

Computational analysis of the post-transcriptional gene regulatory network

by

Mallory Ann Freeberg

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2015

Doctoral Committee:

Associate Professor John K. Kim, Committee Chair
Assistant Professor James D. Cavalcoli
Assistant Professor Aaron C. Goldstrohm
Assistant Professor Jeffrey M. Kidd
Assistant Professor Maureen A. Sartor

“Never do less than your best.”

© Mallory Ann Freeberg 2015
All Rights Reserved

This thesis is dedicated to
my grandfather, Joshua J. Freeberg,
and my grandmother, Ann C. Laub.
I hope I have made you both proud.

ACKNOWLEDGEMENTS

My graduate research experience has been profoundly and forever shaped by the guidance of my mentor, John Kim. His impassioned scientific curiosity, dogged work ethic, and dedication to excellence have inspired me to never stop reaching for my highest potential. I am also greatly appreciative of the support and guidance of my dissertation committee members, James Cavalcoli, Aaron Goldstrohm, Jeff Kidd, and Maureen Sartor.

I am immensely indebted to my close colleagues Allison Billi and Ting Han. Much of my work has directly benefitted from their expertise and critical thinking for which I cannot thank them enough. I am also grateful to my lab mates - Amelia Alessi, Charlotte Choi, Tony Chun, Amanda Day, Vishal Khivansara, Andy Kong, Arun Manoharan, Natasha Weiser, and Danny Yang - for tolerating my scientific musings and proving to be excellent comrades for Red Hawk happy hours. I would like to thank all of my collaborators, of whom there are too many to name, for allowing me access to their data and the freedom to explore it.

I owe a huge debt of gratitude to the Life Sciences Institute's IT department, especially Brad Battey, RJ Connell, Chris Lilienthal, Betty Narvaez-Rodriguez, and Levi Zander. I would like to officially apologize for crashing the server on many occasions, updating to OS X Mavericks, and generally causing/having a lot of "well-I've-never-seen-that-before" issues.

To my Bioinformatics cohort, my dearest friends, Shanshan Cheng, Jinyi Li-Stevenson (honorary member), Ellen Schmidt, Kraig Stevenson, and Artur Veloso: I cannot begin to describe how blessed I am for having you in my life for the last six years. I cherish every memory, both in and out of lab, and thank you all for making my graduate school experience

so wonderful. I am especially grateful to Kraig and Artur for invaluable LaTeX guidance, to Jinyi for sisterly conversations at The Melting Pot, to Shanshan for always encouraging me (add oil!), and to Ellen for keeping me up-to-date on T. Swift. I am grateful for enlightening discussions with my fellow Bioinformatics colleagues Craig Biver, Tony Chun, Matthew Iyer, Andy Kong, and Avinash Shanmugam, as well as with my friends Jeremiah Baron, Brittany Bowman, Alex van Esbroeck, and Chase Weidmann.

Finally, and perhaps most importantly, I am extremely grateful to my amazingly supportive family. To my father, Kevin: thank you for always believing in me and telling me that I am capable of achieving anything. To my mother, BethAnn: thank you for being a patient listener and offering wise advice. To my younger brothers, Christian and Benjamin, and my younger sister, Caitlin: thank you for inspiring me every day with your creativity, your independence, and your love (although I am still jealous that you all finished school before me). To my step-mother, Cheri: thank you for providing a welcome home when I come to visit and sending me the latest viral dachshund videos. To my extended family, both near and far, thank you.

To Tomàs, mi novio: there are no words to express how grateful I am for your love and support during the final stages of my doctoral career. Thank you so much for putting up with, among many other things, the long hours, missing your soccer games, and complaining about the lack of La Croix and Diet Coke in the fridge.

To Grammy: thank you for teaching me patience in the form of knitting, and love in the form of giving to others. To Pap: thank you for teaching me how to be strong, that integrity is everything, and that sometimes you just need to smell the flowers and watch the birds.

PREFACE

The work presented in this dissertation addresses a range of post-transcriptional gene regulation (PTGR) mechanisms in multiple systems. I have included a broad introduction, **Chapter I**, that describes mechanisms and components of PTGR, deep sequencing methods to investigate PTGR, and computational approaches for analyzing sequencing datasets.

Chapter II of this dissertation describes a unique signature of *C. elegans* piRNA expression that distinguishes sub-populations of piRNAs specific to regulating gene expression in the male and female germlines. This work, which includes molecular biology and biochemical experiments by Allison Billi, was published in *PLoS Genetics* in 2013.

Chapter III of this dissertation characterizes transcriptome-wide binding patterns of RNA-binding proteins (RBPs) in budding yeast. This work, which includes molecular biology experiments by Ting Han, was published in *Genome Biology* in 2013. The computational methods developed for this work are in press at *Methods in Molecular Biology*.

Chapter IV of this dissertation describes insights gained into mechanisms of PTGR governed by the conserved Pumilio family of RBPs in budding yeast and represents a close collaboration with Ting Han who generated the sequencing libraries (manuscript submitted).

Chapter V of this dissertation describes collaborative work with Danny Yang to characterize PTGR by PUF-9 and miRNAs in nematodes (manuscript in preparation) and collaborative work with Sung Ki Hong in the lab of Ken Inoki to determine the role of LARP1-mediated PTGR in human cell lines (manuscript in revision). In both studies Ting Han generated the sequencing libraries.

A perspective on future explorations of PTGR mechanisms in eukaryotes, with an emphasis on the importance of RBPs, is included as **Chapter VI**.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
PREFACE	v
LIST OF FIGURES	x
LIST OF TABLES	xiv
LIST OF ABBREVIATIONS	xv
CHAPTER	
I. Introduction to post-transcriptional gene regulation	1
1.1 PTGR governs gene expression in eukaryotes	1
1.2 Proteins and RNAs orchestrate PTGR	2
1.3 Small RNAs and RBPs regulate gene expression networks	8
1.4 Deep sequencing informs mechanisms of PTGR	13
1.5 Deep sequencing data necessitate sophisticated computational methods	18
1.6 Investigating PTGR poses bioinformatic challenges	21
II. Small RNA-mediated PTGR investigated by small RNA-seq	27
2.1 Introduction	27
2.2 A conserved upstream motif orchestrates autonomous, germline-enriched expression of <i>Caenorhabditis elegans</i> piRNAs	30
2.3 Materials and methods	42
2.4 Discussion and concluding remarks	50
2.4.1 Discussion of findings	50
2.4.2 Transcriptional regulation of <i>C. elegans</i> 21U RNAs	55
2.4.3 Investigating PTGR using small RNA-seq	57
2.4.4 Bioinformatics challenges	58
2.5 Data availability	59
2.6 Acknowledgements	60
III. Global characteristics of PTGR investigated by gPAR-CLIP-seq	82

3.1	Introduction	82
3.2	Pervasive and dynamic protein binding sites of the mRNA transcriptome in <i>Saccharomyces cerevisiae</i>	85
3.3	Materials and methods	96
	3.3.1 Specific to this study	96
	3.3.2 Generalizable to any system	110
3.4	Discussion and concluding remarks	122
	3.4.1 Discussion of findings	122
	3.4.2 Investigating PTGR using gPAR-CLIP-seq	125
	3.4.3 Bioinformatic challenges	126
3.5	Data availability	128
3.6	Acknowledgements	128
 IV. Mechanisms of PTGR by Pumilio family RBPs investigated by PAR-CLIP-seq		159
4.1	Introduction	159
4.2	Transcriptome-wide identification and target analysis of the Pumilio family of RNA binding proteins	164
4.3	Materials and methods	182
4.4	Discussion and concluding remarks	186
4.5	Data availability	190
4.6	Acknowledgements	190
 V. Specific RBP-directed PTGR investigated by PAR-/CLIP-seq		217
5.1	PUF-9 regulates developmental timing with miRNAs in <i>C. elegans</i>	218
	5.1.1 Introduction	218
	5.1.2 PUF-9 and miRISC directly interact at sequence-specific and highly structured regions of co-targeted genes	219
5.2	LARP1 displays dual mode of RNA binding to regulate translation of ribosomal protein-encoding genes	224
	5.2.1 Introduction	224
	5.2.2 LARP1 binds 5'UTR pyrimidine-rich regions of translation-related transcripts in a mTOR activity-dependent manner	225
5.3	Materials and methods	230
	5.3.1 PUF-9 HITS-CLIP and data analysis	230
	5.3.2 LARP1 PAR-CLIP-seq and data analysis	236
5.4	Discussion and concluding remarks	239
	5.4.1 PTGR by PUF-9/miRISC is physically linked on 3'UTRs	239
	5.4.2 LARP1 is a model for multiple RBDs promoting multiple PTGR mechanisms	240
5.5	Data availability	241

5.6	Acknowledgements	241
VI.	Perspectives on future investigations of PTGR	252
6.1	Exploring the <i>S. cerevisiae</i> RBPome	252
6.2	RNA binding, glycolytic enzymes, and metabolic flux	257
6.3	Characterizing tissue-specific, multi-system RBP:RNA interactomes .	259
6.4	Bioinformatic considerations for RBP-related sequencing data	261
REFERENCES	270

LIST OF FIGURES

Figure

1.1	Pipeline for computational analysis of small RNA-seq datasets	25
1.2	Pipeline for computational analysis of CLIP-seq datasets	26
2.1	Over 70% of 21U RNAs show distinct germline enrichment	64
2.2	Variation in the core upstream motif correlates with 21U RNA germline enrichment	65
2.3	A transgenic synthetic 21U RNA shows characteristics of endogenous 21U RNAs	67
2.4	A 5' cytidine in the core upstream motif promotes male germline expression pattern of 21UR-synth	68
2.5	21U RNA sequences are specified by the genomic positions of upstream core motifs	69
2.6	21U RNAs represent independent transcriptional units	70
S2.1	Computational identification of male and female germline-enriched 21U RNAs	71
S2.2	Female 21U RNAs are preferentially abundant in embryo	73
S2.3	21U RNA-dependent 22G RNAs	74
S2.4	Transgenic array expression varies across transgenes	75
S2.5	21U RNAs are specifically immunoprecipitated with PRG-1 complexes . . .	76
S2.6	21U RNA expression requires a 5' genomic thymidine	77
S2.7	RNA polymerase II occupancy at 21U RNA loci is below background level	79
S2.8	Model of 21U RNA expression	80

S2.9	Modeling 21U RNA control counts with negative binomial distribution . . .	81
3.1	gPAR-CLIP identifies transcriptome-wide RBP crosslinking sites	130
3.2	gPAR-CLIP captures known RBP crosslinking signatures	131
3.3	RBP crosslinking sites exhibit global sequence conservation	132
3.4	RBP crosslinking sites share global structural characteristics	133
3.5	Nutrient deprivation induces widespread but distinct RBP-crosslinking and mRNA changes	135
3.6	Glucose starvation induces RBP-crosslinking and mRNA changes associated with mitochondrial processes	136
3.7	Nitrogen starvation induces specific RBP-crosslinking and mRNA changes associated with ribosomes and translation-related processes	137
3.8	Pipeline for analysis of gPAR-CLIP-seq data	139
3.9	Pseudocode describing how to generate read clusters from mapped gPAR-CLIP-seq data	140
3.10	Distinct binding events identified within long read clusters	141
3.11	Pseudocode describing how to identify distinct binding events in long read clusters	142
3.12	Total versus T-to-C read coverage for determining gPAR-CLIP-seq binding site FDRs	143
S3.1	Sequencing and mapping statistics	144
S3.2	Pipeline for generating crosslinking scores and crosslinking sites	145
S3.3	Computational identification of crosslinking sites	146
S3.4	Visualization of crosslinking site periodicity	147
S3.5	Analysis and comparison of PAR-CLIP-identified Puf3p targets	148

S3.6	Analysis of crosslinking scores and conservation of genomic Ts in starvation conditions	150
S3.7	Analysis of RNA secondary structure in starvation conditions	151
S3.8	Intra-replicate variation of crosslinking site coverage and global changes in 5'UTR crosslinking sites	153
S3.9	Assessment of crosslinking site and mRNA changes in starvation conditions	154
S3.10	Global changes in 3'UTR crosslinking sites upon glucose starvation	155
S3.11	Changes in 3'UTR crosslinking sites on <i>ALD4</i> and <i>STM1</i> upon glucose starvation	156
S3.12	Global changes in 3'UTR crosslinking sites upon nitrogen starvation	157
S3.13	Changes in 3'UTR crosslinking sites on <i>INO1</i> and <i>AGP3</i> upon glucose starvation	158
4.1	Target binding sites of Puf proteins identified by PAR-CLIP-seq	192
4.2	Puf1/2p binding motifs are correlated with Puf domain architecture	194
4.3	Puf3p displays two distinct modes of target binding	196
4.4	Puf3p binding modes correlate with different biological functions	197
4.5	Puf4p binding characteristics	198
4.6	Puf5p binding characteristics	199
4.7	Puf6p binds snoRNAs and rRNAs in snoRNP complexes	200
4.8	Co-targeting by Pufs is extensive	201
S4.1	Target binding sites of Puf proteins identified by PAR-CLIP-seq	203
S4.2	Puf1/2p binding motifs are correlated with their Puf domain architecture .	205
S4.3	Puf binding sites do not respond to nitrogen starvation.	207
S4.4	Puf3p displays two distinct modes of target binding	209

S4.5	Puf3p binding sites respond uniquely to nutrient starvation	210
S4.6	Gene expression changes in response to nutrient starvation	212
S4.7	Puf4p sites harbor short structural motif	213
S4.8	mRNA stability measurements of Puf5p targets	214
S4.9	Puf6p binds snoRNAs and rRNAs	215
S4.10	Overlap of genes targeted by each Puf at different genic regions	216
5.1	PUF-9 and ALG-1 co-target similar gene sets at directly overlapping target sites.	243
5.2	PUF-9 targets highly structured RNAs.	244
5.3	LARP1 binds pyrimidine-rich 5'UTR regions of translation-related mRNAs	246
S5.1	Pumilio recognition elements are important for PUF-9 binding to targets. .	247
S5.2	PUF-9 targets highly structured RNAs.	248
S5.3	LARP1 function conserved in mouse	251
6.1	Identification of yeast mRBPome	267
6.2	Novel yeast mRBP characteristics	268
6.3	Pfk2p binds mRNAs encoding glycolytic enzymes	269

LIST OF TABLES

Table

1.1	Classical RNA-binding domains play major roles in disease pathologies. . .	24
S2.1	Descriptions of small RNA sequencing libraries used in this study	61
S2.2	<i>p</i> -values for all abundance comparisons between 21U RNAs with different core motifs	62
S2.3	Descriptions of transgenic alleles and features of the transgenes	63
6.1	Examples of novel yeast RBPs and their top mRNA targets by PAR-CLIP-seq.	266

LIST OF ABBREVIATIONS

4sU 4-thiouracil

CDS coding region

CLIP crosslinking and immunopurification (immunoprecipitation)

CLIP-seq CLIP followed by deep sequencing

CLS crosslinking score

endo-siRNA endogenous siRNA

FDR false discovery rate

GO gene ontology

gPAR-CLIP-seq global PAR-CLIP-seq

HITS-CLIP high-throughput sequencing following CLIP

ID intrinsically disordered

LARP1 La-related protein

LC low complexity

m⁶A N⁶-Methyladenosine

miRISC microRNA-loaded RISC

miRNA microRNA

mRBP mRNA-binding protein

mRNA messenger RNA

PAR-CL-MS photoactivatable ribonucleoside-enhanced crosslinking, oligo(dT) affinity purification, and mass spectrometry

PAR-CLIP-seq photoactivatable ribonucleoside-enhanced CLIP-seq

PCR polymerase chain reaction

piRNA Piwi-interacting RNA

PTGR post-transcriptional gene regulation

PUF Pumilio/FBF

RBD RNA-binding domain

RBP RNA-binding protein

RdRP RNA-directed RNA polymerase

RISC RNA-induced silencing complex

RNAi RNA interference

RNP ribonucleoprotein

RPKM RPM per kilobase of transcript

RPM reads per million mapped reads

SD standard deviation

SEM standard error of the mean

siRNA short (small) interfering RNA

snoRNA small nucleolar RNA

snRNA small nuclear RNA

UTR untranslated region

WT wild-type

CHAPTER I

Introduction to post-transcriptional gene regulation

1.1 PTGR governs gene expression in eukaryotes

Generating a protein product from information encoded in a genome constitutes a complex and tightly regulated process and can be modulated in response to intercellular and extracellular cues. For example, gene expression is regulated at the level of transcription by the activity of transcription factors (Chen and Rajewsky, 2007; Levine and Tjian, 2003) and by the remodeling of chromatin (Huisinga et al., 2006). At the translational level, gene expression is modulated by altering ribosomal recruitment to initiate translation (Sonenberg and Hinnebusch, 2009) and affecting the rate of translation elongation through mRNA secondary structures and transfer RNA availability (Gingold and Pilpel, 2011). Post-translational modifications also regulate gene expression through reversible chemical modifications of amino acids (*e.g.* phosphorylation, lysine acetylation, and ubiquitination (Venne et al., 2014)) or non-reversible proteolytic peptide cleavage (Barret et al., 1998).

In addition to these processes, gene expression is intimately regulated at steps between transcription and translation, referred to as **post-transcriptional gene regulation** (PTGR), to control precursor mRNA (pre-mRNA) processing, mRNA export and localization, and mRNA stability and storage. While the paradigm of the Central Dogma of Molecular Biology - that genetic information is transferred from DNA to RNA to protein - is well established, detailed mechanistic insights into the role of post-transcriptional reg-

ulation remains an active area of investigation. This dissertation aims to elucidate aspects and mechanisms of post-transcriptional regulation of gene expression through the analysis of deep sequencing datasets that probe specific and global post-transcriptional regulatory pathways including the role of small noncoding RNAs and RNA-binding proteins.

1.2 Proteins and RNAs orchestrate PTGR

PTGR in eukaryotes is largely directed by a complex network of interactions between a rapidly expanding repertoire of **RNA-binding proteins** (RBPs), highly conserved classes of **small non-coding RNAs**, and the substrate mRNAs being regulated. In order to elicit their regulatory functions, RBPs selectively recognize and interact with their target mRNAs, often forming higher-order ribonucleoprotein (RNP) complexes to co-localize the cellular components required for their function (Mitchell and Parker, 2014). The importance of RBPs and RNP formation in PTGR cannot be understated: RBPs represent a significant portion of any one organism’s proteome. The *C. elegans* genome, for example, is predicted to encode almost 900 RBPs (Tamburino et al., 2013) out of ~19,000 proteins. The budding yeast genome encodes upwards of 800 RBPs (of ~6,000 proteins). The human proteome contains 1,400 known RBPs (of ~20,000 proteins) (Baltz et al., 2012; Castello et al., 2012; Dezso et al., 2008) with many more proteins predicted to have RNA-binding capabilities. Small RNAs, which function in concert with specific RBPs from the conserved Argonaute family, are also pervasive in metazoan species, ranging from hundreds to tens of thousands of unique small RNAs depending on the subclass. Formation and function of RNPs is vital for cellular homeostasis and organismal development. Mutations in regulatory RBPs have been linked to a variety of human disorders including, but not limited to, cancers (Kechavarzi and Janga, 2014; Yang et al., 2006; Xie et al., 2013; Hirschfeld et al., 2014), neurodegenerative

disorders (Buckanovich et al., 1996; Lu et al., 2009), and cardiovascular disease (Ward et al., 2010; Maatz et al., 2014) (Table 1.1 and reviewed in Gerstberger et al., 2014; Lukong et al., 2008). Similarly, some small RNAs, for example microRNAs, have tumor suppressor or oncogenic properties depending on the mRNAs that they target (Chen et al., 2012) while others (*e.g.* Piwi-interacting RNAs) play major roles in regulating mammalian fertility (Heyn et al., 2012).

RNA-binding proteins. RBPs are the major effector components of PTGR. This highly conserved class of proteins has historically been characterized by the presence of one or more **RNA-binding domains** (RBDs), which fall into one of a number of families, summarized in Table 1.1. RBPs share common structural, sequence, and functional characteristics to recognize and bind general, sequence-specific, or structure-specific RNA substrates. RBP:target interactions are often driven by hydrogen bond formation and favorable electrostatic interactions between amino acid side chains within RBDs and a few specific nucleotides within a single-stranded target RNA (Mitchell and Parker, 2014; Auweter et al., 2006). In a second mode of RBP:target interaction, RBPs recognize and bind short, highly structured double-stranded RNA regions potentially agnostic of a particular nucleotide sequence (reviewed in Serganov and Patel, 2008). A common structural theme of RBPs is the modularity of their RBDs. RBPs often contain multiple copies of different RBDs in varying structural arrangements to enable flexibility and specificity of RBP:RNA targeting (Lunde et al., 2007). RBDs can also be arranged within the tertiary structure of an RBP to form a pocket or cleft in which the interacting RNA molecule can favorably reside. In addition to their RBDs, RBPs harbor additional enzymatic or catalytic domains corresponding to the post-transcriptional regulatory function that they impart on their substrate RNAs.

In the last five years, much emphasis has been placed on comprehensively identifying "RBPomes" - the entire collection of RBPs - in multiple systems in order to define key fea-

tures of canonical RBPs and to identify novel RBPs. The earliest RBPome identification studies used high-throughput *in vitro* probing of high-density protein microarrays with fluorescently labeled RNAs to identify proteins that preferentially interact with either total RNA or mRNA in *S. cerevisiae* (Scherrer et al., 2010; Tsvetanova et al., 2010). More recently, *in vivo* approaches have been developed that couple UV crosslinking of RBP:RNA complexes with oligo(dT) purification and mass spectrometry to interrogate mRBPomes in HEK (Baltz et al., 2012), HeLa (Castello et al., 2012), and mouse embryonic stem cell lines (Kwon et al., 2013b). These *in vivo* methods have also been applied to glucose-starved yeast to reveal the RBP components of stress-induced RNP granules (Mitchell et al., 2013).

In parallel to the experimental efforts to identify RBPs, *in silico* approaches have been developed to predict whether a protein has potential RNA-binding ability based on features of known RBPs including amino acid sequences, structural RBDs, homology to known RBPs, and other physio-chemical properties (reviewed in Dieterich and Stadler, 2013; Puton et al., 2012). Recently, computational techniques were applied to the *C. elegans* proteome to predict a comprehensive compendium of almost 900 RBPs (Tamburino et al., 2013). While useful, these computational methods necessarily rely on high-quality information regarding known RBPs, which is still lacking on many fronts. More specifically, canonical RBPs, whose major roles include RNA-binding, contain canonical RBDs and predominate the catalogue of known RBPs. Lacking is information about proteins that have other primary molecular functions but may also bind RNA in specific biological contexts. For example, multiple studies have shown that select metabolic enzymes bind RNA, in some cases in physiologically-meaningful contexts, in a phenomenon referred to as the RNA-enzyme-metabolite (REM) network of regulation (Castello et al., 2012; Cieřla, 2006; Hentze, 1994; Hentze and Preiss, 2010). The prevalence of the REM network, and other regulatory networks involving non-canonical RBPs, remains unknown.

