclusions when studying the role of circRNAs in innate immunity: circRNAs generated from T7 expression plasmids were shown to dampen innate immunity when generated via T4 RNA ligase [13], but triggered immunity when generated via T4 DNA ligase with a DNA splint [15]. Whether this functional discrepancy due to differences in structure at the back-splice junction between circRNAs formed by these methods or specific sequences warrants future investigations.

In addition to the many logistical factors that must be considered when selecting a method to detect and measure circRNA expression, it is important to weigh how well different methods are able to accommodate structured regions of circRNAs and discern linear from circRNA transcripts (Table 1). Similarly, when generating circRNAs with endogenous, modified, or novel sequences, it is important to consider whether the function of the generated circRNA may rely on its structure or, contrarily, whether its function may be negatively affected by structure and select a method accordingly (Table 2). For example, if studying a circRNA with structure spanning the back-splice junction that inhibits PKR, it would be best to generate that circRNA using a method that preserves the structure at the back-splice junction (Section 7.2.1). In contrast, a *de novo* designed circRNA that contains many stable structures may be targeted by SRD and have an unexpectedly short half-life

(Section 7.3.3).

Because structure has diverse ramifications for circRNA function and regulation, it is advisable to interrogate probable circRNA structure before experimenting on or generating any transcript. Computational predictions require only the circRNA sequence and can be generated within minutes, but experimental methods are required to generate high-confidence RNA structure predictions (Section 3). Preemptive circRNA structure predictions of exogenous or endogenous transcripts could save time troubleshooting experiments or explain discrepancies between the effects of endogenous and exogenous circRNAs.

Studying circRNA function and regulation through the lens of its structure using well-considered experimental methods adds a new facet that may help clarify and broaden the rapidly expanding field of circRNA biology.

Funding

This work was supported by Johns Hopkins Bloomberg School of Public Health Start-up fund; the Predoctoral Fellowship in Informatics from the Pharmaceutical Research and Manufacturers of America Foundation; and the National Institute of Health Grant T32GM07814.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] D. Liang, J.E. Wilusz, Short intronic repeat sequences facilitate circular RNA production, Genes Dev. 28 (20) (2014) 2233–2247, https://doi.org/10.1101/gad.251926.114.
[2] X.-O. Zhang, H.-B. Wang, Y. Zhang, X. Lu, L.-L. Chen, L.i. Yang, Complementary Sequence-Mediated Exon Circularization, Cell 159 (1) (2014) 134–147, https://doi.org/10.1016/j.cell.2014.09.001.
[3] S. Kelly, C. Greenman, P.R. Cook, A. Papantonis, Exon Skipping Is Correlated with Exon Circularization, J. Mol. Biol. 427 (15) (2015) 2414–2417, https://doi.org/10.1016/j.jmb.2015.02.018.
[4] H. Suzuki, Y. Zuo, J. Wang, M.Q. Zhang, A. Malhotra, A. Mayeda, Characterization of RNase R-digested cellular RNA source that consists of lariat and circular RNAs from pre-mRNA splicing, Nucleic Acids Res. 34 (2006), e63, https://doi.org/10.1093/nar/gkl151.
[5] M.T. Venø, T.B. Hansen, S.T. Venø, B.H. Clausen, M. Grebing, B. Finsen, I.E. Holm, J. Kjems, Spatio-temporal regulation of circular RNA expression during porcine embryonic brain development, Genome Biol. 16 (2015), https://doi.org/10.1186/s13059-015-0801-3.
[6] L. Szabo, R. Morey, N.J. Palpant, P.L. Wang, N. Afari, C. Jiang, M.M. Parast, C. E. Murry, L.C. Laurent, J. Salzman, Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development, Genome Biol. 16 (2015), https://doi.org/10.1186/s13059-015-0690-5.
[7] W.R. Jeck, J.A. Sorrentino, K. Wang, M.K. Slevin, C.E. Burd, J. Liu, W.F. Marzluff, N.E. Sharpless, Circular RNAs are abundant, conserved, and associated with ALU repeats, RNA 19 (2) (2013) 141–157, https://doi.org/10.1261/rna.035667.112.