An emerging feature of RBPs is the presence of low complexity (LC) and/or intrinsically disordered (ID) domains that are required for RBP aggregation into RNP complexes. Low complexity domains are defined by sequences containing repeats of single amino acids or short amino acid motifs. The specific function of many LC domains is unknown, but they often provide interfaces for protein-protein and protein-nucleic acid interactions (Coletta et al., 2010; Dyson and Wright, 2005). In two seminal papers from the lab of Steven McKnight, LC domains were shown to be both necessary and sufficient to induce a phase transition to a hydrogel state *in vitro* (Han et al., 2012; Kato et al., 2012). These hydrogels exhibit dynamic behaviors in which, dependent upon local concentrations of components, temperature, and phosphorylation state, LC domains drive transitions between soluble and polymerized, amyloid-like fiber phases. In some cancerous cells, translocation events result in the pathogenic addition of RBP LC domains - *e.g.* LC domains from fused in sarcoma (FUS), Ewings sarcoma (EWS), and TAF15, together called FET - to DNA-binding proteins (Lessnick and Ladanyi, 2012; Guipaud et al., 2006; Arvand and Denny, 2001). *In vivo*, the McKnight lab offers evidence that assembly of the FET LC domains into polymeric fibers can recruit and activate the C-terminal domain (CTD) of RNA Pol II to promote transcription initiation. This process is reversible by phosphorylation of the CTD (Kwon et al., 2013a), which releases RNA Pol II to proceed with transcription elongation. This mechanism offers a simple explanation for how the observed RBP LC domain translocations can promote cancer; however, the amount of polymerization required to produce a pathogenic effect is still unclear. The biological importance of LC domains is further underscored by the observation and examination of pathological aggregates. For example, the expansion of LC repeats can cause neurodegenerative disorders such as Huntington's disease (Gatchel and Zoghbi, 2005) as well as developmental disorders (Brown and Brown, 2004). ID domains are regions of proteins that do not form stable secondary or tertiary structures but rather are flexible and

capable of adopting structured configurations driven in part by interactions with binding partners. These regions frequently form multiple interaction interfaces that can promote RNA folding by RNA chaperones (Dyson and Wright, 2005; Tompa and Csermely, 2004) and RNA binding (Phan et al., 2011).

Small RNAs. In addition to RBPs, small noncoding RNAs are potent post-transcriptional regulators of gene expression. Small RNAs are conserved, short (\sim 20-40nt), single-stranded RNAs that do not encode for proteins but rather base pair with perfect or imperfect sequence complementarity to target RNAs to elicit a downstream response. Small RNAs typically function as part of an RNA-induced silencing complex (RISC), the core effector component of which is a member of the conserved Argonaute family of proteins. Small RNAs engaged in the RISC typically promote target silencing by one of a number of distinct mechanisms including endonucleolytic target cleavage, deadenylation, and formation of heterochromatin at target loci. Major subclasses of small RNAs include microRNAs (miRNAs), endogenous short-interfering RNAs (endo-siRNAs), and Piwi-interacting RNAs (piRNAs), all of which can be characterized by their size, 5' chemical modification, associated protein cofactors, function, expression, and patterns of inheritance.

miRNAs are conserved across all metazoan species. In animal systems, miRNAs silence their target mRNAs through imperfectly base-pairing with 3'UTRs to elicit one of two possible mechanisms: mRNA deadenylation and exonucleolytic degradation and/or translational inhibition (Baek et al., 2008; Bartel, 2009; Bazzini et al., 2012; Guo et al., 2010; Hendrickson et al., 2009; Subtelny et al., 2014). Genomically encoded miRNAs are transcribed as long, primary miRNA transcripts and then processed through a tightly regulated pathway until a mature, \sim 22nt miRNA is derived. Loaded into their cognate Argonaute/RISCs, miRNAs form imperfectly-paired RNA duplexes with target 3'UTRs to silence them.

The first identified miRNAs, *lin-4* and *let-7*, were discovered in *C. elegans* in 1993 and

2000, respectively, as regulators of developmental timing (Reinhart et al., 2000; Lee et al., 1993). Although originally thought to be specific to the nematode, hundreds of additional miRNAs were quickly discovered in worm, fly, and human (Kozomara and Griffiths-Jones, 2011; Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). In the years since, miRNAs have been estimated to regulate a vast proportion of any organism's set of mRNAs, including ~60% of mRNAs in humans (Ebert and Sharp, 2012; Friedman et al., 2009). miRNAs have also been linked to pathogenic roles in many human disease such as metabolic disorders, cardiovascular disease, and cancer (miRNA and disease database, Jiang et al., 2009). Thus, understanding the role of miRNAs in PTGR is vital to our efforts to treat human diseases.

piRNAs are a class of conserved small RNAs that are bound by the Piwi subclass of Argonaute proteins. piRNAs are expressed in metazoans most abundantly in animal germline and stem cell populations. In mammalian and fly germ cells, Piwi:piRNA complexes recognize perfectly complementary sites on transposable genetic elements in order to silence them and protect future offspring from their mutagenic effects (Aravin et al., 2007; Siomi et al., 2010). In flies, piRNAs also have an epigenetic role in chromatin modification at transposon genomic loci to allow for trans-generational inheritance of piRNA-mediated PTGR from mother to offspring (Yamanaka et al., 2014; Le Thomas et al., 2014). In the nematode, piRNA-mediated regulation is seemingly more elaborate. Only recently has it been shown using deep sequencing experiments that nematode piRNAs, also called 21U RNAs, interact with the Piwi Argonaute PRG-1 to imperfectly target (up to 4 mismatches) aberrant RNAs (*e.g.* transposable elements, pseudogenes, viral RNAs) in the germline (Bagijn et al., 2012; Lee et al., 2012). This interaction triggers biogenesis of a class of endo-siRNAs which form RISC complexes with their own worm-specific Argonautes to silence the piRNA targets. Regardless of the precise mechanism, piRNAs in all organisms are key post-transcriptional

regulators that actively seek out and destroy harmful RNA elements that threaten genome stability.

The intricate collaborations between RBP- and small RNA-mediated PTGR mechanisms are vital for all cellular processes and can function individually or together as either antagonistic (Jafarifar et al., 2011; Kundu et al., 2012) or synergistic (Kim et al., 2009; Jing et al., 2005) factors (Ho and Marsden, 2014). Specific RBPs and small RNAs can converge upon a single RNA target, interacting with multiple *cis*-regulatory elements to govern organismal development, environmental stress response, and cellular homeostasis. Understanding how PTGR mechanisms cooperate or compete with each other is a difficult task due to the fact that (1) a single RNA can be regulated by multiple factors, (2) a single factor can regulate multiple targets, and (3) many systems have built-in redundancies to protect against disruption of one regulatory pathway (Jens and Rajewsky, 2015). Tedious molecular, genetic, and biochemical experiments have resolved only some of these integrated pathways, and understanding remains an area of intense investigation.

1.3 Small RNAs and RBPs regulate gene expression networks

In this section, major PTGR processes will be described highlighting how small RNAs and RBPs contribute to specific mechanisms.

Pre-mRNA processing. In eukaryotes, the first PTGR mechanisms to occur on a nascent transcript are 5' capping, splicing, and 3' processing. These mechanisms are often co-transcriptional as they occur in the nucleus before transcription is terminated. During 5' capping, a complex of multiple enzymes (*e.g.* Ceg1p which contains an OB-fold RBD) catalyzes reactions that add a 7-methylguanosine to the 5' end of the pre-mRNA. Cap addition is essential for coordinating further pre-mRNA processing, translation, and decay;

therefore, dysregulation at this step can dramatically affect the fate of a mature mRNA. During splicing, transcribed introns are removed and protein-coding or non-coding exons are joined together by the spliceosome RNP complex (reviewed in Zhou et al., 2002; Will and Lührmann, 2011), which contains a number of conserved RBPs (*e.g.* Sm and Lsm RBPs which contain Lsm RBDs) as well as small nuclear RNAs (snRNAs). During pre-mRNA 3' end processing, a protein complex containing specific RBPs (*e.g.* cleavage stimulation factor 2 (CSTF2) which contains an RRM RBD) recognizes and binds sequence elements in the pre-mRNA to direct RNA cleavage and polyadenylation. The poly(A) tail is then bound by poly(A) binding proteins (PAPB, which contains RRM RBDs) to protect the mature mRNA from nuclease digestion. Regulation at any of these levels, for example modulation of a poly(A) tail length, are key steps in PTGR.

mRNA export and subcellular localization. After pre-mRNA processing, mature mRNAs are exported from the nucleus to the cytoplasm so they can be translated into the proteins they encode. Nuclear export involves packaging mRNAs into RNP complexes, translocating the RNPs through the nuclear pore complex, and releasing the mRNAs into the cytoplasm (reviewed in Carmody and Wentz, 2009). Export RNPs contain RNA-binding cofactors such as yeast Mip6p, which contains three RRM RBDs and interacts with nuclear pore complex component Mex67p to shuttle polyadenylated RNAs out of the nucleus (Segref et al., 1997). Some mRNAs are localized to specific subcellular regions with the help of RBPs to allow local translation for a rapid supply of the protein where it is needed. mRNA localization is important in large and/or highly polarized cells such as oocytes/embryos and neurons, especially at dendrites and axons (Steward and Schuman, 2003; Piper and Holt, 2004; Blower, 2013; Du et al., 2007). Localization and translation of tau mRNA to the axon and growth cone requires the RBP HuD (which binds to an AU-rich element in the tau 3'UTR) and a kinesin microtubule-associated motor protein (Aronov et al., 2002).

Regulating localized protein synthesis is far more efficient than eliciting a transcriptional response. In yeast, the RBP She2p (which contains an uncharacterized RBD) localizes approximately 22 mRNAs to the bud tip of dividing yeast cells (Long et al., 2000; Takizawa and Vale, 2000). Also, the RBP Puf3p (which has a PUM RBD) is responsible for localizing mRNAs that encode proteins required for mitochondrial function to mitochondria themselves (Gadir et al., 2011; Saint-Georges et al., 2008).

A common theme among all mRNA subcellular localization mechanisms is the use of cytoskeletal proteins as a framework for shuttling RNPs through the cytoplasm. The RBP FMRP, for example, requires association with microtubule filaments to transport specific mRNAs to dendrites (Dictenberg et al., 2008). Additionally, a genome-wide survey of microtubule-bound mRNAs in *Xenopus laevis* egg and human cells undergoing metaphase revealed many classes of mRNAs associated with these cytoskeletal components (Blower et al., 2007). Identification of both translationally-active and -inactive mRNAs at mitotic spindles suggests that regulation of local protein synthesis as well as proper mRNA segregation during cell division might depend on the RNA-binding ability of mitotic spindles and, more generally, cytoskeletal proteins.

mRNA modifications. The widespread use of transcriptome-wide, reversible RNA modifications as gene expression regulatory mechanisms has only recently begun to be uncovered. In one example, N⁶-methyladenosines (m⁶A) in mRNAs and long non-coding RNAs in eukaryotes regulate RNA stability (reviewed in Xu et al., 2014; Fu et al., 2014). The m⁶A modification is established by a methyltransferase complex (Wang et al., 2014e; Liu et al., 2014), removed by demethylases (Zheng et al., 2013; Jia et al., 2011), and recognized by a recently identified family of RBPs that contain the YTH RBD (Dominissini et al., 2012; Fu et al., 2014), a domain that until recently has had no clearly characterized function. Specifically, YTHDF2 has been characterized as the first ‘reader’ of m⁶A modifications which re-

localizes mRNAs from the pool of actively translated mRNAs to cytoplasmic RNP granules, called processing bodies (P bodies), where mRNA degradation occurs (Wang et al., 2014d). Thus, chemical modifications of individual nucleotides (*e.g.* m⁵C, pseudouridylation, and 2'-OMe) within a protein-coding transcript represent a novel mechanism of PTGR.

mRNA storage and degradation. A cell's ability to either (1) target mRNAs for degradation when they are no longer needed or have been corrupted or (2) stably store mRNAs during intracellular transport or in response to stalled translation is greatly influenced by the formation of specific, non-membrane-bound, cytoplasmic RNP granules. The RBP components of these granules determine how they regulate the mRNAs contained within them. One of the most pervasive types of RNP granules are P bodies, which have been observed across all organisms ranging from fungi to plants to humans (reviewed in Garneau et al., 2007). These granules are locations for many mechanisms responsible for mRNA turnover. Within P bodies, mRNA degradation can be promoted through poly(A) tail shortening, which is accomplished by the enzymatic activity of the CCR4-NOT deadenylase complex (*e.g.* CNOT4 in humans, which contains an RRM RBD). Deadenylation further promotes mRNA instability through one of two ways. First, the yeast complex made up of RBPs Lsm1-7p binds 3'UTRs of mRNAs targeted for degradation and induces decapping by the Dcp1/2p complex in parallel with decapping activator Edc3p. mRNAs are then susceptible to 5' → 3' exonuclease decay by exoribonucleases, for example Xrn1p. Additionally, deadenylation can stimulate 3' → 5' decay by exosomes. Decapping can also occur independently of deadenylation followed by exonuclease activity. Many mRNAs are recruited to P bodies for degradation by the specific interaction of RBPs with AU-rich elements (AREs) in their 3'UTRs (Barreau et al., 2005).

P bodies can serve as sites for nonsense-mediated decay (NMD) of corrupted transcripts and miRNA-induced target silencing. While P bodies contain many components required

for these processes, they are not essential for mRNA decay and contain only a subset of mRNA decay factors. For example, the core NMD factors, UPF proteins, normally do not exist in P bodies, but can accumulate within P bodies when NMD is impaired, perhaps in an effort to increase local efficiency of NMD (Sheth and Parker, 2006). Similarly, many P body components are also localized diffusely in the cytoplasm, suggesting that mRNA decay might be initiated in the cytoplasm or soluble protein:mRNA complexes before full aggregation of these components into P bodies (Eulalio et al., 2007). In support of this idea, mRNA decay has been shown to occur independently of P bodies in mammalian and *Drosophila* S2 cells (Pauley et al., 2006; Eulalio et al., 2007).

Stress granules are a type of RNP complex that forms specifically in response to cellular stress and contain pre-initiation translation complexes and mRNAs that were primed for translation. Unfavorable cellular conditions lead to stalling of translation and storage of the mRNAs in stress granules until either the stress is removed and translation starts again or the mRNA is destroyed. Microtubules, which can function in cytoplasmic mRNA localization, are also required for stress granule formation as loss of them results in failure of stress granules to form (Ivanov et al., 2003). Another important type of RNP complexes are those present in germline cells. These granules often referred to as germ granules or nuage or P granules (in worm) are required for germ cell fate determination (Santos and Lehmann, 2004), fertility (Spike et al., 2008b,a), and germ cell totipotency (Ciosk et al., 2006). The dynamic nature of all RNP complexes is vital for a cell's ability to respond to cellular and environmental signals.

1.4 Deep sequencing informs mechanisms of PTGR

Small RNA quantification. Small RNAs are empirically defined based on the unique sequence of nucleotides that make up the ~ 20 -32nt RNA molecule. While many small RNAs are processed from a longer precursor RNA (either single- or double-stranded RNA), it is the final, short RNA molecule that is functional and thus is the unit that we want to measure. To assess how small RNAs contribute to PTGR mechanisms, it is necessary to quantify the level at which they are expressed in a particular biological sample. Molecular assays, such as Northern blotting or quantitative reverse-transcription PCR, are useful for measuring levels of a one or a few known small RNA species; however, it is impossible to quantify all small RNAs or identify new classes of small RNAs with these low-throughput assays. The emergence of high-throughput, or deep, sequencing technologies has enabled quantification of all small RNAs in a single experiment as well as driven identification of new small RNAs that have not previously been detected. Small RNA sequencing has successfully been implemented in hundreds of studies as assessed by identifying over 600 entries for over 70 eukaryotic species in the Gene Expression Omnibus (searching for “small RNA” or “miRNA” or “piRNA” or “siRNA” within “Expression profiling by high throughput sequencing” datasets on March 6, 2015).

To isolate small RNAs for deep sequencing, total RNA is extracted from a biological sample and small RNAs are separated from long noncoding or coding RNAs by denaturing PAGE. A band corresponding to RNAs ~ 20 -30nt long is excised, and 5' and 3' adapter sequences are ligated to the ends of the captured RNAs in preparation for reverse transcription and PCR amplification to generate a cDNA library. Among different subclasses of small RNAs in different organisms, the 5'-most nucleotide can either be mono- or triphosphorylated depending on the biogenesis mechanism used to generate the small RNA.

5'-monophosphorylated small RNAs are produced by cleavage of a dsRNA precursor by an RNase III-like enzyme called Dicer. 5'-triphosphorylated small RNAs (also called secondary siRNAs) are produced by an RNA-directed RNA polymerase (RdRP) which synthesizes a small RNA from a target RNA template. This distinguishing 5' chemical structure determines which RNAs are isolated during library preparation: only 5'-monophosphorylated small RNAs can be ligated to the 5' adapter, leaving 5'-triphosphorylated RNAs without an adapter and therefore not reverse transcribed. To specifically capture 5'-triphosphorylated small RNAs, size-selected RNAs are treated with a phosphatase to remove all 5' phosphate groups and a kinase to add a single phosphate group back, resulting in the conversion of all small RNAs to a 5'-monophosphorylated state. In *C. elegans*, for example, miRNAs, piRNAs, and the 26G subclass of endo-siRNAs are 5'-monophosphorylated, while the abundant 22G subclass of endo-siRNAs is 5'-triphosphorylated. Small RNA deep sequencing protocols that only capture 5'-monophosphorylated RNAs are termed 5'-monophosphate-dependent, while protocols that capture all small RNAs are termed 5'-monophosphate-independent. Given that 5'-monophosphate-dependent sequencing libraries lack cDNAs corresponding to 5'-triphosphorylated small RNAs, quantified small RNA levels cannot be directly compared across the two types of libraries.

RBP:RNA interaction identification. There are a variety of RNA and protein:RNA capture and sequencing methods available to study global aspects of PTGR. Some of these methods include: mRNA-seq (reviewed in Chu and Corey, 2012) to quantify transcript levels and detect alternate mRNA isoforms; Gro-Seq (Core et al., 2008) and BruChase-Seq (Paulsen et al., 2013) to monitor nascent RNA synthesis rates; PARS-seq (Kertesz et al., 2010) and SHAPE-seq (Lucks et al., 2011) to investigate RNA secondary structure; PARE-seq to explore RNA degradation rates (German et al., 2009); and ribosome profiling (Ribo-seq) (Ingolia et al., 2009) and TRAP-seq (Jiao and Meyerowitz, 2010) to assess ribosome

occupancy and thus translational efficiency of mRNAs. While these approaches are useful for generating insights into RNA biology, Chapters Three through Five address deep sequencing of RNA fragments that physically associate with non-ribosome RBPs. In these methods, proteins and their RNA substrates are crosslinked *in vivo* by use of ultraviolet (UV) light irradiation to induce covalent bond formation between amino acid side chains and nucleotides within Angstroms of each other. RBPs of interest are isolated, bound RNA fragments are RNase digested to the protected RBP footprint size, and then RNAs are subjected to deep sequencing to identify the specific, nucleotide-resolution targets of the RBPs. The low efficiency of UV-induced covalent bond formation is compensated for by the use of PCR to amplify RNAs before deep sequencing. These methods are generally referred to as CLIP-seq (crosslinking and immunopurification followed by deep sequencing) or are derivatives of CLIP-seq methods (Licatalosi et al., 2008).

The use of UV to induce covalent crosslinking of proteins to their closely interacting RNA substrates *in vivo* (Wagenmakers et al., 1980; Greenberg, 1979) followed by immunoprecipitation (Mayrand et al., 1981; Dreyfuss et al., 1984) has widely been used to isolate the protein:RNA complexes. Further isolation of the crosslinked RNAs for cDNA sequencing was initially developed to identify specific RNA targets of the splicing regulator RBP Nova1 in mouse brain (Ule et al., 2003). A few years later, deep sequencing was applied to the RNAs isolated by CLIP to identify the global set of Nova1 RNA substrates in mouse brain, greatly expanding the utility of CLIP to characterize PTGR mechanisms in an *in vivo* system (Licatalosi et al., 2008). Since this pioneering work, CLIP-seq (also referred to as HITS-CLIP for high-throughput sequencing following CLIP) has successfully been implemented in a variety of organisms for endogenous or tagged RBPs to understand complex regulatory pathways including: identification of miRNA/RISC targets in *C. elegans* by Argonaute CLIP-seq (Zisoulis et al., 2010); characterization of FMRP's mRNA substrates and

effect on translation in mouse brain to understand the molecular basis for Fragile X syndrome and autism spectrum disorders (Darnell et al., 2011); and analysis of the mRNA and lncRNA targets of 5' → 3' exonuclease Xrn1p in the yeast RNA degradation pathway (Tuck and Tollervey, 2013). As of March 6, 2015, there are ~300 entries for 7 species (some with multiple cell lines) of CLIP-seq datasets deposited in the NCBI's Gene Expression Omnibus (GEO) data repository (Edgar et al., 2002). Another online database of over 200 CLIP-seq experiments for 49 RBPs in human, mouse, worm, and yeast has been developed independently for exploration of these data (Yang et al., 2015).

Urlaub et al. (2002) described a phenomenon during the reverse transcription phase of CLIP-seq whereby the polymerase loses fidelity in incorporating a nucleotide in the new RNA strand directly across from the covalently crosslinked nucleotide which may still be covalently bound to residual amino acids. König et al. (2010) took advantage of this observation, which was estimated to occur in 8-20% of reads, and developed individual-nucleotide resolution CLIP-seq (iCLIP-seq). In this protocol, reverse transcription is stopped at the UV-induced crosslinked site, creating a prematurely truncated cDNA product. Sequencing of these products results in reads that map directly downstream of the crosslinked nucleotide.

Another modification to the CLIP protocol was recently introduced in which photoactivatable ribonucleoside analogs (*e.g.* 4-thiouridine, 4sU) are incorporated *in vivo* into nascent transcripts at a low rate (Hafner et al., 2010a,b). Incorporated nucleoside analogs facilitate more efficient UV-induced crosslinking and also produce, in the case of 4sU use, a signature T-to-C transition event in resulting sequence reads. This transition event occurs because 4sU preferentially forms hydrogen bonds with guanosine during reverse transcription thus resulting in pairing with cytosine during PCR amplification, which is the nucleotide recovered in deep sequencing libraries. This new technique is called photoactivatable ribonucleoside-enhanced CLIP-seq, or PAR-CLIP-seq and has grown popular very quickly: as of March 6,

2015 there are >300 entries for 7 species (some with multiple cell lines) of PAR-CLIP-seq datasets deposited in GEO.

In Chapter Three, I present computational analyses of sequence data derived from a novel extension of PAR-CLIP that was developed in our lab termed global PAR-CLIP, or gPAR-CLIP. This methodology was developed in an effort to identify all RBP-binding sites across the entire transcriptome in a single organism, budding yeast. By investigating characteristics of these binding sites in wild-type (log-phase growth) yeast and comparing them to characteristics of transcriptome-wide binding sites identified in yeast grown under environmental stress conditions, we can observe dynamic changes in RBP occupancy of the transcriptome, which correlates with remodeling of PTGR networks. To overcome the need to purify RNP complexes by specifically targeting a single RBP with an antibody, we perform chemical biotinylation (covalent attachment of biotin to a molecule) of all proteins and exploit the high-affinity and high-specificity biotin-streptavidin interaction (Green, 1990) to capture all RNP complexes. The RNAs in these complexes can then be deep sequenced to identify all protein:RNA interactions sites transcriptome-wide.

The diversity of the types of RNA isolation and sequencing library preparation protocols used to explore mechanisms of PTGR necessitates the development of rigorous computational tools and methods to probe these large datasets and account for noise and biases in the data. In the next section, I will review some of the most popular tools and techniques available to analyze small RNA-seq as well as CLIP-seq, iCLIP-seq, and PAR-CLIP-seq (hereafter all referred to as “CLIP-seq”) data, highlighting nuances in the data that make analyzing it a bioinformatic challenge.

1.5 Deep sequencing data necessitate sophisticated computational methods

Small RNA quantification. Processing small RNA-seq data requires some basic steps that can be accomplished through company-based software (*e.g.* Strand NGS platform from Illumina), user-developed scripts (*e.g.* written in perl or python), or published tools (*e.g.* ShortStack (Axtell, 2013)). These steps include (Fig. 1.1): (1) removal of low-quality sequencing reads; (2) read de-multiplexing, if needed; (3) removal of 5' and 3' barcode and adapter sequences from reads; (4) read quantification by aggregating unique sequences; (5) read alignment and annotation to a reference sequence; and (6) read normalization to total mapped library size (reported as reads per million mapped reads, RPM). In step 5, sequencing reads may be mapped using one of many available mapping tools (*e.g.* Bowtie 2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009)) to a reference genome, transcriptome, or database of known miRNA sequences (*e.g.* miRBase (Kozomara and Griffiths-Jones, 2014)).

A major post-processing application is to identify the RNA targets of small RNAs. Target identification is accomplished by determining small RNAs that map antisense to an annotated gene, either with perfect complementarity, as in the case of secondary endo-siRNAs, or with imperfect complementary, as in the case of miRNAs. Target identification for miRNAs is a complex computational problem because, while the mechanism of action is fairly well characterized, the rules of miRNA:target interaction are constantly being updated as new targets are identified, necessitating refinement of the base-pairing rules between miRNAs and target mRNAs. This problem is outside the scope of this dissertation but reviews can be found here (Bartel, 2009; Hausser and Zavolan, 2014). A second post-processing analysis is to look for differential expression of small RNAs between samples. Two tools both available as R Bioconductor packages DESeq (Anders and Huber, 2010) and baySeq (Hardcastle and Kelly, 2010) have been shown to perform well for detecting differential expression of small

RNAs when compared to other methods (Cordero et al., 2012). Both approaches assume a negative binomial distribution for the data (quantified counts of each small RNA in a sample). baySeq uses an empirical Bayesian approach to generate posterior probabilities for all small RNAs and then estimate a prior distribution based on all the samples as a whole. The following tools/web servers can also perform differential expression using the noted methods: miRAnalyzer (DESeq) (Hackenberg et al., 2009) and omiRas (DEseq) (Müller et al., 2013). Because of their short length and lack of spliced isoforms, detecting differential expression of small RNAs is more straightforward than it is for mRNAs. Finally, network visualization of differentially expressed small RNAs and their targets is useful and can be performed using the web servers mentioned above.

One of the persistent challenges associated with small RNA sequencing data analysis is related to re-using datasets that have already been published. Many earlier studies lack technical or biological replicate libraries. There is also an inability to directly compare absolute small RNA abundances given the biases in cloning that stem from 5'-monophosphate dependent and independent library preparations. Depth of sequencing can also be an issue, as the number of unique small RNAs vastly outnumbers the number of protein-coding genes or other long non-coding RNAs, so that very low abundant small RNAs might be missed. As sequencing depth and cost continues to improve exponentially, these problems will likely be resolved.

RBP binding site identification. Many tools and algorithms have been published describing ways to analyze CLIP-seq data with the ultimate goal of identifying, with nucleotide-resolution, the boundaries demarcating RNA regions that physically associate with or are bound by RBPs. These sites are typically referred to as “binding sites” or “footprints” but could also arguably represent “crosslinking sites” as the binding of proteins to these sites is only inferred by data from the CLIP-seq experiments, suggesting that closely interacting

nucleotides and amino acids are covalently crosslinked. The major steps of binding site identification from CLIP-seq data are (Fig. 1.2): (1) process sequence data, (2) align reads to a reference sequence, (3) normalize read counts, (4) generate read clusters, (5) refine cluster boundaries to define a binding site (which is on average $\sim 20-40$ nt), (6) annotate binding sites, (7) perform downstream analyses (Fig. 1.2). Steps 1 through 3 are almost identical to processing of small RNA-seq data and so will not be discussed in detail here. Steps 4 and 5 are the subject of many approaches and methodologies, the most popular of which are detailed next. Generally, the major goals of generating and refining read clusters to infer binding sites on RNAs are to eliminate noise (in the form of low read counts) and assign scores to binding sites based on a number of criteria (*e.g.* read counts, T-to-C mismatch rates for PAR-CLIP-seq data).

CLIPZ (Khorshid et al., 2011) is a computational and visualization platform that allows users to explore pre-loaded or user-loaded CLIP-seq data. Clusters are generated from overlapping reads, and summary statistics for these clusters is provided, but no further assessment of binding site confidence is currently provided.

PARalyzer (Corcoran et al., 2011) is a highly popular tool designed specifically for use with PAR-CLIP-seq data. The rate of T-to-C mismatch and the rate of T-to-T non-mismatches along read clusters are smoothed with a non-parametric kernel density estimator. RBP:RNA interaction sites are defined by regions where the likelihood of T-to-C is higher than T-to-T events and a minimum read count threshold is exceeded.

Pirhana (Uren et al., 2012) is a software tools that is generally applicable to all CLIP-seq-related datasets. It utilizes techniques that model the underlying distribution of read counts from individual binding sites and allows incorporation of transcript abundances and other external data sources to improve site identification.

wavClusterR (Sievers et al., 2012; Comoglio et al., 2015) is a computational framework

implemented in R and specifically designed for PAR-CLIP-seq data. In this method, a Bayesian model is used to identify high-confidence T-to-C events, and a false discovery rate is calculated from T-to-C events observed in matched total RNA-seq data.

Many recent CLIP-seq data analysis tools are emerging that incorporate HMMs to interrogate binding site confidence ((Yun et al., 2014), RIPseeker (Li et al., 2013), MiClip (Wang et al., 2014b), dCLIP (Wang et al., 2014c)). The miRTarCLIP algorithm is specific for identifying miRNA:RISC target sites from Argonaute CLIP-seq experiments (Chou et al., 2013). Finally, the CIMS analysis method leverages single nucleotide deletions detected at low frequency in CLIP-seq reads to improve the precision of binding site identification (Zhang and Darnell, 2011). A comprehensive review of these and other CLIP-seq data analysis methods not discussed in this dissertation was published by Reyes-Herrera and Ficarra (2014)).

1.6 Investigating PTGR poses bioinformatic challenges

The study conducted in Chapter Two was motivated by the question of whether there are germline-specific expression patterns for piRNAs in *C. elegans*, how differential expression is achieved, and whether germline-specific piRNAs perform germline-specific PTGR of their targets. To answer the first question, I conducted a preliminary inspection of over 20 small RNA-seq datasets mined from GEO and published in numerous studies addressing other biological questions. This investigation revealed germline-specific characteristics of piRNAs. Since our group did not generate the datasets, and a mix of 5'-monophosphate-dependent and -independent library preparation protocols were used, standard differential expression detection techniques could not be applied. Rigorous testing of differential expression of piRNAs was achieved through the development of an Enrichment Score metric based on multiple comparisons of individual datasets. Enrichment Score thresholds to assign piRNAs as

male or female germline-enriched were empirically derived by analyzing randomized control datasets modeled after a single sequencing dataset and assigning thresholds corresponding to a 1% false discovery rate. Classifications of select piRNAs were validated experimentally and shown to rely on distinct sequence motifs located upstream of their genomic loci.

The studies presented in Chapters Three and Four of this dissertation deal with processing of PAR-CLIP-seq data to identify RBP:mRNA interactions in budding yeast. The studies were motivated by a lack of fundamental knowledge of RBP:RNA biology including: What are the general characteristics of RBP:RNA interactions prevalent throughout all eukaryotes? What can be learned about the specific PTGR mechanisms of select RBPs by defining their precise RNA interaction sites? How can knowledge of general RBP:RNA interactions be used to enhance our understanding of the RNA-binding properties of a conserved family of RBPs?

The major analysis method available at the time of these studies for PAR-CLIP-seq data was PARalyzer, which, as described above, uses kernel density estimators of read coverage across a binding site to establish binding site boundaries. Specifically, read counts covering genomic T positions (technically U positions in RNAs) were separated based on whether a C was sequenced (a signature of direct crosslinking in the PAR-CLIP protocol) or a T was sequenced. Sequenced Ts represent “non-crosslinked” positions in PARalyzer; however, there is a major problem with this assumption: presence of a sequenced T at T positions does not necessarily indicate lack of contact between the 4sU nucleotide and an RBP. Rather, a sequenced T can represent a true RBP-binding position in which (1) a 4sU nucleoside analog was not incorporated into the nascent transcript (which is highly probable given that we estimate 1 in 40 Ts are replaced with 4sU), (2) reverse transcription did not induce a 4sU:G base pairing, or (3) in rare instances, sequencing error switches the 4sU-induced C back to a T. Therefore, by requiring the T-to-C rate to be higher than the T-to-T rate to call a

region an RBP:RNA interaction site requires at minimum roughly half of the read coverage to show a conversion of T to C, a very strict requirement. In the methods of Chapter Three, I present an approach to calling PAR-CLIP-seq binding sites that makes use of kernel density estimators to smooth the read coverage of all reads aligning perfectly or with up to 3 T-to-C mismatches to call *bona fide* binding sites. In the majority of cases, low T-to-C conversion rates are caused by sequencing error. Therefore, I assign an FDR rate to each PAR-CLIP-seq binding site based on the number of T-to-C mismatches observed in binding sites derived from mRNA-seq data, of which all T-to-C mismatches are likely caused by sequencing error. This threshold is unique to sets of binding sites binned on total read coverage since higher coverage sites are more likely to contain more sequencing errors.

While many of the PTGR mechanisms discussed here have been characterized to varying degrees, many general themes and rules regarding PTGR have, before now, only been hypothesized. By applying known and established bioinformatics methods to deep sequencing datasets, and developing and implementing novel computational frameworks based on specific experimental parameters of our sequencing datasets, my doctoral research contributes important insights underlying protein and RNA interactions and their contributions to the regulatory networks that orchestrate PTGR.

Table 1.1: Classical RNA-binding domains play major roles in disease pathologies.

Domain name	Shorthand	Yeast	human RBP	Associated human disorder
C2H2 zinc finger	Znf_C2H2	Nam7p, <u>WT1</u>		Denys-Drash syndrome (Little and Wells, 1997)
CCCH zinc finger	Znf_CCCH	Nab2p, <u>MCPIP1</u>		inflammatory diseases (Liang et al., 2008)
Cold-shock domain	CSD	N/A, <u>YBX1</u>		various cancers (reviewed in Lasham et al., 2013)
K Homology	KH	Khd1p, <u>NOVA1</u>		paraneoplastic motor disorder (Buckanovich et al., 1996)
Like Sm domain	Lsm	Lsm1p, <u>LSM1</u>		breast cancer (Yang et al., 2006)
Lupus La RBD	La	Sro9p, <u>LARP1</u>		hepatocellular carcinoma (Xie et al., 2013)
Oligonucleotide-binding fold	OB-fold	Ceg1p, <u>RPA4</u>		colon cancer (Keshav et al., 1995)
PUA domain	PUA	Cbf5p, <u>DKC1</u>		dyskeratosis congenita (Heiss et al., 1998)
Pumilio-like repeat	PUM	Puf3p, PUM1		N/A
PWI domain	PWI	Snu71p, <u>RBM25</u>		cardiac disease/heart failure (Gao and Dudley, 2013)
Ribosomal protein S1-like	S1	Rrp5p, <u>EIF2S1</u>		MYC-driven cancers (Dey et al., 2013)
RNA Recognition Motif	RRM	Pab1p, <u>ELAVL1</u>		amyotrophic lateral sclerosis (Lu et al., 2009)
Surp module/SWAP	SURP	Prp21p, <u>SFRS8</u>		asthma (Brasch-Andersen et al., 2006)
Thump domain	THUMP	Trm11p, <u>THUMPDI</u>		N/A
Sterile Alpha Motif	SAM	Vts1p, <u>Smaug</u>		myotonic dystrophy type 1 (de Haro et al., 2013)
TROVE module	TROVE	N/A, <u>TEP1</u>		male infertility (Yan et al., 2014)
YTH domain	YTH	Tma19p, <u>YTHDC1</u>		endometrial cancer (Hirschfeld et al., 2014)

Note: Underlined human protein names indicate association or correlation with the listed human disorder.

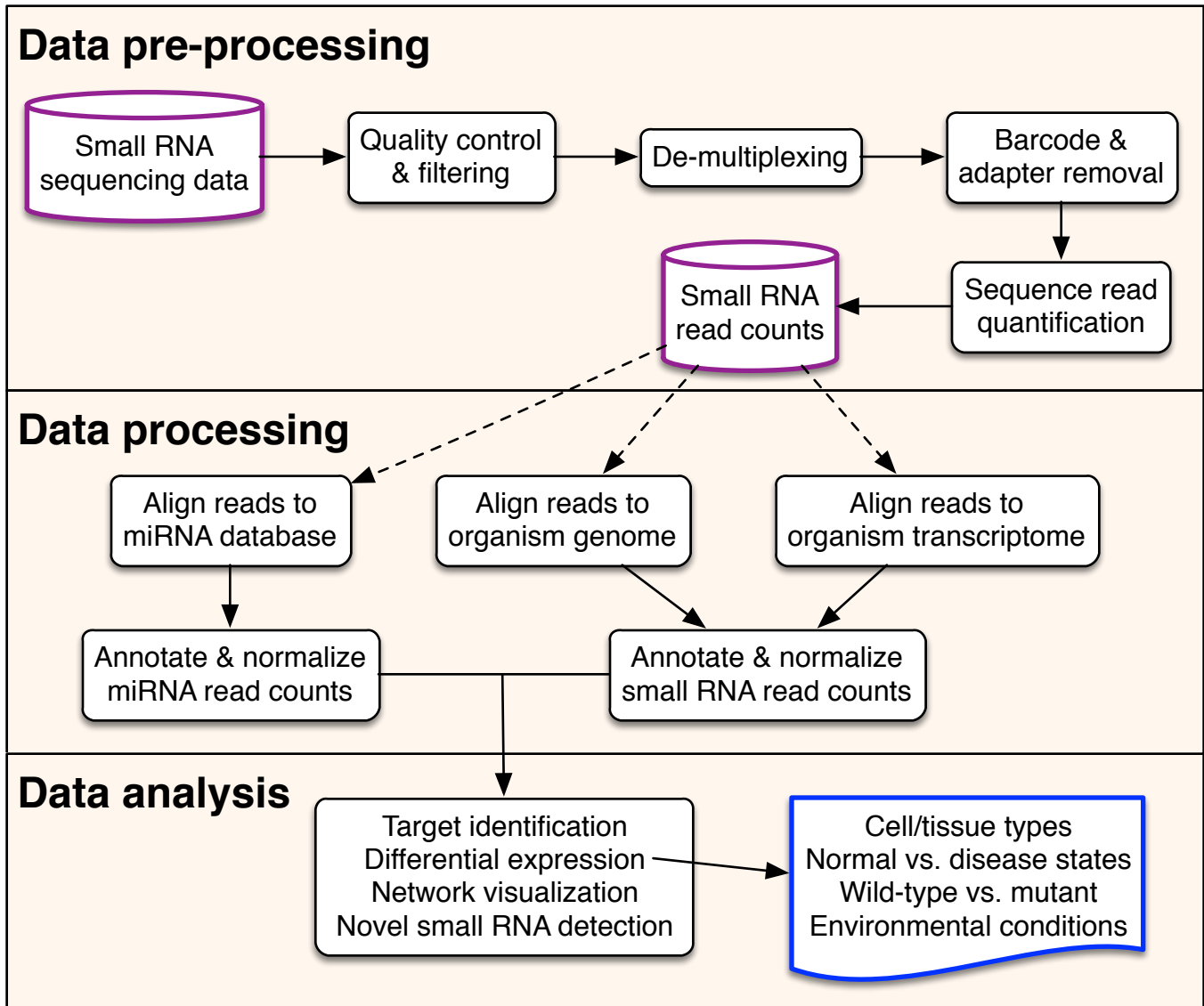


Figure 1.1: Major computational steps required for analysis of small RNA-seq datasets.

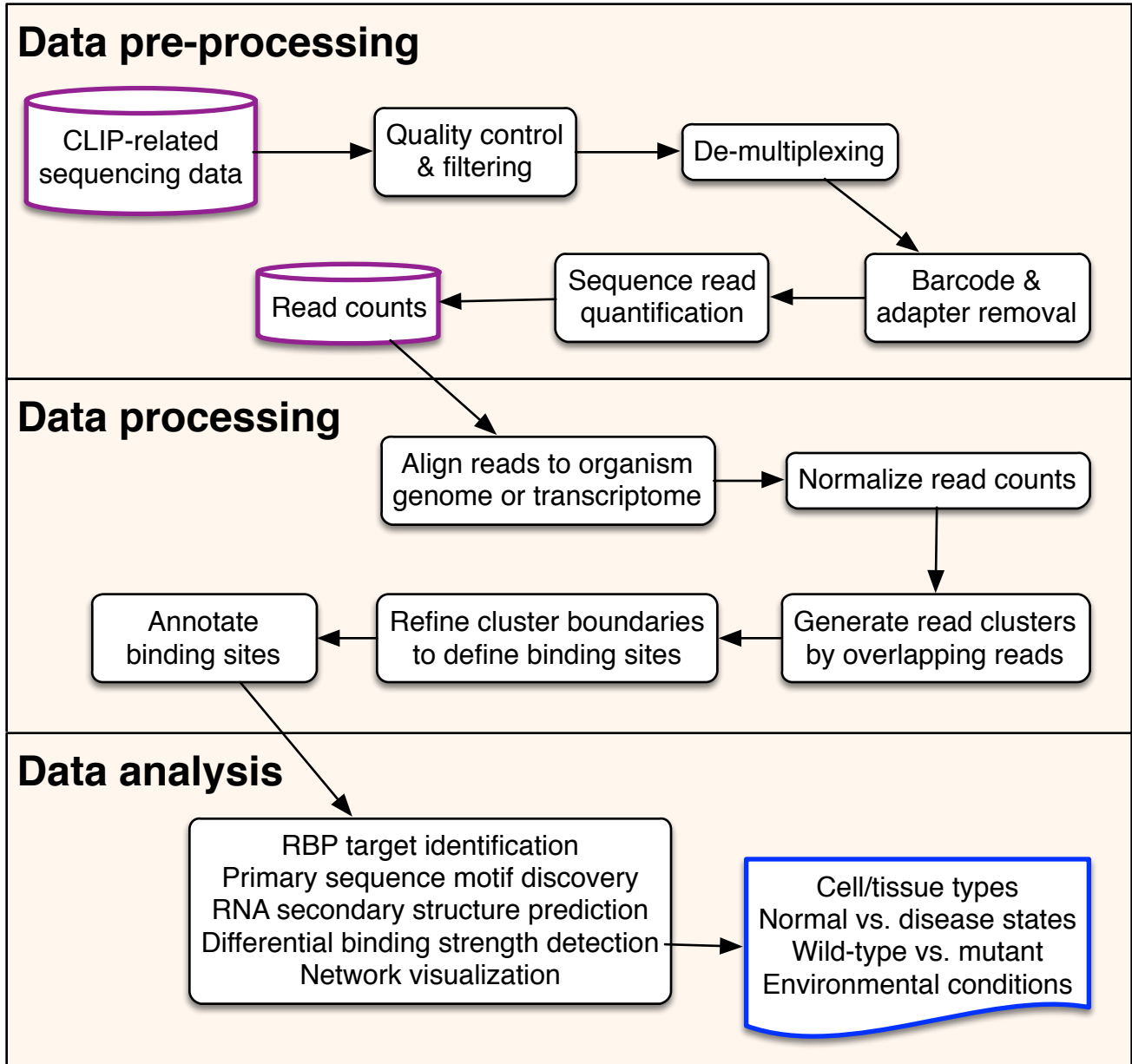


Figure 1.2: Major computational steps required for analysis of CLIP-seq-derived sequencing datasets.

CHAPTER II

Small RNA-mediated PTGR investigated by small RNA-seq

2.1 Introduction

piRNAs and Piwi clade Argonautes arose in the primordial metazoan ancestor (Grimson et al., 2008) and are generally restricted to the germline, where they act in an RNA-induced silencing complex (RISC) to silence foreign genetic elements. From protozoa to mammals, loss of Piwi proteins, and consequently piRNAs, results in abnormal fertility phenotypes or sterility, revealing their highly conserved and essential role in animal reproduction (Carmell et al., 2007; Cox et al., 1998; Kuramochi-Miyagawa et al., 2008; Li et al., 2009a; Lin and Spradling, 1997; Mochizuki et al., 2002). piRNAs are incredibly diverse, with tens of thousands of unique sequences expressed in any single organism. While piRNAs in many organisms map to large, broadly syntenic genomic clusters, the sequences are not conserved among even closely related species, and no unifying sequence features have been identified beyond a bias among primary piRNAs for a 5' uridine (Aravin et al., 2006; Brennecke et al., 2007; de Wit et al., 2009; Girard et al., 2006; Gunawardane et al., 2007; Lau et al., 2006; Ruby et al., 2006).