[8] X. Li, C.-X. Liu, W. Xue, Y. Zhang, S. Jiang, Q.-F. Yin, J. Wei, R.-W. Yao, L.i. Yang, L.-L. Chen, Coordinated circRNA Biogenesis and Function with NF90/NF110 in Viral Infection, Mol. Cell 67 (2) (2017) 214–227.e7, https://doi.org/10.1016/j.molcel.2017.05.023.
[9] S. Conn, K. Pillman, J. Toubia, V. Conn, M. Salmanidis, C. Phillips, S. Roslan, A. Schreiber, P. Gregory, G. Goodall, The RNA Binding Protein Quaking Regulates Formation of circRNAs, Cell 160 (6) (2015) 1125–1134, https://doi.org/10.1016/j.cell.2015.02.014.
[10] T. Aktaş, İ. Avşar Ilık, D. Maticzka, V. Bhardwaj, C. Pessoa Rodrigues, G. Mittler, T. Manke, R. Backofen, A. Akhtar, DHX9 suppresses RNA processing defects originating from the Alu invasion of the human genome, Nature 544 (7648) (2017) 115–119, https://doi.org/10.1038/nature21715.
[11] A. Rybak-Wolf, C. Stottmeister, P. Glažar, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, M. Herzog, L. Schreyer, P. Papavasileiou, A. Ivanov, M. Öhman, D. Refojo, S. Kadener, N. Rajewsky, Circular RNAs in the Mammalian Brain Are Highly Abundant, Conserved, and Dynamically Expressed, Molecular Cell. 58 (5) (2015) 870–885, https://doi.org/10.1016/j.molcel.2015.03.027.
[12] T. Fei, Y. Chen, T. Xiao, W. Li, L. Cato, P. Zhang, M.B. Cotter, M. Bowden, R.T. Lis, S.G. Zhao, Q. Wu, F.Y. Feng, M. Loda, H.H. He, X.S. Liu, M. Brown, Genome-wide CRISPR screen identifies HNRNPL as a prostate cancer dependency regulating RNA splicing, Proc. Natl. Acad. Sci. U.S.A. 114 (2017) E5207–E5215, https://doi.org/10.1073/pnas.1617467114.
[13] C.-X. Liu, X. Li, F. Nan, S. Jiang, X. Gao, S.-K. Guo, W. Xue, Y. Cui, K. Dong, H. Ding, B.o. Qu, Z. Zhou, N. Shen, L.i. Yang, L.-L. Chen, Structure and Degradation of Circular RNAs Regulate PKR Activation in Innate Immunity, Cell 177 (4) (2019) 865–880.e21, https://doi.org/10.1016/j.cell.2019.03.046.
[14] J.W. Fischer, V.F. Busa, Y. Shao, A.K.L. Leung, Structure-Mediated RNA Decay by UPF1 and G3BP1, Mol. Cell (2020), https://doi.org/10.1016/j.molcel.2020.01.021.
[15] Y.G. Chen, M.V. Kim, X. Chen, P.J. Batista, S. Aoyama, J.E. Wilusz, A. Iwasaki, H. Y. Chang, Sensing Self and Foreign Circular RNAs by Intron Identity, Mol. Cell 67 (2) (2017) 228–238.e5, https://doi.org/10.1016/j.molcel.2017.05.022.
[16] C. Zhou, B. Molinie, K. Daneshvar, J.V. Pondick, J. Wang, N. Van Wittenberghe, Y. i. Xing, C.C. Giallourakis, A.C. Mullen, Genome-wide maps of m6A circRNAs identify widespread and cell-type-specific methylation patterns that are distinct from mRNAs, Cell Rep. 20 (9) (2017) 2262–2276, https://doi.org/10.1016/j.celrep.2017.08.027.