Official citation:

Billi AC*, **Freeberg MA***, Day AM, Chun SY, Khivansara V, and Kim JK. (2013) A Conserved Upstream Motif Orchestrates Autonomous, Germline-Enriched Expression of *Caenorhabditis elegans* piRNAs. *PLoS Genet* **9**(3): e1003392. doi:10.1371/journal.pgen.1003392. (*equal contribution)

The mechanisms of de novo piRNA biogenesis remain elusive. In fly and mouse, primary piRNAs appear to be processed from long, single-stranded RNA precursors (Aravin et al., 2006; Girard et al., 2006; Brennecke et al., 2007). This long transcript is cleaved by the endoribonuclease Zucchini with little or no sequence specificity to generate candidate piRNA 5' ends (Ipsaro et al., 2012; Nishimasu et al., 2012), which are likely subsequently purified according to the binding preferences of the Piwi proteins that bind primary piRNAs (Kawaoka et al., 2011). Silkworm data suggest that the 3' ends of these piRNA precursors are then trimmed by a 3' to 5' exonuclease until the 3' end is sufficiently short for anchoring by Piwi to protect against further trimming (Kawaoka et al., 2011). The 3' end is then methylated to prevent degradation (Houwing et al., 2007; Kirino and Mourelatos, 2007; Kurth and Mochizuki, 2009; Ohara et al., 2007; Vagin et al., 2006). While recent studies have shed light on the biogenesis of primary piRNAs in many animal models, little is known in any organism about how primary piRNA expression is regulated or how specific sequences are designated as piRNAs.

21U RNAs, a class of germline-enriched small RNAs, represent the piRNAs of *C. elegans*. They are terminally methylated (Billi et al., 2012a; Kamminga et al., 2012; Montgomery et al., 2012), show a 5' uridine bias (Ruby et al., 2006), and are dependent upon and bound by the Piwi Argonaute PRG-1 (Batista et al., 2008; Das et al., 2008), which is required for normal fertility (Cox et al., 1998). Yet *C. elegans* piRNAs exhibit some unusual features. While the vast majority of 21U RNAs map to two large genomic clusters on chromosome IV, the loci do not exhibit prominent strand biases (Ruby et al., 2006). The 21U RNAs also do not appear to play a prominent role in silencing transposable elements, a main function of mouse and fly piRNAs, nor do they engage a ping-pong amplification mechanism (Batista et al., 2008; Das et al., 2008). Rather, PRG-1 and the 21U RNAs target aberrant and coding transcripts broadly via imperfect complementarity, triggering production of secondary

endogenous siRNAs (Bagijn et al., 2012; Batista et al., 2008; Billi et al., 2012b; Das et al., 2008; Lee et al., 2012). These 21U RNA-dependent 22G RNAs can induce chromatin changes to establish dominant, heritable target silencing (Ashe et al., 2012; Luteijn et al., 2012; Shirayama et al., 2012). 21U RNAs evolve rapidly, presumably constrained only by selection against sequences that silence mRNAs; thus, mismatch-tolerant 21U RNAs constitute an epigenetic memory of self versus non-self. Finally, a conserved motif lies upstream of 21U RNA genomic loci (Ruby et al., 2006). This stretch of sequence, which includes an eight-nucleotide (nt) core motif approximately 40 nt upstream of the 21U RNA locus, is conserved across divergent nematodes (de Wit et al., 2009; Ruby et al., 2006). Recently, Cecere et al. (2012) found that this motif is bound by forkhead family transcription factors and that deletion of the core motif abrogates 21U RNA expression, but it is still unknown how 21U RNA sequences are defined and how their expression is regulated.

Here, we demonstrate that piRNAs are expressed autonomously in *C. elegans*. Combining computational and transgenic approaches, we find that the conserved core motif defines the piRNA transcriptional cassette, specifying expression of 21U RNAs from genomic thymidines situated at an optimal distance downstream to determine which genomic sequences are expressed as *C. elegans* piRNAs. Core motifs also encode information dictating germline-specific expression of 21U RNAs. We show that more than 70% of *C. elegans* piRNAs are preferentially enriched in male or female germline. Unexpectedly, this germline enrichment appears to be enforced by a single nucleotide position within the core motif. We demonstrate autonomous expression of synthetic 21U RNAs from multiple minimal transgenic cassettes consisting only of the 8 nt core motif, the, 40 nt intervening genomic spacer, the 21U RNA sequence, and 50-100 nt of flanking genomic context. Finally, we use single-copy transgenes integrated in genomic isolation to show that the clustered organization of endogenous piRNA loci is entirely dispensable for robust piRNA expression. Together, our

results suggest that each 21U RNA locus encodes all of the information necessary for driving independent, autonomous transcription from more than 15,000 unique piRNA loci in *C. elegans*.

2.2 A conserved upstream motif orchestrates autonomous, germline-enriched expression of *Caenorhabditis elegans* piRNAs

A majority of 21U RNAs are male or female germline-enriched. To investigate the mechanisms regulating piRNA expression, we first identified 21U RNA subclasses by performing a meta-analysis of over 50 million reads from published small RNA deep sequencing datasets (Batista et al., 2008; Conine et al., 2010; Gent et al., 2009, 2010; Gu et al., 2009; Han et al., 2009; Kato et al., 2009; Stoeckius et al., 2009) (Table 2.1). Using the pipeline shown in Fig. 2.1A, we determined that a majority of the 13,711 21U RNAs represented in our composite dataset show differential germline enrichment, distinguishing 7,677 (56.0%) unique male and 2,171 (15.8%) unique female germline-enriched 21U RNAs (hereafter, male and female 21U RNAs) (Materials and methods). The distribution of 21U RNA Enrichment scores is skewed toward the male (Fig. S2.1A), whereas randomly generated 21U RNA count data show no significant skewing (Binomial test, $p=0.245$) and define a false discovery rate below 1% (Fig. S2.1B). To assess the reliability of the Enrichment score in classifying germline enrichment, we quantified the average relative abundance of every male 21U RNA between each pair of male and female libraries (Fig. S2.1C); the reciprocal calculation was performed for female 21U RNAs (Fig. S2.1D). On average, the abundance of male 21U RNAs is 6.8-fold higher in male libraries than female, whereas the abundance of female 21U RNAs is 2.4-fold higher in female libraries than male. Average abundance of 21U RNAs not classified as male or female (hereafter, non-enriched 21U RNAs) is approximately equal in male and

female libraries (Fig. S2.1E). Taqman RT-qPCR of select 21U RNAs in *fem-1(hc17)* adult female versus *him-8(e1489)* or *fog-2(q71)* adult male animals shows segregation of 21U RNAs according to germline enrichment classification (Fig. 2.1B,C), endorsing our computational discovery of germline-enriched piRNA subclasses in *C. elegans*.

Male and female 21U RNAs show different expression profiles in embryo.

Our meta-analysis also revealed a subpopulation of 21U RNAs highly abundant in embryo. Comparison of the abundances of male and female 21U RNAs in mixed stage embryo sequencing libraries showed that female 21U RNAs were overrepresented in embryo relative to male. A higher proportion of unique female 21U RNAs were detected in embryo (χ^2 test, $p=9.2e-245$) (Fig. S2.2A,B). Furthermore, unique female 21U RNAs were on average 4.4-fold more abundant in embryo than unique male species (Welch's *t*-test, $p=3.4e-2148$). The trend is corroborated by Taqman analysis showing depletion of male 21U RNAs and enrichment of female 21U RNAs in embryo (Fig. S2.2C-E). These data suggest that female piRNAs are preferentially inherited into *C. elegans* embryo, consistent with previous observations in fly (Brennecke et al., 2008; Chambeyron et al., 2008; Grentzinger et al., 2012). Parallel classification and embryonic enrichment analysis of 26G RNAs, germline-enriched primary endo-siRNAs, recapitulated previously observed inheritance patterns (Han et al., 2009) and validated the ability of our pipeline to identify germline-enriched small RNA subclasses (Fig. S2.2F,G).

Male 21U RNA targets reflect spermatogenic gonad restriction.

21U RNAs target transcripts with imperfect complementarity of up to three mismatches to trigger production of antisense 22G RNAs proximal to the targeting site (Ashe et al., 2012; Bagijn et al., 2012; Lee et al., 2012; Shirayama et al., 2012). The lax complementarity requirement

for piRNA-mediated silencing predicts widespread targeting capacity. Compartmentalization of piRNA expression to the male and female germline may help to confer specificity. To investigate the biological significance of germline-enriched 21U RNA subclasses, we first examined whether male and female 21U RNAs target distinct subsets of genes. We analyzed the overlap between their respective dependent 22G RNAs by identifying 22G RNAs that map antisense to within 40 nt of 21U RNA target sites (Lee et al., 2012) (Materials and methods). Ignoring 22G RNAs detected in *prg-1(n4357)* deep sequencing datasets, as these are likely not 21U RNA-dependent, we identified 11,377 (72.3%) unique 22G RNAs that are likely male 21U RNA-dependent and 3,855 (24.5%) unique 22G RNAs that are likely female 21U RNA-dependent (Fig. S2.3A). Only 494 (3.1%) unique 22G RNAs lie within 40 nt of both a male and female 21U RNA target site, precluding assignment to either category. This overlap is less than expected when 22G RNAs from random but similarly sized sets of 21U RNAs are compared (χ^2 test, $p=0.012$). We then compared the 5,956 male and 1,387 female 21U RNA targets identified in young adult (Bagijn et al., 2012) and gravid (Lee et al., 2012) animals, respectively. Overlap between targets (149 overlapping targets) is significantly lower compared to random sets of genes (294 overlapping and 6,756 non-overlapping targets; χ^2 test, $p=7.7e-213$) (Fig. S2.3B).

Because targets of 21U RNAs are subject to transgenerational silencing (Ashe et al., 2012; Luteijn et al., 2012; Shirayama et al., 2012), 21U RNAs are unlikely to evolve to target transcripts required in the germline. Similarly, male 21U RNAs would not be expected to target transcripts required for spermatogenesis; however, temporal separation of the spermatogenic and oogenic gonads might permit evolution of male 21U RNAs capable of targeting transcripts required for oogenesis. We examined our data for evidence of this evolutionary signature. As comprehensive lists of genes required for spermatogenesis and oogenesis have yet to be assembled, we used as a proxy lists of transcripts identified by microarray studies as

enriched during spermatogenesis (865 transcripts) or oogenesis (1,030) (Reinke et al., 2004). Comparing male 21U RNA targets to randomly generated gene lists, we found that male 21U RNA targets are indeed depleted of spermatogenesis transcripts (χ^2 test, $p=0.044$), but neither enriched nor depleted for oogenesis transcripts (χ^2 test, $p=0.76$) (Fig. S2.3C). Curiously, we do not observe the same signature for female 21U RNAs (Fig. S2.3D). Their targets are neither enriched nor depleted for spermatogenesis transcripts (χ^2 test, $p=0.27$), as expected, but female 21U RNA targets are significantly enriched for oogenesis transcripts (χ^2 test, $p=0.0017$). These differences between male and female 21U RNA targeting suggest that the evolutionary pressures acting on male and female 21U RNA sequences may differ (Discussion).

Male and female 21U RNAs have distinct core upstream motifs. To investigate how 21U RNA germline enrichment information is genetically encoded, we analyzed the genomic loci of the 13,387 21U RNAs that map uniquely to the genome. Comparison of male and female 21U RNA sequences identified no differences in content; therefore, we evaluated the 21U RNA upstream region. The 8 nt core motif, with consensus sequence CTGTTTCA, is separated from the 21U RNA locus by an A/T-rich spacer of, 35 to 42 nt (Ruby et al., 2006). Scanning the 60 nt upstream of each 21U RNA for the best conserved central GTTTC of the core motif, we found that 6,615 of 7,677 (88%) male 21U RNAs show a canonical, GTTTC-containing core motif, compared to only 1,119 of 2,171 (54%) female 21U RNAs. While the length of the A/T-rich spacer does not differ between male and female 21U RNAs (Fig. 2.2A), core motif sequence analysis revealed a striking difference: only the core motifs of male 21U RNAs are enriched for a 5' cytidine. 5,765 of 7,677 (77%) male 21U RNAs are located downstream of canonical, GTTTC-containing core motifs with a 5' cytidine, compared to only 443 of 2,171 (21%) female 21U RNAs (χ^2 test, $p=7.9e-137$) (Fig. 2.2B). To

examine whether this 5' core motif position influences 21U RNA expression, we calculated the average abundance of male and female 21U RNAs grouped by 5' core motif nt. Male 21U RNAs with 5' cytidine core motifs are significantly more abundant than all other male 21U RNAs (Fig. 2.2C, Welch's *t*-test *p*-values in Table 2.2), consistent with the previous observation that 21U RNAs whose core motifs better match the consensus sequence are more highly expressed (Ruby et al., 2006). No other subgroup differs significantly in abundance from all others among the male, female, and non-enriched 21U RNAs (Fig. 2.2C,D, Table 2.2), suggesting that GTTTC-containing core motifs with a 5' cytidine are overrepresented among male 21U RNAs and may drive male germline expression.

A transgenic synthetic 21U RNA recapitulates features of endogenous 21U RNAs. To explore the significance of variation at the 21U RNA upstream motif, we developed a transgenic system to express synthetic 21U RNAs from high-copy, integrated arrays *in vivo* (Fig. 2.3A). 2-3 kilobase regions of genomic sequence from a chromosome IV piRNA cluster were cloned, and a central 21U RNA (male 21U RNA 21UR-1258 or female 21U RNA 21UR-2502) was mutated to a unique synthetic 21 nt sequence (21UR-synth) to distinguish transgenic from endogenous expression. The sequences were then further mutated to generate the panel of transgenes shown in Table 2.3. Transgenes are named for the endogenous 21U RNA replaced by 21UR-synth, with prefixes to indicate transgene type (*e.g.*, ♀Tg2502 represents the otherwise wild-type transgene encoding 21UR-synth in place of 21UR-2502). These transgenes are carried by the vector pCFJ178 (Frokjaer-Jensen et al., 2008), which also expresses the *C. briggsae unc-119* gene (Fig. S2.4A), enabling gross normalization for variable array expression.

To validate our transgenic system, we examined whether 21UR-synth recapitulates all of the known features and genetic sensitivities of endogenous 21U RNAs. 21U RNAs are

2'-O-methylated at the 3' terminus by the conserved methyltransferase HENN-1 (Billi et al., 2012a; Kamminga et al., 2012; Montgomery et al., 2012). Northern blot for 21UR-synth in transgenic strains identified a 21 nt species that is terminally methylated in a *henn-1*-dependent manner (Fig. 2.3B). Robust, specific detection of the 3' terminus by Taqman RT-qPCR (Nolan et al., 2006) confirms that this species corresponds to 21UR-synth (Fig. 2.3C). Levels of endogenous 21U RNAs 21UR-1258 and 21UR-2502 are largely unaffected by expression of the transgenes (Fig. S2.4B). Endogenous 21U RNAs are generated in the germline and require PRG-1 for accumulation (Batista et al., 2008; Das et al., 2008). Accordingly, 21UR-synth is highly depleted by loss of *prg-1* and in the *glp-4(bn2)* germline-deficient mutant (Fig. 2.3C-E). 21UR-synth and endogenous 21U RNAs are also specifically detected in immunoprecipitated PRG-1 complexes, while a microRNA control is not (Fig. 2.3F,G, Fig. S2.5). To rule out the unlikely possibility that transgenic products corresponding to the 21UR-synth sequence might be generated by an alternative, Dicer-dependent mechanism, we assayed 21UR-synth accumulation in a null mutant of *rde-4*. This gene encodes a dsRNA binding protein that is a key cofactor of Dicer in siRNA biogenesis (Sijen et al., 2007; Tabara et al., 2002; Vasale et al., 2010), but dispensable for 21U RNA production (Fig. S2.4B). Loss of *rde-4* does not impair 21UR-synth expression (Fig. 2.3C-E), suggesting that 21UR-synth does not represent an siRNA generated from the high-copy transgenic array.

Finally, we examined whether the core motif is required for 21UR-synth expression. We scrambled the core motif to eliminate any resemblance to the consensus sequence (σ Scram1258 and ϕ Scram2502 transgenes; Table 2.3). 21UR-synth levels in these strains are depleted by more than 100-fold after normalization for array expression (Fig. 2.3H,1), consistent with previous findings that deletion of the core motif depletes 21U RNA expression (Cecere et al., 2012). Together, these data demonstrate that 21UR-synth represents a *bona fide* 21U RNA and support the use of this transgenic system for exploring 21U RNA biology *in vivo*.

21U RNA core upstream motif variation influences germline enrichment. We then used our transgenic system to test whether variation at the core motif 5' position affects germline expression of 21UR-synth (Fig. 2.4A). Endogenous male 21U RNA 21UR-1258, which lies downstream of a CTGTTTCA core motif, peaks in expression during spermatogenesis (52h time point) and is highly expressed in *him-8(e1489)* male adult; in contrast, expression of endogenous 21UR-2502, with an ATGTTTCA core motif, peaks after the spermatogenesis-to-oogenesis transition in adulthood (~72h) and is highly expressed in *fem-1(hc17)* female adult (Fig. 2.4B). Accordingly, the σ Tg1258 and φ Tg2502 transgenes express 21UR-synth in similar male and female patterns, respectively (Fig. 2.4C,D, colored lines/bars). Toggling the core motif from CTGTTTCA to ATGTTTCA (σ C>A1258 transgene) or ATGTTTCA to CTGTTTCA (φ A>C2502) disrupts these germline-specific expression patterns. Whereas 21UR-synth expression from σ Tg1258 plummets after spermatogenesis, loss of the core motif 5' cytidine in the σ C>A1258 transgenic strain results in sustained 21UR-synth expression through oogenesis; the σ C>A1258 transgene also preferentially expresses 21UR-synth in *fem-1(hc17)* female (Fig. 2.4C). Thus, mutating the 5' cytidine of a male 21U RNA core motif results in a failure to restrict 21U RNA expression to spermatogenesis. Similarly, introducing a 5' cytidine into a female 21U RNA core motif impairs restriction of expression to oogenesis: while φ Tg2502 expression of 21UR-synth increases dramatically during the spermatogenesis-to-oogenesis transition, gain of the motif 5' cytidine in the φ A>C2502 transgene dampens this increase (Fig. 2.4D). These results suggest that this single nucleotide orchestrates the accurate switching of 21U RNA expression in the hermaphroditic germline. However, 21UR-synth expression from the φ A>C2502 transgene is still high in *fem-1(hc17)* female, indicating that other elements contribute to female 21U RNA expression patterns. This is consistent with our finding that female 21U

RNA core motifs show no bias at the 5' nucleotide, and indeed, 21% of female 21U RNA core motifs show a 5' cytidine (Fig. 2.2B). As expected, 21UR-synth expression from the σ C>A1258 and φ A>C2502 transgenes is still dependent upon *prg-1* (Fig. 2.4E).

A 5' thymidine is required for robust expression from the 21UR-synth locus.

It is not yet known how individual genomic sequences are selected for expression as piRNAs. As the core motifs, but not the sequences, of 21U RNAs are conserved across *Caenorhabditis* species, it seemed possible that the core motifs themselves might determine what sequences are expressed as 21U RNAs by directing their expression from genomic thymidines located an optimal distance downstream. We explored this hypothesis by mutating the genomic thymidines encoding the first nucleotide of 21UR-synth to adenosine (21U>A transgenes) or guanosine (21U>G transgenes), such that the transgenes encode 21[U>A]R-synth or 21[U>G]R-synth, respectively (Fig. 2.5A, Fig. S2.6A). These putative products emulate the 5' nucleotide identity of microRNAs (predominantly 5' uridine and adenosine) and endo-siRNAs (predominantly 5' guanosine). Small RNAs expressed from these transgenes and recognized by the 21UR-synth northern blot probe differ in size from and are less abundant than wild-type 21UR-synth (Fig. 2.5B, Fig. S2.6B). By Taqman analysis, 21[U>A]R-synth and 21[U>G]R-synth are detected at levels more than 150-fold lower than 21UR-synth after normalization for array expression (Fig. 2.5C, Fig. S2.6C), suggesting that 21[U>A]R-synth and 21[U>G]R-synth are poorly transcribed, stabilized, or both.

The genomic positioning of core motifs specifies 21U RNA sequences. We hypothesized that 21U RNA expression from a particular genomic thymidine may simply be a function of distance from a core motif (*i.e.*, length of the intervening genomic spacer). Therefore, the presence of multiple thymidines within the optimal genomic window down-

stream of a core motif might result in expression of multiple, overlapping 21U RNAs. Indeed, many *C. elegans* piRNAs map to proximal genomic thymidines as members of “miniclusters” of overlapping 21U RNAs that appear to share an upstream core motif. To explore the relationship between core motif position and expression, we extracted read count information from deep sequencing of wild-type adult animals (Batista et al., 2008) for uniquely mapping 21U RNAs and analyzed their corresponding genomic loci. After separating 21U RNAs into those that share a core motif with at least one other uniquely mapping 21U RNA (“miniclustered”; 4,550 21U RNAs) and those that do not (“solitary”; 8,837 21U RNAs), we grouped 21U RNAs by length of genomic spacer and examined their abundance. For both miniclustered and solitary 21U RNAs, the resulting distributions peak at a 39 nt spacer length and decrease as the spacer lengthens or shortens (Fig. 2.5D). The evident correlation between spacer length and robustness of expression explains previous observations that miniclustered 21U RNAs routinely show great variation in abundance (Kato et al., 2009).

We also observed that miniclustered 21U RNAs with 37-40 nt spacers are more abundant than solitary 21U RNAs at matched positions (Fig. 2.5D, asterisks), suggesting that 21U RNA miniclusters may arise when expression is driven more robustly. To investigate this further, we compared the core motifs associated with miniclustered 21U RNAs (“shared” motifs) versus solitary 21U RNAs (“non-shared” motifs). We found that a significantly larger proportion of miniclustered 21U RNAs (3,580 of 4,550, 79%) than solitary 21U RNAs (5,667 of 8,837, 64%) are associated with canonical, GTTTC-containing core motifs (χ^2 test, $p=1.4e-266$). Additionally, we observed significantly greater thymidine richness in the optimal genomic windows 35-42 nt downstream of shared GTTTC-containing motifs versus non-shared (Welch’s *t*-test, $p=4.0e-295$) (Fig. 2.5E). Therefore, particular sequences of 21U RNAs may not be specified intrinsically; rather, core motifs may simply direct expression of 21U RNAs from one or more downstream thymidines, depending on the strength of the

motif and the number of optimally positioned thymidines.

To further confirm the association between the core motif and germline enrichment, we analyzed miniclusters consisting of two germline-enriched 21U RNAs (1,026 pairs). Random assortment of these 21U RNAs would predict 66% male:male, 4% female:female, and 31% male:female pairs; however, we observed 73% male:male, 12% female:female, and only 15% male:female pairs. Thus, 85% of pairs showed matching enrichment classification (Fig. 2.5F), a significant departure from the 69% expected by random assortment (χ^2 test, $p=9.6e-228$). We note that this paucity of mixed male:female 21U RNA miniclusters likely contributes to the low number of 22G RNAs that can be attributed to both male and female 21U RNAs (Fig. S2.3A).

Each upstream motif and 21U RNA sequence constitutes a tiny, autonomous transcriptional unit. The absence of long, unidirectional 21U RNA clusters in the *C. elegans* genome and the presence of the conserved upstream motif have generated speculation that 21U RNAs represent autonomously transcribed units (Ruby et al., 2006; Batista et al., 2008; Das et al., 2008). This is further suggested by our and others' findings that scrambling or deleting the core motif abrogates 21U RNA expression (Fig. 2.3H,I and (Cecere et al., 2012)). To test whether 21U RNAs express independently, we generated transgenes representing putative minimal 21U RNA transcriptional units. Each of these Min transgenes encodes only a single core motif, spacer, and 21U RNA, with limited 5' and 3' genomic context Fig. 2.3A. Strikingly, 21UR-synth expressed from this minimal context shows the same size, *prg-1* dependence, *rde-4* independence, and germline enrichment as endogenous 21U RNAs (Fig. 2.6A,B), indicating that the sequence features conferring these 21U RNA characteristics are contained within a single 21U RNA transcriptional unit. To ensure that the 5' nucleotide of the core motif still influences germline enrichment within this minimal

context, we also generated and tested an independent set of minimal 21UR-synth transgenes with core motif intact (σ Min1415) or first nucleotide toggled (σ MinC>A1415). These transgenes also showed impaired male germline enrichment upon toggling of the core motif 5' nucleotide (Fig. 2.6C), reaffirming our conclusions that a core motif 5' cytidine helps to orchestrate 21U RNA male germline enrichment.

The 21U RNA transcriptional unit is autonomous. To explore the autonomy of the 21U RNA transcriptional unit further, we generated additional transgenes carrying <300 nt of genomic sequence encoding two adjacent 21U RNA transcriptional units on the same strand (Fig. 2.6D). To create the “wild-type” Tg1415-Tg2109 transgene, the upstream 21U RNA locus, corresponding to 21UR-1415, was mutated to encode 21UR-synth, and the downstream locus, corresponding to 21UR-2109, was mutated to encode a different unique synthetic 21U RNA (21UR-synthB). We then scrambled the core motif of the upstream 21U RNA locus to generate the Scram1415-Tg2109 transgene and measured relative expression of the two synthetic 21U RNAs from each transgene. Much as expression of 21UR-synth is vastly decreased by loss of the core motif in the σ Scram1258 and φ Scram2502 transgenes above (Fig. 2.3H,I), 21UR-synth is expressed at far lower levels than 21UR-synthB from the Scram1415-Tg2109 transgene, whereas expression of the synthetic 21U RNAs from the Tg1415-Tg2109 transgene is comparable (Fig. 2.6E). This experiment specifically pursues a recent finding by Cecere et al. (2012) that deletion of the core motif of one 21U RNA does not abrogate expression of neighboring 21U RNAs, although the species assessed were distant, separated by multiple 21U RNA loci, and encoded on both strands.

C. elegans 21U RNA loci, like the piRNA loci of mouse and fly (Aravin et al., 2006; Brennecke et al., 2007; Girard et al., 2006), are genomically clustered. The overwhelming majority of 21U RNAs map to two large regions on chromosome IV, and GTTTC, the most

highly conserved five nt of the core motif, occurs much more frequently on chromosome IV at these regions (4.0 occurrences per kilobase, occ/kb) than on chromosome IV outside these regions (0.4 occ/kb) or on other chromosomes (0.2 occ/kb). Furthermore, 21U RNAs encoded on chromosome IV are detected at much higher abundance (mean abundance: 148 RPM) than those encoded on other chromosomes (1 RPM) (Welch's t -test, $p=2.4e-269$). These observations suggest the possibility of a positional requirement for expression of 21U RNA loci: a privileged genomic environment might contribute to the expression of 21U RNAs. To investigate the significance of 21U RNA genomic organization, we carried out rough mapping of the genomic insertion sites of several of the high-copy transgenic arrays. None of the integration loci mapped to chromosome IV (Table 2.3), indicating that these strains are not expressing 21UR-synth from the context of the 21U RNA genomic clusters. Yet the transgenic arrays themselves could represent 21U RNA-rich genomic microenvironments, much like the chromosome IV 21U RNA clusters. We therefore tested the true autonomy of the 21U RNA by using the MosSCI technique (Frokjaer-Jensen et al., 2008) to insert single-copy transgenes at a locus on chromosome IV not contained within the 21U RNA genomic clusters. Local 21U RNA concentration at the integration site is low, and no 21U RNAs are annotated as mapping to the homology arms encoded on the pCFJ178 MosSCI plasmid. Unexpectedly, single-copy insertions of σ Tg1258 and φ Tg2502 transgenes express 21UR-synth at levels easily detectable, albeit tenfold lower than the high-copy arrays. As observed for the high-copy arrays, scrambling of the core motif severely diminishes expression of 21UR-synth from the single-copy transgenic insertions (Fig. 2.6F). Finally, to exclude the remote possibility that chromosome IV origin itself is essential for 21U RNA expression, we used an alternative MosSCI plasmid to insert onto chromosome II a single copy of the φ Min2502 transgene, which encodes no other 21U RNAs. Like the chromosome IV transgene insertions, φ Min2502 expresses 21UR-synth robustly (Fig. 2.6F), confirming that 21U RNAs can be

autonomously transcribed.

2.3 Materials and methods

Strains

C. elegans were maintained according to standard procedures. The Bristol strain N2 was used as the standard wild-type strain. The alleles used in this study, listed by chromosome, are: unmapped: *xkIs11*[♂ *Scram1258 cb-unc-119(+)*], *xkIs12*[♀ *Scram2502 cb-unc-119(+)*], *xkIs14*[♂ *21U>A1258 cb-unc-119(+)*], *xkIs15*[♂ *21U>G1258 cb-unc-119(+)*], *xkIs16*[♂ *Min1258 cb-unc-119(+)*], *xkIs17*[♀ *21U>A2502 cb-unc-119(+)*], *xkIs18*[♀ *21U>G2502 cb-unc-119(+)*], *xkIs19*[♀ *Min2502 cb-unc-119(+)*], *xkIs20*[♂ *Min1415 cb-unc-119(+)*], *xkIs21*[♂ *MinC>A1415 cb-unc-119(+)*], *xkIs22*[*Tg1415-Tg2109 cb-unc-119(+)*], *xkIs23*[*Scram1415-Tg2109 cb-unc-119(+)*]; LGX: *xkIs10*[♀ *A>C2502 cb-unc-119(+)*]; LGI: *glp-4(bn2)*, *prg-1(tm872)*, *xkIs5*[♀ *Tg2502 cb-unc-119(+)*]; LGII: *xkSi30* [♀ *Min2502 cb-unc-119(+)*], *xkIs6*[♂ *C>A1258 cb-unc-119(+)*]; LGIII: *rde-4(ne301)*, *henn-1(tm4477)*; LGIV: *xkSi3*[♂ *Tg1258 cb-unc-119(+)*], *xkSi23*[♂ *Scram1258 cb-unc-119(+)*], *xkSi17*[♂ *C>A1258 cb-unc-119(+)*], *xkSi13*[♀ *Tg2502 cb-unc-119(+)*], *xkSi28*[♀ *Scram2502 cb-unc-119(+)*], *xkSi20*[♀ *A>C2502 cb-unc-119(+)*], *fem-1(hc17)*, *him-8(e1489)*; LGV: *fog-2(q71)*, *xkIs1*[♂ *Tg1258 cb-unc-119(+)*].

Transgenic allele details and corresponding strain names are shown in [Table 2.3](#).

Sample collection and small RNA analysis

C. elegans samples were generated as previously described (Billi et al., 2012a). Samples for Taqman RT-qPCR validation of 21U RNA germline enrichment classification analysis were collected in biological duplicate. Samples collected for RNA-immunoprecipitation (RIP) analysis were collected in biological duplicate and analyzed in independent experiments with

technical duplicates. All other samples were collected in biological triplicate. All samples analyzed represent adult animals unless otherwise stated.

RNA isolation, beta-elimination, northern blot analysis, Taqman RT-qPCR, and mRNA quantitation were performed as previously described (Billi et al., 2012a). RIP analysis was performed as follows: A custom rabbit polyclonal anti-PRG-1 antibody was generated by Proteintech Group, Inc using an N-terminal peptide antigen (MASGSGRGRGRGSGSNNS (C)) conjugated to keyhole limpet hemocyanin (KLH) carrier protein. Antisera were affinity purified using Affi-Gel 10 gel (Bio-Rad). PRG-1 was purified from synchronized gravid animals using this anti-PRG-1 rabbit polyclonal antibody. For each IP, 10 g of anti-PRG-1 antibody was cross-linked to Dynabeads Protein A (Invitrogen) and incubated with lysate prepared from 0.3 ml of frozen worms at 4°C for 1 hr. Beads were washed 4X with RIP wash buffer (50 mM Tris-HCL pH 7.5, 200 mM KCL and 0.05% NP-40). After final wash, beads were split into equal volumes for RNA extraction and western blot procedure. For western blot analysis: 30 μ l of 1X Tris-glycine SDS sample buffer (Invitrogen) without DTT was added directly to beads and incubated at 50°C for 10 min. 0.1 M DTT was then added to samples and boiled for 5 min before loading on gel. Proteins immobilized on Immobilon-FL transfer membrane (Millipore) were probed with anti-PRG-1 rabbit polyclonal antibody or anti-gamma-tubulin rabbit polyclonal antibody (LL-17) (Sigma) (1:2,000). Peroxidase-AffiniPure goat anti-rabbit IgG secondary antibody was used at 1:10,000 (Jackson ImmunoResearch Laboratories) for detection using Pierce ECL Western Blotting Substrate (Thermo Scientific). For RNA extraction: 1 ml of TRI-Reagent (Ambion) was directly added to beads and incubated at room temperature for 5 min. RNAs were precipitated in isopropanol for 1 hr at -30°C followed by three washes with 70% ethanol.

Small RNA quantitation was performed as previously described (Billi et al., 2012a). All 21U RNA qPCR data from transgenic studies were normalized to miR-1 levels. As a result of

this normalization, some small RNAs whose levels are not detectable (cycle number >36) appear to be detected due to small variation in detection of miR-1. 21UR-synth is not detectable in non-transgenic animals in any stage at which it was assessed. All *Cbr-unc-119* qPCR data were normalized to *act-1* mRNA levels. The sequence of 21UR-synth is 5' TGATATGCGATGTAGTAGACT 3'. The sequence of 21UR-synthB is 5' TTAGTCGTATGTGACGCTGCC 3'. Full small RNA sequences were submitted to Applied Biosystems for design of Taqman assays. Northern blot probe sequences used for this study: miR-1 5' TACATACTTCTTTA-CATTCCA /3StarFire/ 3'; ♂21UR-2502 5' CAGCAGTCTACTACAATTTCA /3StarFire/ 3'; 21UR-synth 5' AGTCTACTACATCGCATATCA /3StarFire/ 3'. RT-qPCR primer sequences used for this study are as follows: *act-1* F 5' CCAGGAATTGCTGATCGTATGCAGAA 3', R 5' TGGAGAGGGAAGCGAGGATAGA 3'; *Cbr-unc-119* F 5' AACGACGTTT-TAGCACTTCCG 3', R 5' GGATTTGGAAGTTGGTGAAGTTCG 3'.

C. elegans transgenesis

To generate the base of the 1258 transgene, sequence spanning genomic coordinates IV:14390835-14393692 was used; IV:14392513-14392673 was used for the ♂Min1258 transgene. To generate the base of the 2502 transgene, sequence spanning genomic coordinates IV:15395699-15397722 was used; IV:15396667-15396886 was used for the ♀Min2502 transgene. To generate the base of the Min1415 transgene, sequence spanning genomic coordinates IV:16564187-16564395 was used. To generate the base of the Tg1415-Tg2109 transgene, sequence spanning genomic coordinates IV:16564133-16564395 was used; the Tg1415-Tg2109 and Scram1415-Tg2109 transgenes carry a 13 nt deletion downstream of both 21U RNA loci. Coordinates were taken from the *C. elegans* genome WS220. The mutations described in [Table 2.3](#) were introduced through site-directed mutagenesis or inverse PCR with phosphorylated primers. Transgenes were then subcloned into the pCFJ178 (IV) or pCFJ151 (II) vector. The chro-

mosome IV transgene insertion site lies outside the larger 21U RNA genomic clusters, and the homology arms of chromosome IV MosSCI vector pCFJ178 do not encode any annotated 21U RNAs. Transgenes were confirmed by sequencing and injected into animals with pharyngeal and/or body wall muscle coinjection markers to distinguish transgenic animals. High-copy arrays were integrated through ultraviolet irradiation. MosSCI single-copy insertions were generated as previously described (Frokjaer-Jensen et al., 2008).

Small RNA sequencing data acquisition and linker removal

Raw data files from 24 small RNA sequencing experiments (Batista et al., 2008; Conine et al., 2010; Gent et al., 2009, 2010; Gu et al., 2009; Han et al., 2009; Kato et al., 2009; Stoeckius et al., 2009) were downloaded from NCBI Gene Expression Omnibus (Barrett et al., 2011). Artificial linker sequences were removed using an in-house linker removal pipeline. We first searched each sequence for a perfect match to the linker. If a perfect match was not found, we searched for an alignment to the linker with 1 mismatch. If not found, we searched for a perfect alignment between the last 5 nt of the sequence and the first 5 nt of the linker. If not found, we repeated this search allowing 1 mismatch. We continued this pattern to align 4 and 3 nt. Sequences with no linker alignment were discarded ($\sim 20\%$ of reads).

Small RNA read alignment to genome and annotation to 21U RNAs

Reads were aligned to the reference *C. elegans* genome version WS220 using Bowtie (Langmead et al., 2009) with the following parameters: `-f -v 2 -k 50 -best -strata`. Mapped read counts in each library were normalized to the number of total mapped reads in that library and to the number of mapped genomic loci. Sequence abundance is reported as reads per million mapped reads (RPM). To determine 21U RNA abundance, we first generated 21U RNA genomic coordinates by aligning 15,703 known 21U RNA sequences (Batista et al.,

2008) to the *C. elegans* genome version WS220 using Bowtie. Perfect, full-length alignments for 15,093 of these sequences were considered valid 21U RNA coordinates. Reads mapping entirely within these coordinates were annotated to 21U RNAs.

Enrichment Score calculations

Germline enrichment classifications of 21U RNAs were generated based on read counts in 17 germline libraries: 14 male germline libraries prepared from isolated spermatogenic cells, isolated spermatids, or whole adult males; and 3 female germline libraries prepared from purified oocytes or whole adult hermaphrodites defective in sperm production (Table 2.1). 1,198 21U RNAs had no read counts in any of these libraries and were removed from our analysis. 184 21U RNAs had higher read counts in a *prg-1(tm872)* young adult library compared to an N2 young adult library (Batista et al., 2008) and were removed from our analysis, leaving 13,711 21U RNAs for which we assessed germline enrichment. Libraries generated using a 5'-monophosphate-dependent (5 male, 1 female) versus -independent (9 male, 2 female) protocol were separated for calculation of the Enrichment Score as follows: For each 21U RNA, we calculated fold abundance difference between every male and female library, for a total of 23 comparisons. Each 21U RNA began with an Enrichment Score of 0. For every comparison, if the 21U RNA was more than 5-fold abundant in the male library, the Enrichment Score decreased by 1; if the 21U RNA was more than 5-fold abundant in the female library, the Enrichment Score increased by 1. Male 21U RNAs were defined as those with Enrichment Scores ≤ -3 , while female 21U RNAs were defined as those with Enrichment Scores ≥ 3 . Remaining 21U RNAs were classified as non-enriched. To validate enrichment classifications, the fold abundance differences for each 21U RNA were averaged across all 23 comparisons. Less than 1% of 21U RNAs classified as male or female do not show enrichment by average fold abundance in their respective libraries. These 21U RNAs

were reclassified as non-enriched for subsequent analyses. 21U RNA Enrichment scores and germline enrichment classifications are in [Data Table S1 \(click to download\)](#).

Determination of false discovery rate

To approximate the number of 21U RNAs falsely classified as male or female germline-enriched by our method, we performed Enrichment Score calculations on randomly generated count data modeled from an N2 young adult library (Batista et al., 2008). 11,458 21U RNAs are represented in this library. Because 17 germline libraries were used for the real analysis, we generated 17 control libraries as follows: For each 21U RNA, 17 random counts were generated from a Poisson distribution with $\lambda=\alpha$ (where α is set to the 21U RNA count in the N2 library) and assigned to one of 17 control libraries. After all counts were assigned, the 17 control libraries were randomly grouped to represent the number of male or female and 5'-monophosphate-dependent or -independent libraries used above. Enrichment Score calculations were then performed on these control libraries as described above, and the number of 21U RNAs classified as germline-enriched was calculated. This protocol was repeated 1,000 times. On the basis of this randomized data, we defined an Enrichment Score threshold of ± 3 , inclusive, for classifying 21U RNAs as male or female germline-enriched, respectively. Application of this threshold to the randomized data resulted in classification of, on average, only 0.76% (101 of 11,458) of 21U RNAs as germline-enriched, corresponding to a false discovery rate below 1%. This value is consistent with the less than 1% of 21U RNAs classified as male or female that do not show enrichment by average fold abundance in their respective libraries.

Enrichment Score calculations performed on 26G RNAs

26G RNA annotations were taken from Han et al. (2009). The abundances of 4,002 26G

RNAs were measured in 13 of the 17 libraries used for 21U RNA Enrichment Score calculations. Four male libraries (GSM465244, GSM503843, GSM459329, and GSM459331) were excluded because the animals used in preparation of the libraries carried mutations in genes required for 26G RNA expression (Conine et al., 2010; Gent et al., 2009, 2010). Enrichment Score calculations were performed on the 13 remaining libraries as above, for a total of 16 male:female comparisons. We retained the Enrichment Score threshold for classifying 26G RNAs as male or female germline-enriched.

Analysis of 21U RNA-dependent 22G RNAs and 21U RNA targets

21U RNA target and 22G RNA information for young adult animals (N2 and *prg-1(n4357)*) was obtained from Bagijn et al. (2012); raw sequencing data files for gravid adult animals (N2 and *prg-1(n4357)*) were downloaded from GEO (Lee et al., 2012). Raw sequences were processed as described above, and reads 22 nt long and starting with guanosine were annotated as 22G RNAs. 21U RNA targets were defined as transcripts with 0-3 mismatches to a 21U RNA sequence. 21U RNA-dependent 22G RNAs were defined as 22G RNAs that map antisense to transcripts within 40 nt of a 21U RNA target site. The number of 22G RNAs that map to both male and female 21U RNA target sites was compared to a control number of 22G RNAs that map to both a random set of male and a random set of female 21U RNA target sites. These random target sites were defined as the target sites of 7,677 randomly selected 21U RNAs representing “male” 21U RNAs and the target sites of 2,171 randomly selected (and not overlapping random male) 21U RNAs to represent “female” 21U RNAs. This random selection was repeated 1,000 times. A similar randomization process was repeated to compare with the number of genes targeted by both male and female 21U RNAs.

Core motif visualization

Core motifs of 21U RNAs were visualized using WebLogo and correcting for *C. elegans* genome nucleotide composition (Crooks et al., 2004). To account for variability in the location of core motifs relative to their 21U RNA loci, upstream regions were aligned by the central 3 Ts of the core motif. If no core motif was identified within 60 nt upstream of a 21U RNA, we aligned position -44 relative to the 21U RNA locus to the G of the core motif, corresponding to the previously identified most common position of the G (Ruby et al., 2006). Only 21U RNAs that map to a single locus in the genome (13,387 of 13,711 21U RNAs, 97.6%) were analyzed since 21U RNAs that map to more than 1 locus may have different upstream sequences.

Identification of genomic features for nucleosome and Pol II occupancy profiling

Nucleosome and Pol II occupancy profiling for 21U RNA loci was centered on the genomic thymidine encoding the 21U RNA 5' uridine. Profiling for transcripts was centered on transcription start sites (TSS) defined as the start of 5'UTRs annotated in the Ensembl66 database (Flicek et al., 2012). Intergenic regions were defined as regions absent of an annotated 5'UTR, exon, intron, 3'UTR or small RNA transcript that were partitioned into randomly distributed, non-overlapping 1,000 nt windows. Profiling for intergenic regions was centered on these 1,000 nt windows.

Young adult TSS expression was calculated as fragments per kilobase per million mapped reads (FPKM) using biological replicates from a transcriptomic sequencing experiment (Hillier et al., 2009). Transcriptome sequence data were removed of linkers and aligned to the *C. elegans* genome version WS220 using TopHat (Trapnell et al., 2009). Cufflinks (Trapnell et al., 2010) was used to calculate transcript isoform expression. Transcripts with an annotated 5'UTR were extracted from the Ensembl66 database. Average transcript FPKM across

the two libraries was calculated, and the isoform with the highest expression was chosen for nucleosome and Pol II occupancy analyses. For isoforms with equivalent expression, a single isoform was randomly chosen.

Analysis of nucleosome and Pol II occupancy

Published nucleosome occupancy data (Valouev et al., 2008) were downloaded from UCSC, and the genomic coordinates were lifted over from WS170 to WS220. Adjusted nucleosome occupancy data centered on 21U RNAs, TSS, and an intergenic background control were averaged for each nucleotide. Pol II ChIP-seq data from young adult worms were downloaded from the modEncode repository (Celniker et al., 2009). Pol II signal to input ratios on chromosome IV were averaged for each nucleotide. TSS were further filtered to only include transcripts with at least 5 FPKM as calculated above.

Pol II ChIP-seq data from young adult worms were downloaded from the modEncode repository. Pol II signal to input ratios were averaged for each nucleotide separately for male and female 21U RNAs on chromosome IV.

2.4 Discussion and concluding remarks

2.4.1 Discussion of findings

piRNAs are transcribed as tiny, autonomous transcriptional units. Our data support a 21U RNA biogenesis mechanism wherein the upstream motif and 21U RNA sequence constitute a tiny, independent transcriptional unit that encodes regulated germline expression. The upstream motif as initially identified by Ruby et al. (2006) is necessary for autonomous expression of a 21U RNA from one or more optimally situated downstream genomic thymidines. Importantly, this genomic thymidine may not represent a transcrip-

tional requirement but rather reflect the binding preferences of the Argonaute PRG-1: a heterogeneous pool of candidate 21U RNA sequences may be transcribed and subsequently purified through preferential stabilization by PRG-1. Our transgenic studies showing greatly decreased expression when 21UR-synth is mutated to 21[U>A/G]R-synth cannot differentiate between a transcriptional or post-transcriptional requirement for a 5' uridine; however, findings in other organisms support the latter mechanism. In mouse and fly, the prevailing model posits that Zucchini generates candidate primary piRNA 5' ends with very little sequence specificity during the processing step, and then Piwi preferentially binds 5' uridine piRNAs during the loading step (Ipsaro et al., 2012; Nishimasu et al., 2012). This is consistent with *in vitro* data showing that Siwi, the silkworm ortholog of PRG-1, preferentially incorporates ssRNAs bearing a 5' uridine (Kawaoka et al., 2011).