[17] Y. Yang, X. Fan, M. Mao, X. Song, P. Wu, Y. Zhang, Y. Jin, Y.i. Yang, L.-L. Chen, Y. Wang, C.CL. Wong, X. Xiao, Z. Wang, Extensive translation of circular RNAs driven by N6-methyladenosine, Cell Res. 27 (5) (2017) 626–641, https://doi.org/10.1038/cr.2017.31.
[18] I.L. Hofacker, P.F. Stadler, Memory efficient folding algorithms for circular RNA secondary structures, Bioinformatics 22 (10) (2006) 1172–1176, https://doi.org/10.1093/bioinformatics/btl023.
[19] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, I.L. Hofacker, ViennaRNA Package 2.0, Algorithms Mol. Biol. 6 (2011) 26, https://doi.org/10.1186/1748-7188-6-26.
[20] J. Salzman, C. Gawad, P.L. Wang, N. Lacayo, P.O. Brown, Circular RNAs Are the Predominant Transcript Isoform from Hundreds of Human Genes in Diverse Cell Types, PLoS ONE 7 (2012), e30733, https://doi.org/10.1371/journal.pone.0030733.
[21] D. Tsitsipatis, M. Gorospe, Practical guide for circular RNA analysis: Steps, tips, and resources, WIREs RNA. n/a (2020), e1633, https://doi.org/10.1002/wrna.1633.
[22] W.R. Jeck, N.E. Sharpless, Detecting and characterizing circular RNAs, Nat. Biotechnol. 32 (5) (2014) 453–461, https://doi.org/10.1038/nbt.2890.
[23] X. Li, L. Yang, L.-L. Chen, The Biogenesis, Functions, and Challenges of Circular RNAs, Mol. Cell. 71 (3) (2018) 428–442, https://doi.org/10.1016/j.molcel.2018.06.034.
[24] D.B.T. Cox, J.S. Gootenberg, O.O. Abudayyeh, B. Franklin, M.J. Kellner, J. Joung, F. Zhang, RNA Editing with CRISPR-Cas13, Science 358 (6366) (2017) 1019–1027, https://doi.org/10.1126/science:aaq0180.
[25] O.O. Abudayyeh, J.S. Gootenberg, P. Essletzbichler, S. Han, J. Joung, J.J. Belanto, V. Verdine, D.B.T. Cox, M.J. Kellner, A. Regev, E.S. Lander, D.F. Voytas, A.Y. Ting, F. Zhang, RNA targeting with CRISPR-Cas13a, Nature 550 (7675) (2017) 280–284, https://doi.org/10.1038/nature24049.
[26] S. Li, X. Li, W. Xue, L. Zhang, L.-Z. Yang, S.-M. Cao, Y.-N. Lei, C.-X. Liu, S.-K. Guo, L. Shan, M. Wu, X. Tao, J.-L. Zhang, X. Gao, J. Zhang, J. Wei, J. Li, L. Yang, L.-L. Chen, Screening for functional circular RNAs using the CRISPR–Cas13 system, Nat. Methods (2020) 1–9, https://doi.org/10.1038/s41592-020-01011-4.
[27] Z. Zheng, M. Reichel, I. Deveson, G. Wong, J. Li, A.A. Millar, Target RNA Secondary Structure Is a Major Determinant of miR159 Efficacy, Plant Physiol. 174 (3) (2017) 1764–1778, https://doi.org/10.1104/pp.16.01898.
[28] K. Felekkis, E. Touvana, C. Stefanou, C. Deltas, microRNAs: a newly described class of encoded molecules that play a role in health and disease, Hippokratia. 14 (2010) 236–240.
[29] A.C. Panda, I. Grammatikakis, K.M. Kim, S. De, J.L. Martindale, R. Munk, X. Yang, K. Abdelmohsen, M. Gorospe, Identification of senescence-associated circular RNAs (SAC-RNAs) reveals senescence suppressor CircPVT1, Nucleic Acids Res. 45 (2017) 4021–4035, https://doi.org/10.1093/nar/gkw1201.
[30] T.B. Hansen, E.D. Wiklund, J.B. Bramsen, S.B. Villadsen, A.L. Statham, S.J. Clark, J. Kjems, miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA, EMBO J. 30 (2011) 4414–4422, https://doi.org/10.1038/emboj.2011.359.
[31] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L.H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, N. Rajewsky, Circular RNAs are a large class of animal RNAs with regulatory potency, Nature 495 (7441) (2013) 333–338, https://doi.org/10.1038/nature11928.
[32] J. Jiao, Y. Xiang, C. Duan, Y. Liu, C. Li, G. Li, Lighting Up CircRNA Using a Linear DNA Nanostructure, Anal. Chem. 92 (18) (2020) 12394–12399, https://doi.org/10.1021/acs.analchem.0c02146.
[33] J.U. Guo, V. Agarwal, H. Guo, D.P. Bartel, Expanded identification and characterization of mammalian circular RNAs, Genome Biol. 15 (2014), https://doi.org/10.1186/s13059-014-0409-z.
[34] H. Le Hir, E. Izaurralde, L.E. Maquat, M.J. Moore, The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mRNA exon–exon junctions, EMBO J. 19 (2000) 6860–6869, https://doi.org/10.1093/emboj/19.24.6860.
[35] O.H. Park, H. Ha, Y. Lee, S.H. Boo, D.H. Kwon, H.K. Song, Y.K. Kim, Endoribonucleolytic Cleavage of m6A-Containing RNAs by RNase P/MRP Complex, Mol. Cell 74 (3) (2019) 494–507.e8, https://doi.org/10.1016/j.molcel.2019.02.034.
[36] J.L. Litke, S.R. Jaffrey, Highly efficient expression of circular RNA aptamers in cells using autocatalytic transcripts, Nat. Biotechnol. 37 (6) (2019) 667–675, https://doi.org/10.1038/s41587-019-0090-6.
[37] C. Roost, S.R. Lynch, P.J. Batista, K. Qu, H.Y. Chang, E.T. Kool, Structure and Thermodynamics of N6-Methyladenosine in RNA: A Spring-Loaded Base

Modification, J Am Chem Soc. 137 (2015) 2107–2115, https://doi.org/10.1021/ja513080v.
[38] N. Liu, K.I. Zhou, M. Parisien, Q. Dai, L. Diatchenko, T. Pan, N 6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein, Nucleic Acids Res. 45 (2017) 6051–6063, https://doi.org/10.1093/nar/gkx141.
[39] Y.G. Chen, R. Chen, S. Ahmad, R. Verma, S.P. Kasturi, L. Amaya, J.P. Broughton, J. Kim, C. Cadena, B. Pulendran, S. Hur, H.Y. Chang, N6-methyladenosine modification controls circular RNA immunity, Mol Cell. 76 (1) (2019) 96–109.e9, https://doi.org/10.1016/j.molcel.2019.07.016.
[40] M. Lei, G. Zheng, Q. Ning, J. Zheng, D. Dong, Translation and functional roles of circular RNAs in human cancer, Mol Cancer. 19 (2020) 30, https://doi.org/10.1186/s12943-020-1135-7.
[41] G. Di Timoteo, D. Dattilo, A. Centrón-Broco, A. Colantoni, M. Guarnacci, F. Rossi, D. Incarnato, S. Oliviero, A. Fatica, M. Morlando, I. Bozzoni, Modulation of circRNA Metabolism by m6A Modification, Cell Reports. 31 (2020), 107641, https://doi.org/10.1016/j.celrep.2020.107641.
[42] R. Perriman, M. Ares, Circular mRNA can direct translation of extremely long repeating-sequence proteins in vivo, RNA 4 (9) (1998) 1047–1054.
[43] Y. Yang, Z. Wang, IRES-mediated cap-independent translation, a path leading to hidden proteome, J. Mol Cell Biol. 11 (2019) 911–919. 10.1093/jmcb/mjz091.