On the evidence for transcription of 21U RNAs by RNA polymerase II. The upstream motif differences of male and female 21U RNAs suggest that germline enrichment could be achieved through selective transcription in male versus female germlines. Recently, Cecere et al. (2012) reported that 21U RNA upstream regions are depleted of nucleosomes. They further observed that RNA polymerase II (Pol II) occupancy shows local peaks in this region, rising steadily over the interval of 2300 nt to 250 nt from the genomic thymidine encoding the 5' uridine of the 21U RNA. Analyzing the same ChIP-seq dataset as Cecere et al. (2012), we noticed that the amplitude of the changes in Pol II occupancy at 21U RNA loci is quite modest. Analyzing randomly generated intergenic windows from chromosome IV, we determined that the Pol II ChIP-seq background actually exceeds the “signal” at 21U RNA loci (Fig. S2.7A,B), indicating relative Pol II depletion. This overall depletion of Pol II occupancy at 21U RNA loci may indicate that transcription of 21U RNAs is a more transient process than transcription of genes with canonical promoter elements. Thus the ChIP-seq

might capture only a small fraction of interactions between Pol II and DNA. However, the Pol II occupancy profiles for the loci encoding the top 25% and bottom 25% of 21U RNAs by abundance are virtually indistinguishable (Fig. S2.7C). Again, this is in stark contrast to mRNA coding loci, for which Pol II occupancy at the top 25% of mRNAs by abundance is much higher than at the bottom 25% (Fig. S2.7D). An alternative possibility is that the open chromatin of the nucleosome-depleted regions upstream of 21U RNA loci is more susceptible to incidental binding by Pol II, causing the modest increase in local occupancy observed by Cecere et al. (2012). Should this be the case, the products of Pol II transcription at these loci could be unrelated to 21U RNAs. Cecere et al. (2012) also identify a transcript whose 5' end extends 2 nt upstream of a 21U RNA locus and note that deep sequencing of 5' capped RNAs reveals many more such transcripts. While these transcripts may represent 21U RNA precursors, they may also represent the products of incidental transcription from 21U RNA loci exposed due to local nucleosome depletion. The levels of such long putative precursors were below the threshold of our detection, precluding further study. Nevertheless, the uncertain 5' nucleotide identity of the nascent 21U RNA transcript does not affect the interpretation of our results. Further studies, including identification of a cleavage mechanism for the 2 nt 5' overhang, are needed to confirm these capped transcripts as *bona fide* 21U RNA precursors. The Zucchini endoribonuclease, thought to generate piRNA 5' ends in mouse and fly (Ipsaro et al., 2012; Nishimasu et al., 2012), is not a likely candidate, as it has no obvious homolog in *C. elegans* and shows very little sequence specificity, nor is there any evidence in *C. elegans* for processing of a long 21U RNA precursor into multiple species.

How are the male and female subsets of 21U RNAs differentially expressed?

We show that the 5' nucleotide of the conserved core motif influences germline enrichment of the dependent 21U RNA species (Figures 2.4 and 2.6). This differential expression of male

and female 21U RNAs may be orchestrated by DNA-binding proteins that differ in germline expression patterns and/or binding affinity for 5' cytidine core motifs. Recently, Cecere et al. (2012) demonstrated that the forkhead transcription factors UNC-130, FKH-3, and FKH-5 specifically bind a CTGTTTCA-containing substrate dsDNA probe *in vitro*. However, male and female 21U RNAs do not appear to be differentially sensitive to depletion of these forkhead proteins, nor do 21U RNAs with and without 5' cytidine motifs (data not shown and (Cecere et al., 2012)). Cecere et al. (2012) propose that these forkhead proteins play a redundant role in transcription of 21U RNAs. While these are dispensable for viability and fertility, other forkhead proteins are required for development of the germline, precluding testing for a role in transcribing 21U RNAs; these additional forkhead proteins could indeed represent germline-specific or motif-specific transcription factors (Fig. S2.8).

Why are autonomous 21U RNA transcriptional units genomically clustered?

The autonomy of the *C. elegans* piRNA gene raises the questions of why 21U RNA loci exhibit genomic clustering on chromosome IV and why 21U RNAs encoded on chromosome IV are expressed more robustly. Perhaps the high density of 21U RNAs within these genomic clusters evolved as such: 21U RNA loci, defined by 21U RNA core motifs flanked by A/T richness, accumulated randomly on ancestral chromosome IV. Targeting of any overlapping genes resulted in silencing, subjecting the coding sequences of these genes to drift and eventual elimination. This would deplete the region of genes, reducing selection upon the genomic sequence and thereby permitting further accumulation of 21U RNA loci. The lack of selective pressure related to conservation of protein-coding genes might also explain why chromosome IV loci express 21U RNAs most robustly: the high density of coding and regulatory elements on other chromosomes likely constrains the evolution of features such as flanking A/T-richness that might enhance 21U RNA expression. It is also possible that

different transcriptional machineries or different chromatin configurations are required to transcribe 21U RNAs versus other elements.

Genomic clustering of piRNA loci has been proposed to provide a “trap” for mobile elements (Brennecke et al., 2007). In organisms such as mouse and fly where these clusters are transcribed to generate long precursors from which piRNAs are processed (Aravin et al., 2006; Brennecke et al., 2007; Girard et al., 2006), the trapping function of the genomic piRNA cluster is readily apparent. Although the 21U RNAs are independently transcribed, Bagijn et al. (2012) have identified a similar mechanism acting in *C. elegans*: the genome shows evidence of recent transposon integration downstream of the conserved upstream 21U RNA motif, sometimes generating 21U RNAs that are antisense to the transposon 3' end and capable of silencing it. Each conserved upstream motif can therefore serve as an independent trap, with the result that increased accumulation of motifs enhances protection against mobile elements. While retroelements comprise over 40% of the human genome, they appear to have been strongly counterselected in *C. elegans*, where they constitute only 0.2% of the genome (Bessereau, 2006). Perhaps the autonomous piRNA mechanism at play in *C. elegans* has rendered the animal less susceptible to this kind of mobile element over an evolutionary time scale. Intriguingly, however, *C. elegans* shows significantly higher rates of gene duplication than fly (Lynch and Conery, 2000), and the *C. elegans* genome shows substantial expansions of gene families; for example, the *C. elegans* Argonaute family has expanded to over two dozen members, with the evolution of a worm-specific clade. As gene duplications, like mobile elements, may also be targeted by piRNAs, the preponderance of gene family expansions in *C. elegans* could suggest that this system confers enhanced protection against transposons at the expense of enhanced tolerance for gene duplications. Identification of additional organisms that use similar mechanisms for generating piRNAs will reveal whether this is a pattern or a peculiarity of *C. elegans*.

Additional evidence for transcription of 21U RNAs by Pol II. Gu et al. (2012) recently identified global candidate RNA polymerase II transcription start sites by deep sequencing of capped RNAs. For a large proportion of annotated 21U RNAs, the authors identified 5' capped, ~26 nt putative precursors with a 2 nt 5' overhang. Longer RNA reads (70-90 nt) were identified overlapping a very small minority of 21U RNA loci. Abundance of these longer reads correlated poorly with 21U RNA abundance, while the abundance of the short, ~26 nt reads correlated well, suggesting they are likelier to represent 21U RNA precursors. The 5' cap structure of the putative 21U RNA precursor indeed suggests transcription by Pol II, although our analysis of Pol II occupancy data is inconclusive.

2.4.2 Transcriptional regulation of *C. elegans* 21U RNAs

In this Chapter I describes the computational identification of male and female germline-specific subclasses of 21U RNAs in *C. elegans* (Fig. 2.1) and how these subclasses differentially target germline-specific mRNAs (Fig. S2.3). In parallel work, Lee et al. (2012) and Bagijn et al. (2012) show that 21U RNAs and PRG-1 trigger production of a subclass of endo-siRNAs the 22G RNAs that, in complex with a worm-specific class of Argonaute (WAGO) proteins, directly silence target RNAs. In their work, the authors showed that 21U RNAs and these WAGO 22G RNAs mainly silence transposable elements and pseudogenes. In contrast, a separate subclass of 22G RNAs, which are bound by the CSR-1 Argonaute, target genomic loci expressing germline-specific protein-coding genes to protect them from 21U RNA/WAGO 22G RNA silencing. Thus, 21U RNAs are part of a key germline-specific PTGR mechanism required for maintaining fertility and protecting germline genomes against harmful genetic elements. I co-authored a review of these studies, and additional work regard-

ing 21U RNA-dependent heritable post-transcriptional gene silencing, which was published in *Genome Biology* in 2012 (Billi et al., 2012b).

A major outstanding question in the field of nematode 21U RNAs is how specific 21U RNA sequences are selected for expression in each germline. While this question is rooted in transcriptional mechanisms, the answer is nonetheless relevant to understanding the post-transcriptional regulatory role of 21U RNAs. Work by Kasper et al. (2014) provides evidence for a Myb-like DNA-binding protein, SNPC-4, to bind genomic 21U RNA clusters and be required for 21U RNA expression. The authors are careful, though, to not refer to SNPC-4 as the “transcription factor” required for 21U RNA expression. They do not provide evidence that SNPC-4 preferentially and specifically binds to the 21U RNA upstream motif, leaving open the possibility that an additional factor or factors are responsible for sequence recognition and the direct binding of the 21U RNA upstream motif.

SNPC-4 expression and DNA binding is not restricted to the germline. To explain developmental regulation of SNPC-4 for promoting 21U RNA expression, Kasper et al. (2014) identify the novel 21U RNA biogenesis and germline-specific factor, PRDE-1, as required for germline concentration of SNPC-4 at genomic 21U RNA loci. No distinction was made, however, between SNPC-4 binding upstream of male or female germline-specific 21U RNAs, despite clear evidence in our analysis that the 5'-most nucleotide of the 21U RNA upstream motif orchestrates 21U RNA germline expression. Therefore, the key factor(s) that is (are) required for upstream motif-dependent regulation of 21U RNA expression remain to be identified.

A recent screen for genes required for 21U RNA biogenesis identified 22 candidates, including *snp-4* (Goh et al., 2014). Mutants for some of these candidate genes, termed TOFUs (Twenty-One-u Fouled Ups), were shown to affect pre-21U RNA levels, suggesting that they have direct roles in 21U RNA transcription. Further characterization of these candidates is

needed to determine which, if any, recognize and bind the 21U RNA upstream motif, differentiating between male and female germline expression, to direct autonomous transcription of 21U RNAs.

2.4.3 Investigating PTGR using small RNA-seq

The ability to quantify entire subclasses of small RNAs and identify novel small RNAs through the use of deep sequencing is a great advantage over low-throughput molecular and biochemical experiments. With each published small RNA-seq dataset, additional novel small RNAs are being discovered, many of which are unique to certain cell types or extracellular conditions and some of which function in specific pathogenic mechanisms. The application of small RNA sequencing to single-cell studies could be highly informative of the variability of small RNA expression across cells in the same tissue. For example, one could investigate small RNA levels across cells from solid tumor tissue and nearby non-tumor tissue or across specific cells in the anterior and posterior ends of developing embryos.

In Chapter Two, the strength of the analysis of male and female 21U RNA expression patterns lies in the ability to reuse data generated by other groups for different purposes. Without investing significant amounts of money and time into collecting samples for and generating small RNA sequencing libraries, I was able to screen all 21U RNAs *in silico* to draw conclusions about their germline expression patterns and generate testable hypotheses regarding how expression might be orchestrated. Sequencing data re-use is not uncommon among the small RNA field and has been employed to investigate many areas of small RNA biology including: screening for pathogens among host organisms that might be responsible for emerging infectious diseases (Kumar et al., 2015; Szcześniak et al., 2012); building databases and develop tools for efficient exploration of small RNAs (Yang et al., 2011;

Szcześniak et al., 2012); exploring biogenesis and effector mechanisms of known small RNA subclasses (Wang et al., 2014a; Sakaguchi et al., 2014; Li et al., 2012); and discovering novel small RNA subclasses (Kumar et al., 2014).

2.4.4 Bioinformatics challenges

The major bioinformatic challenge faced in this study was how to robustly compare 21U RNA expression across small RNA-seq datasets generated from 5'-monophosphate-dependent and -independent library preparation protocols. The approach I developed is simple in its interpretation: the abundance of each 21U RNA is compared in all 1-to-1 comparisons of a male germline and female germline library within the same library type to derive a score representing overall enrichment. The problem then becomes what parameters to use as thresholds for (1) determining enrichment (fold-change) in a single 1-to-1 comparison and (2) the overall Enrichment score. To empirically assign values to these thresholds, I generated control small RNA sequencing datasets to match the number of male/female and independent/dependent libraries used in the study. I based count data for these control datasets on the distribution of 21U RNA reads from a single library, in effect, creating a situation where no 21U RNAs should be differentially expressed. Values for the two parameters were then chosen such that <1% of the control 21U RNAs showed differential expression, corresponding to an FDR of 1%. I chose to model these distributions from a wild-type adult *C. elegans* small RNA-seq sample since both male and female germline tissues are present in this sample and >80% of 21U RNAs were robustly quantified. The distribution I chose to model the 21U RNA counts after was a Poisson distribution, which has been used to model sequencing count data (Wang et al., 2010; Marioni et al., 2008). The only parameter supplied to a Poisson distribution is lambda, which is estimated by the mean; therefore, I

set this parameter to the mean 21U RNA counts in the wild-type adult library.

Modeling sequencing count data with a Poisson distribution, however, has been shown to predict smaller variance than is actually observed in the data. In an effort to compensate for this over-dispersion problem, newer methods have modeled count data based on a negative binomial distribution, which has an additional parameter to model the variance (Hashimoto et al., 2014; Anders and Huber, 2010). Indeed, when I fit a Poisson distribution to the control data and test for goodness-of-fit (GOF) with a chi-squared test (using the residual deviance and degrees of freedom) I get $p=0$ indicating that a Poisson model is not a good fit for these data. Fitting a negative binomial distribution to the control data gives a dispersion parameter of 2.08, indicating that the data are, in fact, over-dispersed. A chi-squared GOF test returns a $p=0.1819$, suggesting that negative binomial is a good model for these data. I then generated new control 21U RNA counts under the negative binomial distribution and optimized the two thresholds. Setting the fold-change cutoff to 10 (from 5) and changing the Enrichment score cutoff to 4 (from 3), I calculated differential expression for 4.8% of the control 21U RNAs corresponding to an FDR <5% (Fig. 2.15). Applying these new parameters to the actual 21U RNA count data revealed little change in the overall numbers of male- and female-enriched 21U RNAs. Although individual 21U RNAs might have lost their germline enrichment classification, the general observations still hold true and the experimental assays performed on endogenous and transgenic 21U RNAs based on the original classifications remain valid.

2.5 Data availability

The following published data series were re-used for this study and are available through the NCBI Gene Expression Omnibus: GSE11738, GSE13339, GSE17153, GSE18215, GSE18429,

GSE18729, GSE19414, and GSE20341.

2.6 Acknowledgements

Allison Billi performed a majority of the genetic and molecular biology experiments associated with this study. Sang Young Chun performed the Pol II occupancy analyses. Germano Cecere and Alla Grishok generously shared experimental methods. This manuscript benefited from thoughtful discussions with Sylvia Fischer, Patrick Hu, Sundeep Kalantry, John Moran, and Chi Zhang. Manuscript was written by ACB, MAF, and JKK.

MAF was supported by the National Science Foundation Open Data IGERT Grant 0903629. ACB was supported by the NIH Genetics Training Grant graduate fellowship T32GM007544-34. AMD was supported by the NIH Cellular and Molecular Biology Training Grant T32GM007315. SYC was supported by the National Institutes of Health Bioinformatics Training Grant T32GM070449-06. VK and JKK were supported by grants from the National Institute of General Medical Sciences (NIGMS) R01GM088565 and the Pew Charitable Trusts.

<i>GEO Accessions</i>											
<i>Dataset</i>	<i>Library</i>	<i>Genotype</i>	<i>Developmental Stage</i>	<i>Extraction protocol</i>	<i>Sequencing platform</i>	<i>Raw reads</i>	<i>Mapped reads</i>	<i>%</i>	<i>21U RNA reads</i>	<i>%</i>	<i>Use</i>
GSE20341	GSM510085	N2	mixed-stage embryos	Dep	Illumina	4,661,109	3,721,872	79.8	31,837	0.9	Embryo analysis
	GSM509932/ GSM510089	<i>him-8(e1489)</i>	isolated spermatogenic cells	Dep	Illumina/454	9,596,732	1,312,607	13.7	25,024	1.9	Germline enrichment
	GSM509933/ GSM510090	<i>fer-1(hc1)</i>	purified unfertilized oocytes	Dep	Illumina/454	6,488,731	2,165,341	33.4	47,731	2.2	Germline enrichment
	GSM297742	N2	mixed-stage embryos	Dep	Illumina	2,730,450	2,382,829	87.3	52,072	2.2	Embryo analysis
GSE11738	GSM297751	N2	young adult	Dep	Illumina	3,533,717	3,169,078	89.7	333,587	10.5	Random control, spacer analyses
	GSM297755	<i>prg-1(tm872)</i>	young adult	Dep	Illumina	3,588,293	3,303,711	92.1	2,577	0.1	21U RNA filtering
GSE18215	GSM297753	<i>fog-2(q71)</i>	young adult	Dep	Illumina	3,387,268	2,960,986	87.4	297,715	10.1	Germline enrichment
	GSM455395	<i>fem-1(hc17)</i>	purified oocytes	Indep	Illumina	8,496,639	7,575,752	89.2	53,848	0.7	Germline enrichment
GSE19414	GSM503834	<i>fem-1(hc17)</i>	adult	Indep	Illumina	389,636	369,130	94.7	224	0.1	Germline enrichment
	GSM503842	<i>mut-16(mg461); fem-3(q20)</i>	isolated spermatogenic cells	Dep	Illumina	425,438	399,905	94.0	94,863	23.7	Germline enrichment
	GSM503843	<i>rrf-3(pk1426); fem-3(q20)</i>	isolated spermatogenic cells	Dep	Illumina	650,621	608,159	93.5	46,864	7.7	Germline enrichment
GSE17153	GSM427297	N2	mixed-stage embryos	Dep	Illumina	2,159,213	1,681,110	77.9	6,926	0.4	Embryo analysis
GSE13339	GSM336052	N2	mixed-stage embryos	Dep	Illumina	6,391,734	2,746,387	43.0	29,378	1.1	Embryo analysis
	GSM336086	<i>dpy-28(y1); him-8(e1489)</i>	young adult	Dep	Illumina	3,653,638	1,357,061	37.1	21,778	1.6	Germline enrichment
GSE18729	GSM465244	<i>alg-3(tm1155); alg-4(ok1041); fog-2(q71)</i>	adult	Indep	Illumina	3,216,031	3,003,318	93.4	237,635	7.9	Germline enrichment
	GSM465245	<i>fog-2(q71)</i>	adult	Indep	Illumina	821,513	757,771	92.2	45,592	6.0	Germline enrichment
	GSM465246	<i>fog-2(q71)</i>	adult	Indep	Illumina	2,740,511	2,562,914	93.5	157,991	6.2	Germline enrichment
	GSM465247	<i>fem-3(q20)</i>	isolated spermatids	Indep	Illumina	10,478,418	7,131,378	68.1	256,516	3.6	Germline enrichment
GSE18429	GSM459328	<i>fem-3(q20)</i>	isolated spermatogenic cells	Indep	Illumina	375,816	341,766	90.9	8,029	2.3	Germline enrichment
	GSM459329	<i>rrf-3(pk1426); him-8(e1489)</i>	young adult	Indep	Illumina	1,756,561	1,673,756	95.3	15,702	0.9	Germline enrichment
	GSM459330	<i>him-8(e1489)</i>	young adult	Indep	Illumina	1,709,934	1,614,576	94.4	3,515	0.2	Germline enrichment
	GSM459331	<i>rrf-3(pk1426); him-8(e1489)</i>	young adult	Indep	Illumina	1,492,360	1,366,804	91.6	9,238	0.7	Germline enrichment
	GSM459332	<i>him-8(e1489)</i>	young adult	Indep	Illumina	755,623	695,239	92.0	2,191	0.3	Germline enrichment

Table S2.1: GEO Accessions for datasets and libraries used are listed. Libraries generated using 5'-monophosphate-dependent (Dep) or -independent (Indep) RNA extraction protocols are indicated along with how the library was used in this study ("Use" column).

Library type	Enrichment classification	A vs. C	A vs. G	A vs. T	A vs. N	C vs. G	C vs. T	C vs. N	G vs. T	G vs. N	T vs. N
Male (5'-mP _r -Dep.)	Male	2.4E-13 C	3.5E-01	3.5E-03 A	2.0E-08 A	1.2E-03 C	6.0E-11 C	1.3E-56 C	3.6E-01	2.2E-01	7.7E-01
	Non-enriched	1.9E-01	4.2E-01	2.8E-01	3.3E-05 A	1.5E-01	3.0E-02	2.1E-12 C	9.0E-01	3.1E-01	5.3E-02
Female (5'-mP _r -Dep.)	Female	6.1E-03 A	2.4E-02	9.8E-01	1.4E-01	4.0E-01	1.9E-02	6.8E-02	3.1E-02	9.6E-02	2.2E-01
	Non-enriched	6.0E-03 A	5.0E-01	9.3E-01	4.6E-03 A	5.9E-01	3.6E-02	9.5E-01	4.9E-01	5.7E-01	3.2E-02
Male (5'-mP _r -indep.)	Male	4.06E-09 C	2.48E-01	6.40E-04 A	5.14E-07 A	1.09E-03 C	1.90E-11 C	8.21E-45 C	2.72E-01	3.28E-01	6.79E-01
	Non-enriched	4.04E-03 A	8.76E-01	9.41E-01	1.84E-04 A	2.59E-01	2.99E-02	2.73E-01	8.49E-01	1.33E-01	5.31E-03 T
Female (5'-mP _r -indep.)	Female	1.53E-05 A	2.62E-03 A	4.17E-01	4.44E-04 A	4.20E-01	5.98E-03 T	1.06E-01	1.72E-02	1.18E-01	6.69E-02
	Non-enriched	4.12E-06 A	8.19E-01	8.04E-01	6.90E-04 A	9.71E-02	1.99E-03 T	1.10E-01	9.42E-01	2.40E-01	2.63E-02

Table S2.2: Highlighted are p -values <0.01 . Identity of the 5' nt corresponding to higher 21U RNA abundance is indicated below each significant p -value. All t -tests are two-tailed. Boxed p -values/letters show that male 21U RNAs with 5'-cytidine motifs are more abundant than any other 5'-nt. mPi-indep: monophosphate independent. mPi-dep: monophosphate dependent. Comparisons of abundances in 5' -monophosphate dependent and -independent libraries were performed separately.