[44] N.R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, S. Shenzis, M. Samson, G. Dittmar, M. Landthaler, M. Chekulaeva, N. Rajewsky, S. Kadener, Translation of CircRNAs, Mol. Cell 66 (1) (2017) 9–21.e7, https://doi.org/10.1016/j.molcel.2017.02.021.
[45] R.A. Wesselhoeft, P.S. Kowalski, D.G. Anderson, Engineering circular RNA for potent and stable translation in eukaryotic cells, Nat Commun. 9 (2018), https://doi.org/10.1038/s41467-018-05096-6.
[46] D.H. Mathews, D.H. Turner, Prediction of RNA secondary structure by free energy minimization, Curr. Opin. Struct. Biol. 16 (3) (2006) 270–278, https://doi.org/10.1016/j.sbi.2006.05.010.
[47] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (13) (2003) 3406–3415.
[48] S. Busan, C.A. Weidmann, A. Sengupta, K.M. Weeks, Guidelines for SHAPE Reagent Choice and Detection Strategy for RNA Structure Probing Studies, Biochemistry 58 (23) (2019) 2655–2664, https://doi.org/10.1021/acs.biochem.8b0121810.1021/acs.biochem.8b01218.s00110.1021/acs.biochem.8b01218.s002.
[49] X. Weng, J. Gong, Y. Chen, T. Wu, F. Wang, S. Yang, Y. Yuan, G. Luo, K. Chen, L. Hu, H. Ma, P. Wang, Q.C. Zhang, X. Zhou, C. He, Keth-seq for transcriptome-wide RNA structure mapping, Nat. Chem. Biol. (2020) 1–4, https://doi.org/10.1038/s41589-019-0459-3.
[50] C. Twittenhoff, V.B. Brandenburg, F. Righetti, A.M. Nuss, A. Mosig, P. Dersch, F. Narberhaus, Lead-seq: transcriptome-wide structure probing in vivo using lead (II) ions, Nucleic Acids Res. (2020), https://doi.org/10.1093/nar/gkaa404.
[51] S.A. Mortimer, K.M. Weeks, A Fast-Acting Reagent for Accurate Analysis of RNA Secondary and Tertiary Structure by SHAPE Chemistry, J. Am. Chem. Soc. 129 (14) (2007) 4144–4145, https://doi.org/10.1021/ja0704028.
[52] D.A. Peattie, W. Gilbert, Chemical probes for higher-order structure in RNA, Proc. Natl. Acad. Sci. U.S.A. 77 (8) (1980) 4679–4682.
[53] N.A. Siegfried, S. Busan, G.M. Rice, J.A.E. Nelson, K.M. Weeks, RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP), Nat Methods. 11 (9) (2014) 959–965, https://doi.org/10.1038/nmeth.3029.
[54] J.T. Low, K.M. Weeks, SHAPE-Directed RNA Secondary Structure Prediction, Methods 52 (2) (2010) 150–158, https://doi.org/10.1016/j.ymeth.2010.06.007.
[55] M.J. Smola, G.M. Rice, S. Busan, N.A. Siegfried, K.M. Weeks, Selective 2′-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile, and accurate RNA structure analysis, Nat Protoc. 10 (11) (2015) 1643–1669, https://doi.org/10.1038/nprot.2015.103.
[56] M.J. Smola, K.M. Weeks, In-cell RNA structure probing with SHAPE-MaP, Nat Protoc. 13 (6) (2018) 1181–1195, https://doi.org/10.1038/nprot.2018.010.
[57] A.C. Panda, S. De, I. Grammatikakis, R. Munk, X. Yang, Y. Piao, D.B. Dudekula, K. Abdelmohsen, M. Gorospe, High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs, Nucleic Acids Res. 45 (2017), e116, https://doi.org/10.1093/nar/gkx297.
[58] M.-S. Xiao, J.E. Wilusz, An improved method for circular RNA purification using RNase R that efficiently removes linear RNAs containing G-quadruplexes or structured 3′ ends, Nucleic Acids Res. 47 (2019) 8755–8769, https://doi.org/10.1093/nar/gkz576.
[59] T.B. Hansen, M.T. Venø, C.K. Damgaard, J. Kjems, Comparison of circular RNA prediction tools, Nucleic Acids Res. 44 (2016), e58, https://doi.org/10.1093/nar/gkv1458.
[60] L. Blazquez, W. Emmett, R. Faraway, J.M.B. Pineda, S. Bajew, A. Gohr, N. Haberman, C.R. Sibley, R.K. Bradley, M. Irimia, J. Ule, Exon Junction Complex Shapes the Transcriptome by Repressing Recursive Splicing, Mol Cell. 72 (3) (2018) 496–509.e9, https://doi.org/10.1016/j.molcel.2018.09.033.
[61] Y. Yue, G. Li, Y. Yang, W. Zhang, H. Pan, R. Chen, F. Shi, Y. Jin, Regulation of Dscam exon 17 alternative splicing by steric hindrance in combination with RNA secondary structures, RNA Biol. 10 (12) (2013) 1822–1833, https://doi.org/10.4161/rna.27176.
[62] J.N. Vo, M. Cieslik, Y. Zhang, S. Shukla, L. Xiao, Y. Zhang, Y.-M. Wu, S. M. Dhanasekaran, C.G. Engelke, X. Cao, D.R. Robinson, A.I. Nesvizhskii, A. M. Chinnaiyan, The Landscape of Circular RNA in Cancer, Cell 176 (4) (2019) 869–881.e13, https://doi.org/10.1016/j.cell.2018.12.021.
[63] C. Cocquerelle, B. Mascrez, D. Hétuin, B. Bailleul, Mis-splicing yields circular RNA molecules, FASEB J. 7 (1) (1993) 155–160, https://doi.org/10.1096/fasebj.7.1.7678559.
[64] Y. Gao, J. Wang, F. Zhao, CIRI: an efficient and unbiased algorithm for de novo circular RNA identification, Genome Biol. 16 (2015), https://doi.org/10.1186/s13059-014-0571-3.
[65] R. Xin, Y. Gao, Y. Gao, R. Wang, K.E. Kadash-Edmondson, B. Liu, Y. Wang, L. Lin, Y. Xing, isoCirc catalogs full-length circular RNA isoforms in human transcriptomes, Nat. Commun. 12 (2021), https://doi.org/10.1038/s41467-020-20459-8.
[66] Y. Gao, J. Wang, Y. Zheng, J. Zhang, S. Chen, F. Zhao, Comprehensive identification of internal structure and alternative splicing events in circular RNAs, Nat. Commun. 7 (2016) 12060, https://doi.org/10.1038/ncomms12060.
[67] T.B. Hansen, T.I. Jensen, B.H. Clausen, J.B. Bramsen, B. Finsen, C.K. Damgaard, Jørgen Kjems, Natural RNA circles function as efficient microRNA sponges, Nature 495 (7441) (2013) 384–388, https://doi.org/10.1038/nature11993.
[68] K. Wang, B. Long, F. Liu, J.-X. Wang, C.-Y. Liu, B. Zhao, L.-Y. Zhou, T. Sun, M. Wang, T. Yu, Y. Gong, J. Liu, Y.-H. Dong, N. Li, P.-F. Li, A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223, Eur. Heart J. 37 (33) (2016) 2602–2611, https://doi.org/10.1093/eurheartj/ehv713.
[69] V. Conn, S.J. Conn, SplintQuant: a method for accurately quantifying circular RNA transcript abundance without reverse transcription bias, RNA 25 (2019) 1202–1210, https://doi.org/10.1261/rna.070953.119.
[70] M. Dahl, I. Daugaard, M.S. Andersen, T.B. Hansen, K. Grønbæk, Jørgen Kjems, L. S. Kristensen, Enzyme-free digital counting of endogenous circular RNA molecules in B-cell malignancies, Lab Invest. 98 (12) (2018) 1657–1669, https://doi.org/10.1038/s41374-018-0108-6.