High-copy transgenes								
Transgene name	Transgene description	Size (kb)	Upstream motif	1 st nt	Chr	Allele	Strain	Backgrounds
ϕTg1258	Cluster with synthetic 21U RNA replacing 21UR-1258	2.87	attaagc <u>CTGTTTC</u> acattttt	U	V	<i>xkls1</i>	QK7	WT, <i>prg-1, rde-4, glp-4, him-8, fem-1, henn-1</i>
ϕScram1258	ϕTg1258 with upstream motif scrambled	2.87	attattcagaccgctattttt	U	N/D	<i>xkls11</i>	QK8	WT
ϕC>A1258	ϕTg1258 with upstream motif first nt mutated C>A	2.87	attaagc <u>ATGTTTC</u> acattttt	U	II	<i>xkls6</i>	QK9	WT, <i>prg-1, him-8, fem-1</i>
ϕ21U>A1258	ϕTg1258 with synthetic 21U RNA first nt mutated U>A	2.87	attaagc <u>CTGTTTC</u> acattttt	A	N/D	<i>xkls14</i>	QK10	WT
ϕ21U>G1258	ϕTg1258 with synthetic 21U RNA first nt mutated U>G	2.87	attaagc <u>CTGTTTC</u> acattttt	G	N/D	<i>xkls15</i>	QK11	WT
ϕMin1258	Minimal 21U RNA construct from ϕTg1258	0.17	attaagc <u>CTGTTTC</u> acattttt	U	N/D	<i>xkls16</i>	QK12	WT, <i>prg-1, rde-4, glp-4</i>
ϕTg2502	Cluster with synthetic 21U RNA replacing 21UR-2502	2.04	aaataaa <u>ATGTTTC</u> Aactagtc	U	I	<i>xkls5</i>	QK13	WT, <i>prg-1, rde-4, glp-4, him-8, fem-1, henn-1</i>
ϕScram2502	ϕTg2502 with upstream motif scrambled	2.04	aaataaa <u>ggacacttattattc</u>	U	N/D	<i>xkls12</i>	QK14	WT
ϕA>C2502	ϕTg2502 with upstream motif first nt mutated C>A	2.04	aaataaa <u>CTGTTTC</u> Aactagtc	U	X	<i>xkls10</i>	QK15	WT, <i>prg-1, him-8, fem-1</i>
ϕ21U>A2502	ϕTg2502 with synthetic 21U RNA first nt mutated U>A	2.04	aaataaa <u>ATGTTTC</u> Aactagtc	A	N/D	<i>xkls17</i>	QK16	WT
ϕ21U>G2502	ϕTg2502 with synthetic 21U RNA first nt mutated U>G	2.04	aaataaa <u>ATGTTTC</u> Aactagtc	G	N/D	<i>xkls18</i>	QK17	WT
ϕMin2502	Minimal 21U RNA construct from ϕTg2502	0.23	aaataaa <u>ATGTTTC</u> Aactagtc	U	N/D	<i>xkls19</i>	QK18	WT, <i>prg-1, rde-4, glp-4</i>
ϕMin1415	Minimal 21U RNA construct generated from 21UR-1415	0.22	ttttcgc <u>CTGTTTC</u> Aaggagtt	U	N/D	<i>xkls20</i>	QK19	WT
ϕMinC>A1415	ϕMin1415 with upstream motif first nt mutated C>A	0.22	ttttcgc <u>ATGTTTC</u> Aaggagtt	U	N/D	<i>xkls21</i>	QK20	WT
Tg1415–Tg2109	Minimal 21U RNA construct with 21UR-synth replacing 21UR-1415 and 21UR-synthB replacing 21UR-2109	0.26	ttttcgc <u>CTGTTTC</u> Aaggagtt taatctc <u>CTGTTTC</u> acaatatt	U	N/D	<i>xkls22</i>	QK21	WT
Scram1415–Tg2109	Tg1415–Tg2109 with 21UR-1415 upstream motif scrambled	0.26	ttttcgtaggt <u>taccctgtagtt</u> taatctc <u>CTGTTTC</u> acaatatt	U	N/D	<i>xkls23</i>	QK22	WT
MosSCI transgenes								
Transgene name	Transgene description	Size (kb)	Upstream motif	1 st nt	Chr	Allele	Strain	
ϕTg1258	Cluster with synthetic 21U RNA replacing 21UR-1258	2.87	attaagc <u>CTGTTTC</u> acattttt	U	IV	<i>xkSi3</i>	QK23	
ϕScram1258	ϕTg1258 with upstream motif scrambled	2.87	attattcagaccgctattttt	U	IV	<i>xkSi23</i>	QK24	
ϕC>A1258	ϕTg1258 with upstream motif first nt mutated C>A	2.87	attaagc <u>ATGTTTC</u> acattttt	U	IV	<i>xkSi17</i>	QK25	
ϕTg2502	Cluster with synthetic 21U RNA replacing 21UR-2502	2.04	aaataaa <u>ATGTTTC</u> Aactagtc	U	IV	<i>xkSi13</i>	QK26	
ϕScram2502	ϕTg2502 with upstream motif scrambled	2.04	aaataaa <u>ggacacttattattc</u>	U	IV	<i>xkSi28</i>	QK27	
ϕA>C2502	ϕTg2502 with upstream motif first nt mutated C>A	2.04	aaataaa <u>CTGTTTC</u> Aactagtc	U	IV	<i>xkSi20</i>	QK28	
ϕMin2502	Minimal 21U RNA construct from ϕTg2502	0.23	aaataaa <u>ATGTTTC</u> Aactagtc	U	II	<i>xkSi30</i>	QK29	

Table S2.3: Both high-copy and MosSCI transgenes used in this study are listed with a short description, sequence characteristics, integration information, and strain notation. Full transgene data are listed in Materials and Methods. Bolded letters indicate mutated nucleotides. Eight nt core upstream motifs are capitalized while motif positions are underlined. N/D, not determined.

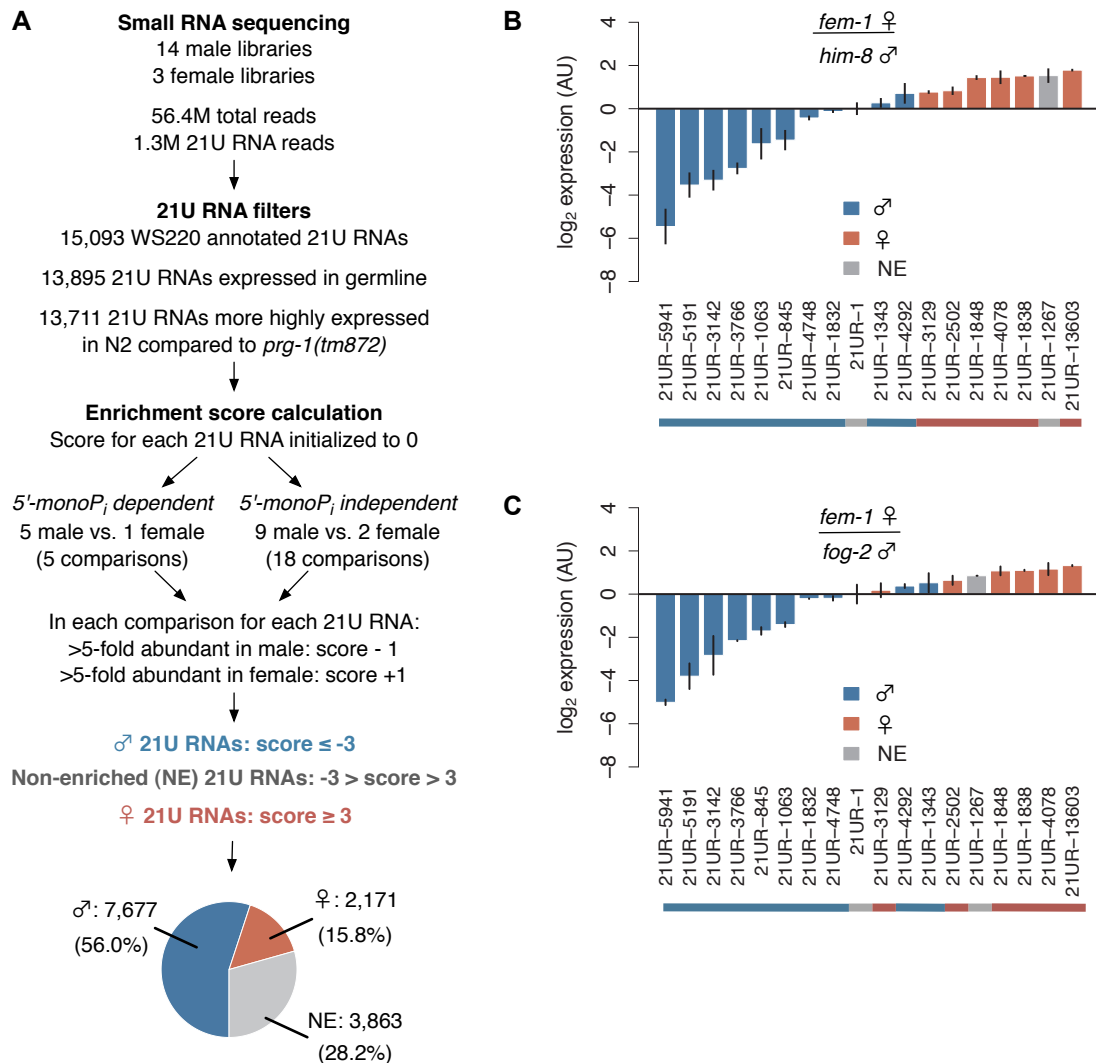


Figure 2.1: (A) Pipeline for computational identification of male and female 21U RNAs. A majority of 21U RNAs are classified as male or female germline-enriched. Pie chart depicts classification as proportion of 13,711 21U RNAs analyzed. (B,C) Male 21U RNAs are more highly expressed in male animals, and female 21U RNAs are more highly expressed in female animals. Relative expression of representative 21U RNAs was assayed by Taqman RT-qPCR in *him-8(e1489)* (B) and *fog-2(q71)* (C) male versus *fem-1(hc17)* female animals and normalized to non-enriched 21U RNA 21UR-1. Error bars: ± 1 standard deviation (SD) of two biological replicates. AU: arbitrary units.

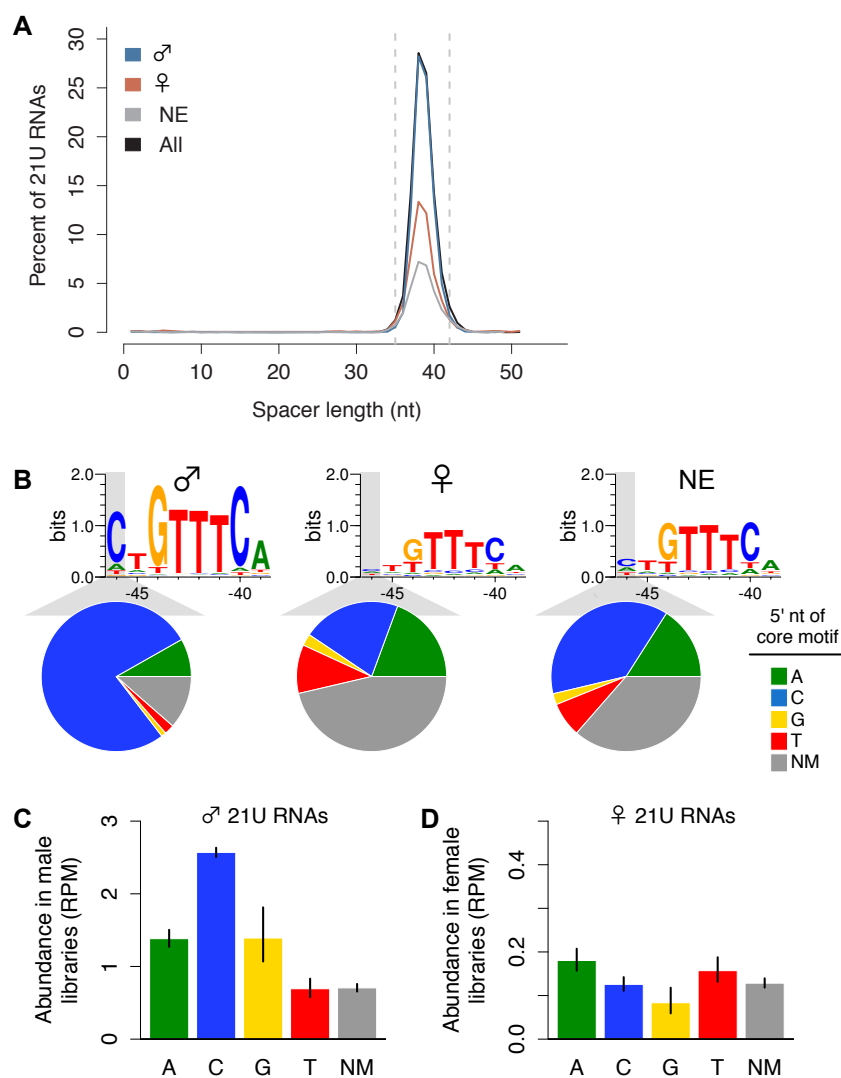


Figure 2.2: (A) Spacer lengths follow expected distribution for all enrichment classifications. Dotted lines: canonical spacer length range (35-42 nt). (B) Male, but not female, 21U RNA coding loci show enrichment for core motifs with 5' cytidines. Significantly fewer female 21U RNAs exhibit a GTTTC-containing core motif than male. Top: Weblogo plots illustrate core motif differences. Bottom: Pie charts depict proportions of 21U RNAs with GTTTC-containing core motifs indicating the 5' nt (colors) or with no GTTTC-containing core motif (NM, no motif, dark grey). (C) Core motif variations correlate with male 21U RNA abundance in 5'-monophosphate-dependent libraries. Average 21U RNA abundance was calculated based on the 5' nt of the core motif. Error bars: ± 1 standard error of the mean (SEM). (D) Core motif variations do not correlate with female 21U RNA abundance in 5'-monophosphate-dependent libraries. Average 21U RNA abundance was calculated as in (C).

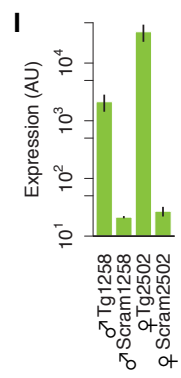
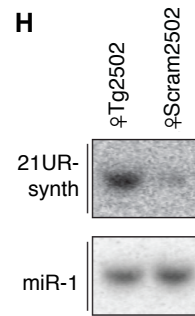
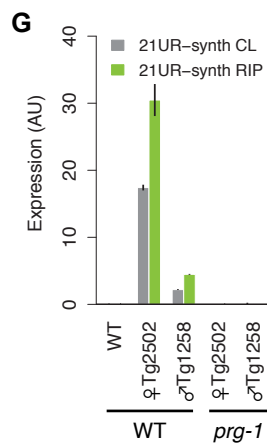
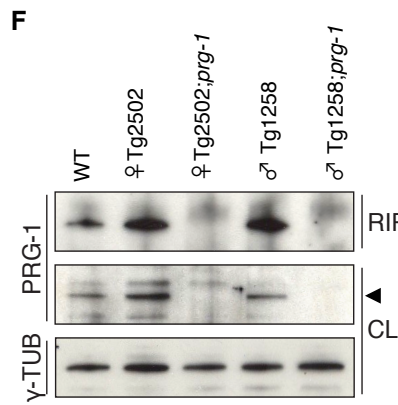
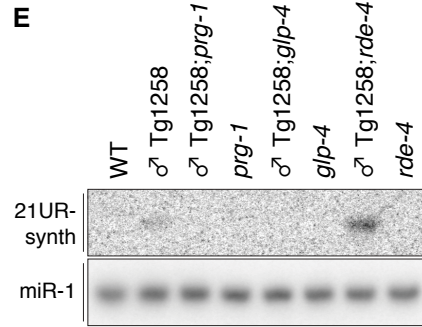
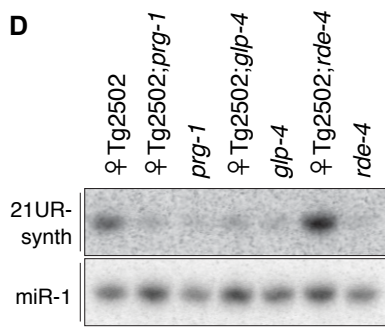
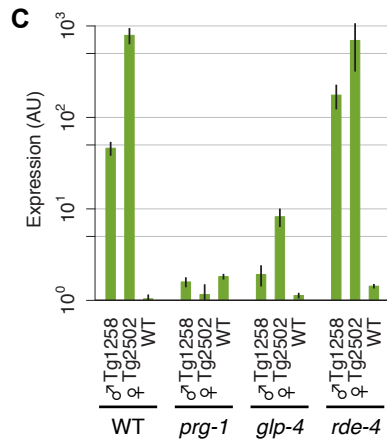
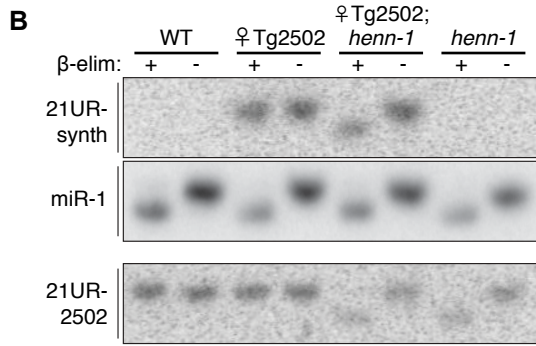
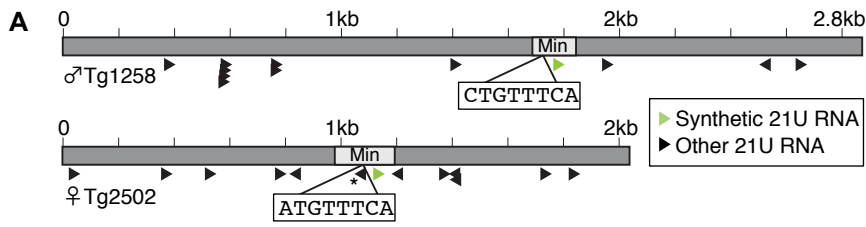


Figure 2.3: (A) Diagram of Tg (dark grey) and Min (light grey) transgenes with core motif sequences shown. Asterisk indicates a 21U RNA whose core motif is disrupted by 21UR-synth and is therefore predicted not to express. (B) 21UR-synth is methylated by HENN-1. 21UR-synth is specifically detected in transgenic strains and is susceptible to β -elimination only in the *henn-1(tm4477)* background. Arrowhead represents migration of a 21 nt size marker. 21UR-synth blot was reprobed for miR-1. Endogenous ϕ 21UR-2502 is shown as a control. (C-E) 21UR-synth is a *prg-1*-dependent, germline-enriched 21U RNA. 21UR-synth detection by Taqman RT-qPCR (C) and northern blot (D,E) is greatly decreased in *prg-1(tm872)* and *glp-4(bn2)* germline-deficient mutant animals, but intact in *rde-4(ne301)* mutant animals. Error bars: ± 1 SD of three biological replicates. (F) anti-PRG-1 antibody immunopurifies PRG-1 complexes. CL: crude lysate, RIP: RNA immunoprecipitation. (G) 21UR-synth is bound by endogenous PRG-1. Error bars: ± 1 SD of two technical replicates; data are representative of two independent experiments. (H,I) Loss of the core motif dramatically decreases 21UR-synth expression by northern blot (H) and Taqman qRT-PCR (I). Error bars: ± 1 SD of three biological replicates.

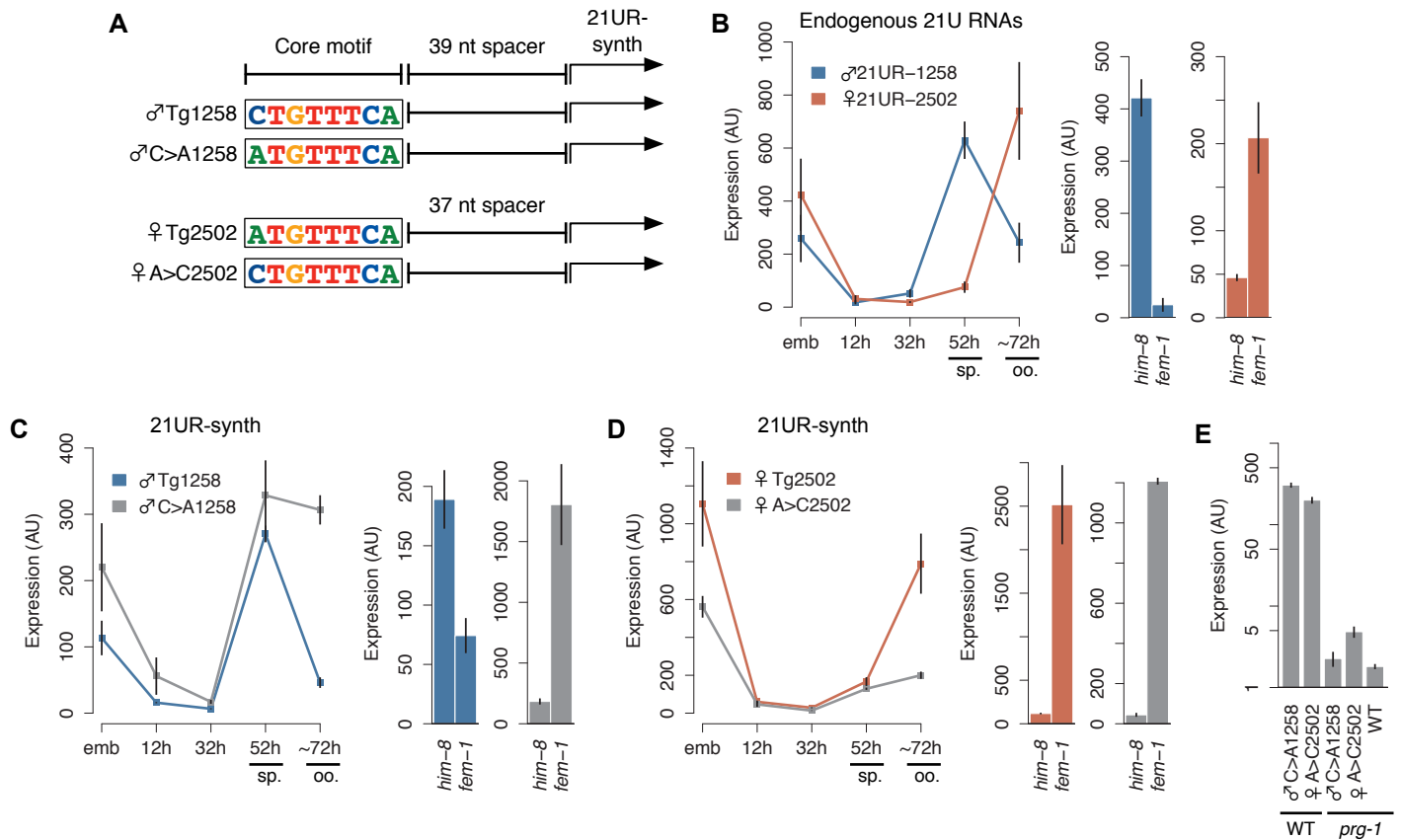


Figure 2.4: (A) Schematic of transgenes with 5' nt of core motif mutated. (B) Left: Endogenous σ 21UR-1258 and ϕ 21UR-2502 peak during spermatogenesis (sp.) and oogenesis (oo.), respectively. Right: Germline enrichment patterns are recapitulated in *him-8(e1489)* male and *fem-1(hc17)* female animals. Error bars: \pm 1 SD of three biological replicates. (C) The male expression pattern of 21UR-synth from σ Tg1258 is disrupted by core motif mutation in σ C>A1258. Error bars: \pm 1 SD of three biological replicates. (D) The female expression pattern of 21UR-synth from ϕ Tg2502 is disrupted by core motif mutation in ϕ A>C2502, but expression in *fem-1(hc17)* female is not lost. Error bars: \pm 1 SD of three biological replicates. (E) Mutating the 5' nt of the core motif does not affect 21UR-synth *prg-1* dependence.