[71] U. Ahmadov, M.M. Bendikas, K.K. Ebbesen, A.M. Sehested, J. Kjems, H. Broholm, L.S. Kristensen, Circular RNA expression profiles in pediatric ependymomas, MedRxiv. (2020) 2020.08.04.20167312. 10.1101/2020.08.04.20167312.
[72] T. Li, Y. Shao, L. Fu, Y. Xie, L. Zhu, W. Sun, R. Yu, B. Xiao, J. Guo, Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection, J. Mol. Med. 96 (1) (2018) 85–96, https://doi.org/10.1007/s00109-017-1600-y.
[73] L.-L. Chen, The expanding regulatory mechanisms and cellular functions of circular RNAs, Nat. Rev. Mol. Cell Biol. (2020) 1–16, https://doi.org/10.1038/s41580-020-0243-y.
[74] S. Petkovic, S. Müller, RNA circularization strategies in vivo and in vitro, Nucleic Acids Res. 43 (2015) 2454–2465, https://doi.org/10.1093/nar/gkv045.
[75] J. Breuer, O. Rossbach, Production and Purification of Artificial Circular RNA Sponges for Application in Molecular Biology and Medicine, Methods Protoc. 3 (2020), https://doi.org/10.3390/mps3020042.
[76] R. Russell, RNA misfolding and the action of chaperones, Front Biosci. 13 (13) (2008) 1, https://doi.org/10.2741/2557.
[77] I. Jost, L.A. Shalamova, G.K. Gerresheim, M. Niepmann, A. Bindereif, O. Rossbach, Functional sequestration of microRNA-122 from Hepatitis C Virus by circular RNA sponges, RNA Biol. 15 (2018) 1032–1039, https://doi.org/10.1080/15476286.2018.1435248.
[78] M. Puttaraju, M. Been, Group I permuted intron-exon (PIE) sequences self-splice to produce circular exons, Nucleic Acids Res. 20 (20) (1992) 5357–5364.
[79] E. Ford, M. Ares, Synthesis of circular RNA in bacteria and yeast using RNA cyclase ribozymes derived from a group I intron of phage T4, PNAS 91 (8) (1994) 3117–3121, https://doi.org/10.1073/pnas.91.8.3117.
[80] X.-O. Zhang, R. Dong, Y. Zhang, J.-L. Zhang, Z. Luo, J. Zhang, L.-L. Chen, L. Yang, Diverse alternative back-splicing and alternative splicing landscape of circular RNAs, Genome Res. 26 (9) (2016) 1277–1287, https://doi.org/10.1101/gr.202895.115.
[81] R.M. Meganck, E.K. Borchardt, R.M. Castellanos Rivera, M.L. Scalabrino, J. E. Wilusz, W.F. Marzluff, A. Asokan, Tissue-Dependent Expression and Translation of Circular RNAs with Recombinant AAV Vectors In Vivo, Mol. Ther Nucleic Acids. 13 (2018) 89–98, https://doi.org/10.1016/j.omtn.2018.08.008.
[82] V.N.S. Garikipati, S.K. Verma, Z. Cheng, D. Liang, M.M. Truongcao, M. Cimini, Y. Yue, G. Huang, C. Wang, C. Benedict, Y. Tang, V. Mallaredy, J. Ibetti, L. Grisanti, S.M. Schumacher, E. Gao, S. Rajan, J.E. Wilusz, D. Goukassian, S.R. Houser, W. J. Koch, R. Kishore, Circular RNA CircFndc3b modulates cardiac repair after myocardial infarction via FUS/VEGF-A axis, Nat Commun. 10 (2019), https://doi.org/10.1038/s41467-019-11777-7.
[83] A. Lavenniah, T.D.A. Luu, Y.P. Li, T.B. Lim, J. Jiang, M. Ackers-Johnson, R.-Y. Foo, Engineered Circular RNA Sponges Act as miRNA Inhibitors to Attenuate Pressure Overload-Induced Cardiac Hypertrophy, Mol. Ther. 28 (6) (2020) 1506–1517, https://doi.org/10.1016/j.ymthe.2020.04.006.
[84] R.M. Meganck, J. Liu, A.E. Hale, K.E. Simon, M.M. Fanous, H.A. Vincent, J. E. Wilusz, N.J. Moorman, W.F. Marzluff, A. Asokan, Engineering highly efficient backsplicing and translation of synthetic circRNAs, Mol. Ther. Nucleic Acids (2021), https://doi.org/10.1016/j.omtn.2021.01.003.
[85] Z. Lu, G.S. Filonov, J.J. Noto, C.A. Schmidt, T.L. Hatkevich, Y. Wen, S.R. Jaffrey, A. G. Matera, Metazoan tRNA introns generate stable circular RNAs in vivo, RNA 21 (9) (2015) 1554–1565, https://doi.org/10.1261/rna.052944.115.
[86] J.J. Noto, C.A. Schmidt, A.G. Matera, Engineering and expressing circular RNAs via tRNA splicing, RNA Biol. 14 (8) (2017) 978–984, https://doi.org/10.1080/15476286.2017.1317911.
[87] R. Dong, X.-K. Ma, L.-L. Chen, L. Yang, Increased complexity of circRNA expression during species evolution, RNA Biol. 14 (8) (2017) 1064–1074, https://doi.org/10.1080/15476286.2016.1269999.
[88] H. Kristiansen, H.H. Gad, S. Eskildsen-Larsen, P. Despres, R. Hartmann, The Oligoadenylate Synthetase Family: An Ancient Protein Family with Multiple Antiviral Activities, J. Interferon Cytokine Res. 31 (2010) 41–47, https://doi.org/10.1089/jir.2010.0107.

[89] S.R. Nallagatla, R. Toroney, P.C. Bevilacqua, RNA structure and regulation of innate immunity through protein kinase PKR, Curr. Opin Struct. Biol. 21 (2011) 119–127, https://doi.org/10.1016/j.sbi.2010.11.003.
[90] X. Zheng, P.C. Bevilacqua, Activation of the protein kinase PKR by short double-stranded RNAs with single-stranded tails, RNA 10 (2004) 1934–1945, https://doi.org/10.1261/rna.7150804.
[91] G. Hutvágner, P.D. Zamore, A microRNA in a Multiple-Turnover RNAi Enzyme Complex, Science 297 (2002) 2056–2060, https://doi.org/10.1126/science.1073827.
[92] S. Yekta, I. -hung Shih, D.P. Bartel, MicroRNA-Directed Cleavage of HOXB8 mRNA, Science. 304 (2004) 594–596. 10.1126/science.1097434.
[93] Q. Fei, P. Li, C. Teng, B.C. Meyers, Secondary siRNAs from Medicago NB-LRRs modulated via miRNA–target interactions and their abundances, Plant J. 83 (3) (2015) 451–465, https://doi.org/10.1111/tpj.2015.83.issue-310.1111/tpj.12900.
[94] X. Wang, Z. Lu, A. Gomez, G.C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, C. He, m6A-dependent regulation of messenger RNA stability, Nature 505 (2014) 117–120, https://doi.org/10.1038/nature12730.
[95] S.R. Lee, G. Pratt, F. Martinez, G.W. Yeo, J. Lykke-Andersen, Target discrimination in nonsense-mediated mRNA decay requires Upf1 ATPase activity, Mol Cell. 59 (2015) 413–425. 10.1016/j.molcel.2015.06.036.
[96] F. Fiorini, D. Bagchi, H. Le Hir, V. Croquette, Human Upf1 is a highly processive RNA helicase and translocase with RNP remodelling activities, Nat Commun. 6 (2015), https://doi.org/10.1038/ncomms8581.
[97] H. Tourrière, I. Gallouzi, K. Chebli, J.P. Capony, J. Mouaikel, P. van der Geer, J. Tazi, RasGAP-Associated Endoribonuclease G3BP: Selective RNA Degradation and Phosphorylation-Dependent Localization, Mol Cell Biol. 21 (2001) 7747–7760, https://doi.org/10.1128/MCB.21.22.7747-7760.2001.