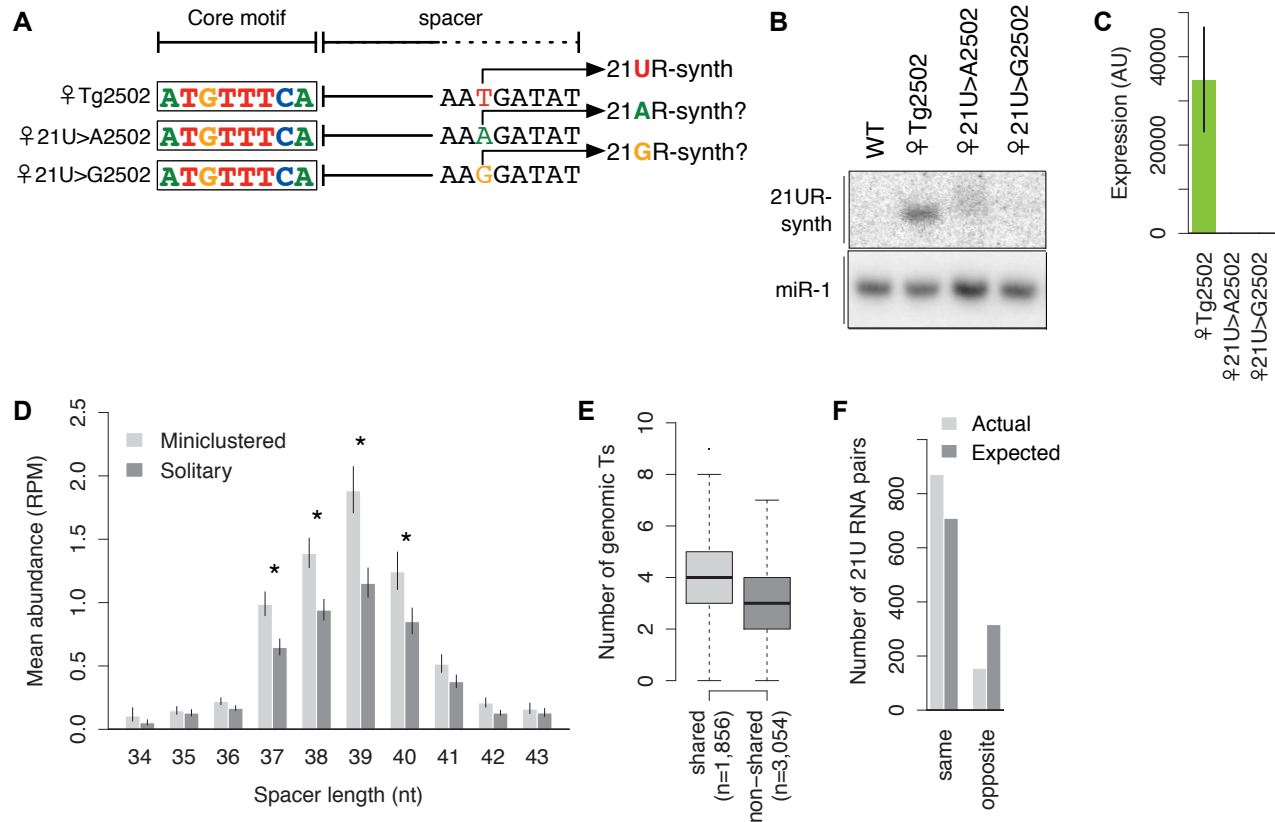


Figure 2.5: (A) Schematic of transgenes with 5' nt of 21U RNA mutated. (B-C) Mutation of the 5' genomic thymidine disrupts expression of 21UR-synth by northern blot (B) and Taqman assay (C). (D) 21U RNA abundances correlate with distances downstream of core motifs. Miniclustered 21U RNAs with 37-40 nt spacer lengths are more abundant than solitary 21U RNAs. Asterisks indicate Welch's *t*-tests, $p < 0.05$. Error bars: ± 1 SEM. (E) Optimal downstream windows are more thymidine-rich for shared core motifs than non-shared (Welch's *t*-test, $p = 2.5 \times 10^{-46}$). The number of genomic thymidines located 35-42 nt downstream of each GTTTC-containing motif was counted. (F) 21U RNA miniclusters are significantly biased for being composed of 21U RNAs with the same, as opposed to opposite, germline enrichment than expected if the same 21U RNAs were randomly paired.

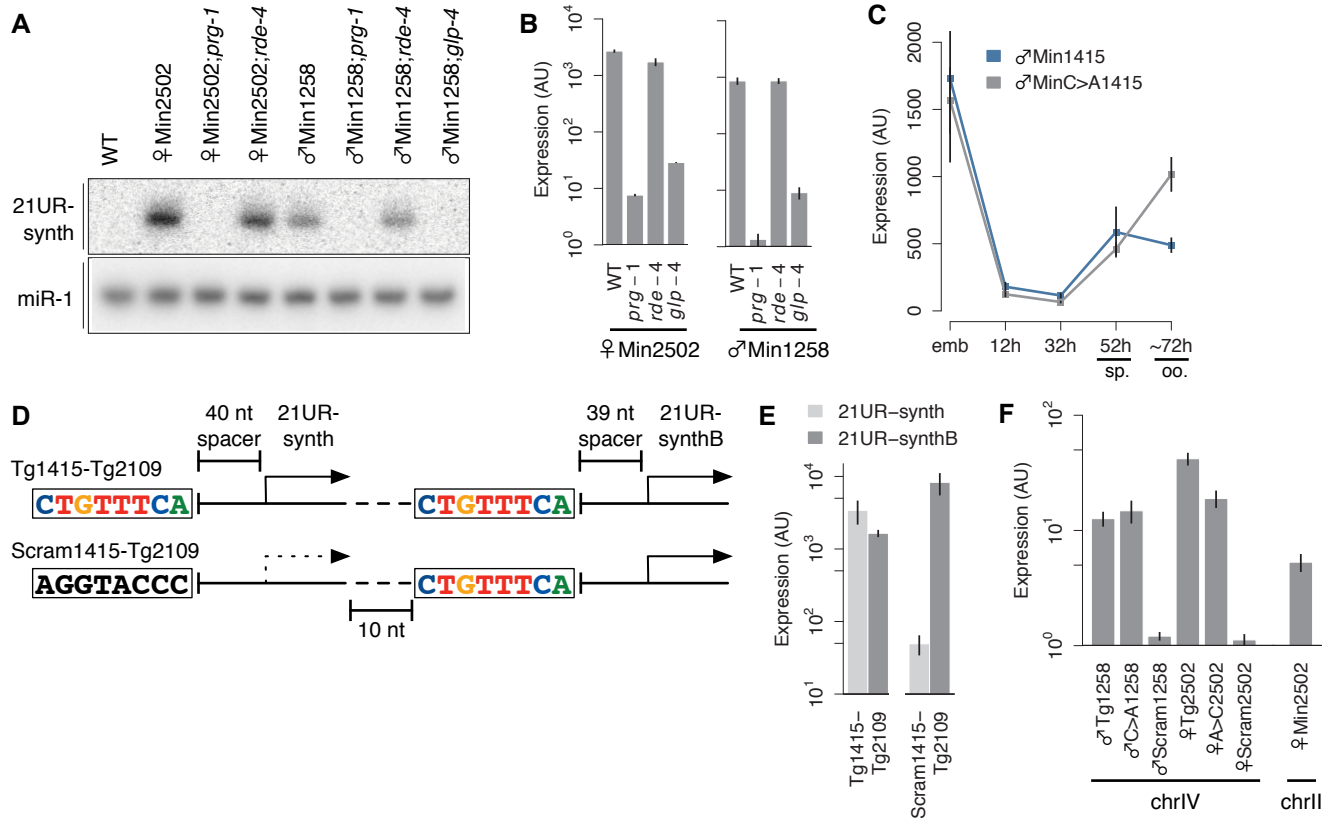


Figure 2.6: (A-B) 21UR-synth expressed from a minimal transcriptional cassette shows *prg-1* dependence, *rde-4* independence, and germline enrichment by northern blot (A) and Taqman assay (B). (C) The male expression pattern of 21UR-synth from ♂Min1415 is disrupted by core motif mutation in ♂MinC>A1415. Error bars: ± 1 SD of three biological replicates. (D) Schematic of transgenes encoding two closely adjacent 21U RNAs. (E) Scrambling the core motif upstream of 21UR-synth abrogates 21UR-synth, but not 21UR-synthB, expression levels. (F) WT and swap transgenes, but not scam transgenes, express from single copy MosSCI on chrIV. Minimal transgenes also expresses from MosSCI on chrII.

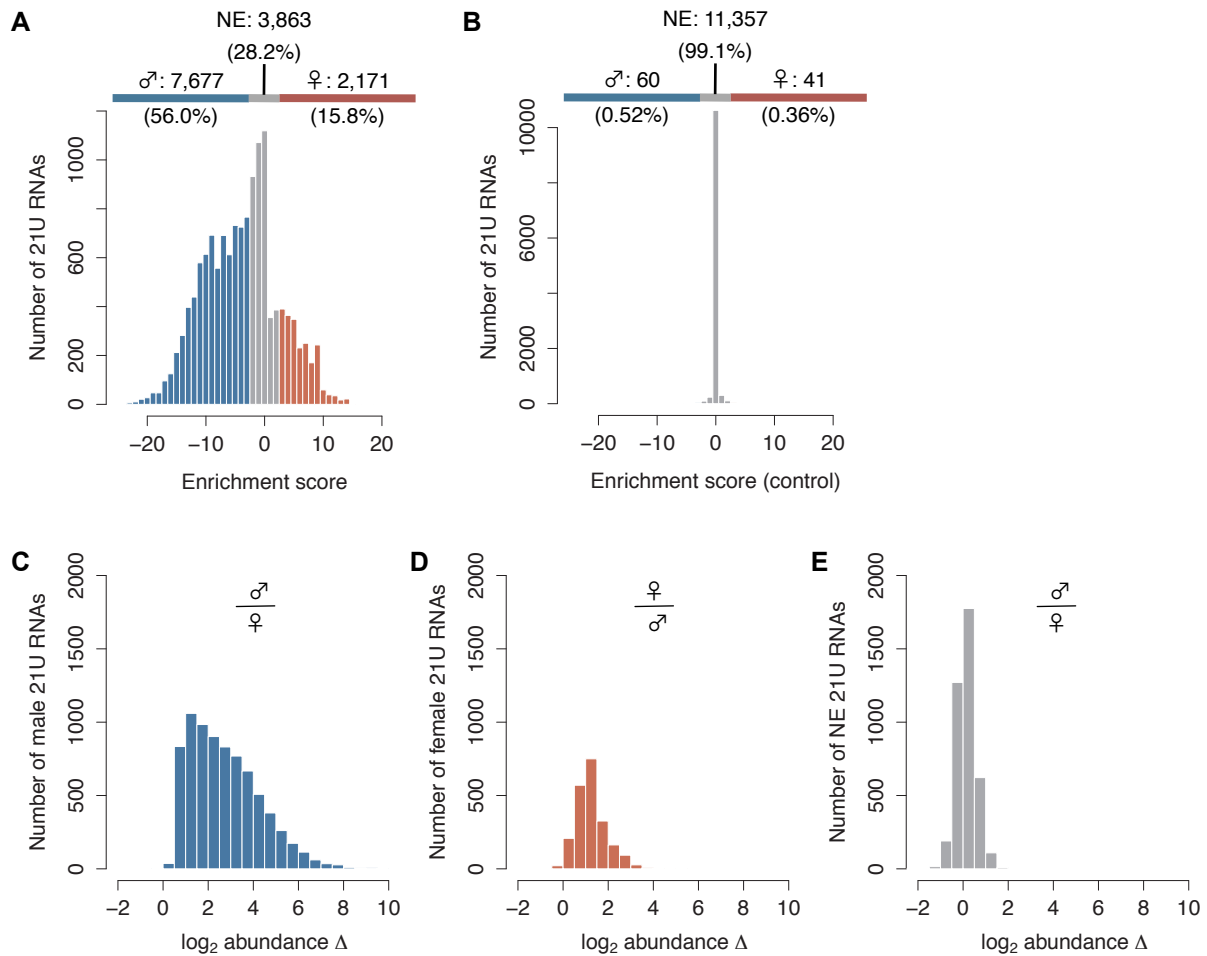


Figure S2.1: (A) Enrichment Score calculations performed on 17 small RNA sequencing libraries classify a majority of 21U RNAs as male (blue) or female (red) germline-enriched. Non-enriched (NE) 21U RNAs, grey. Numbers indicate percent of 13,711 21U RNAs analyzed. (B) Enrichment Score calculations performed on control data classify <1% of 21U RNAs as male or female germline-enriched indicating a 1% false discovery rate. Numbers indicate percent of 11,458 21U RNAs analyzed. (C) Male 21U RNAs are more abundant in male libraries. Average relative abundance of each male 21U RNA was calculated between each of the 23 male:female library comparisons. (D) Female 21U RNAs are more abundant in female libraries. (E) Non-enriched 21U RNAs are equally abundant in male and female libraries.

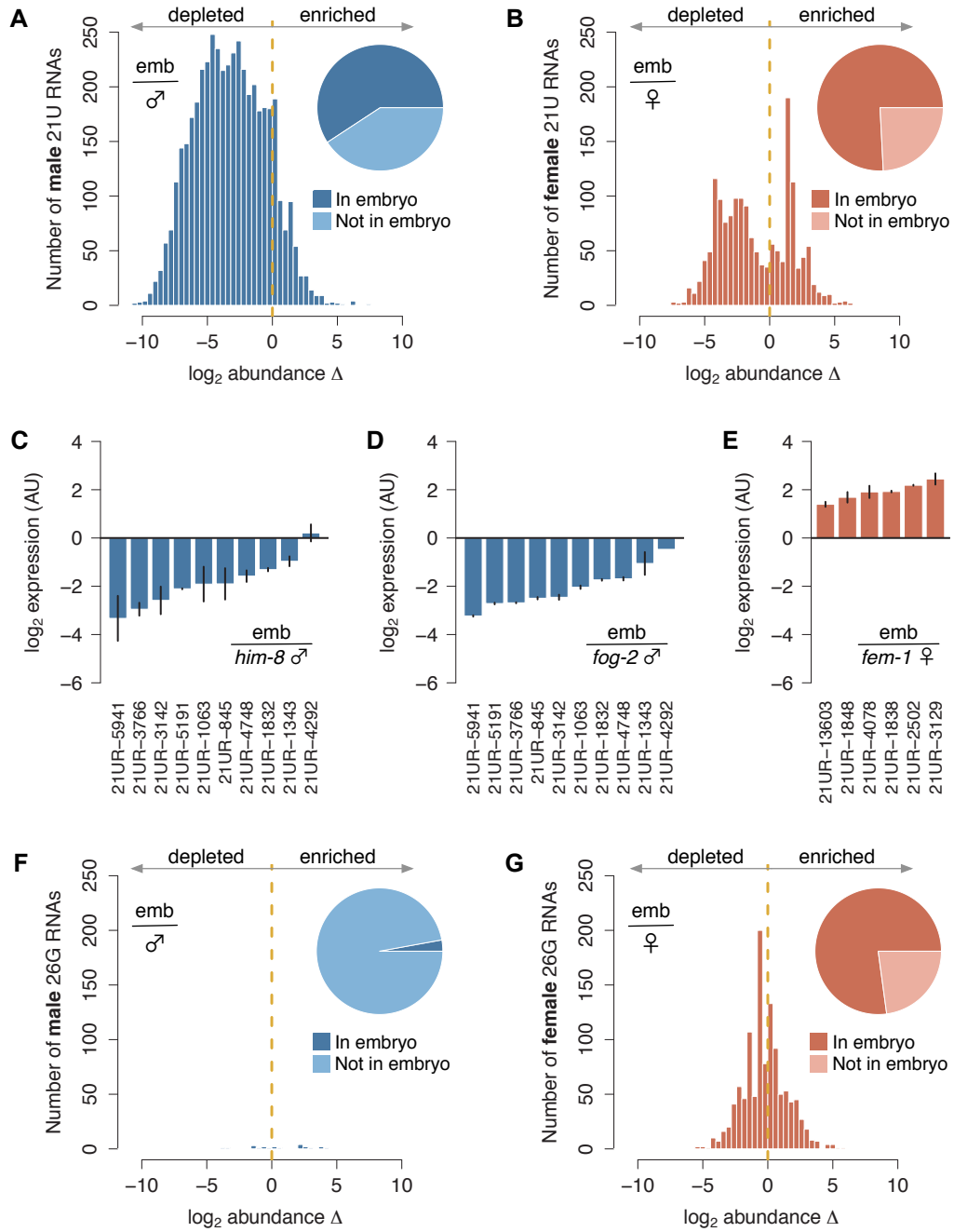


Figure S2.2: (A) Relative male 21U RNA abundance is decreased in embryo. Average relative abundance of each male 21U RNA was calculated between each of 5 male and 4 mixed stage embryo libraries. Dotted line indicates equal male and embryo abundance. Pie chart depicts proportion of male 21U RNAs with reads in at least one embryo library (dark blue). (B) A population of female 21U RNAs shows increased abundance in embryo. Average relative abundance of each female 21U RNA was calculated between each of 1 female and 4 mixed stage embryo libraries. Pie chart depicts proportion of female 21U RNAs with reads in at least one embryo library (dark red). (C,D) Taqman RT-qPCR analysis corroborates male 21U RNA depletion in embryo. Expression of representative male 21U RNAs was assayed by Taqman in *him-8(e1489)* (E) and *fog-2(q71)* (F) male animals and N2 embryos. Error bars represent ± 1 SD from two biological replicates. (E) Taqman RT-qPCR analysis corroborates female 21U RNA enrichment in embryo. Expression of representative female 21U RNAs was assayed by Taqman in *fem-1(hc17)* female animals and N2 embryos. (F) Male germline-enriched 26G RNAs are generally absent in embryo. Average relative abundance of each male 26G RNA was calculated between each of 4 male and 4 mixed stage embryo libraries. (G) Female germline-enriched 26G RNAs are robustly expressed in embryo. Average relative abundance of each female 26G RNA was calculated between each of 1 female and 4 mixed stage embryo libraries.

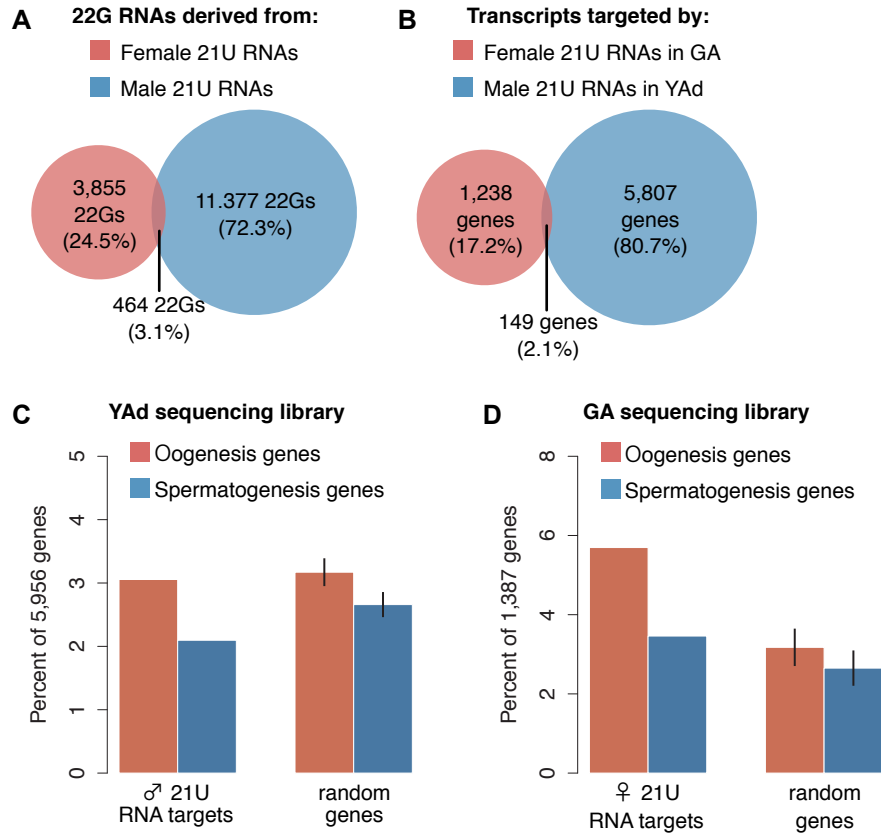


Figure S2.3: (A) 22G RNAs are almost exclusively derived from either male or female 21U RNAs, but not both. The number of unique 22G RNAs derived from both male and female 21U RNAs is significantly less than expected if 22G RNAs are selected at random (Fisher's exact test, $p=1.2e-02$). (B) Male and female 21U RNAs target significantly fewer overlapping genes compared to selecting random sets of genes (Fisher's exact test, $p=7.7e-13$). (C) 5,956 genes targeted by male 21U RNAs in young adult animals are depleted of spermatogenesis genes compared to a random set of 5,956 genes. (D) 1,387 genes targeted by female 21U RNAs in gravid adult animals are enriched for oogenesis genes compared to a random set of 1,387 genes.

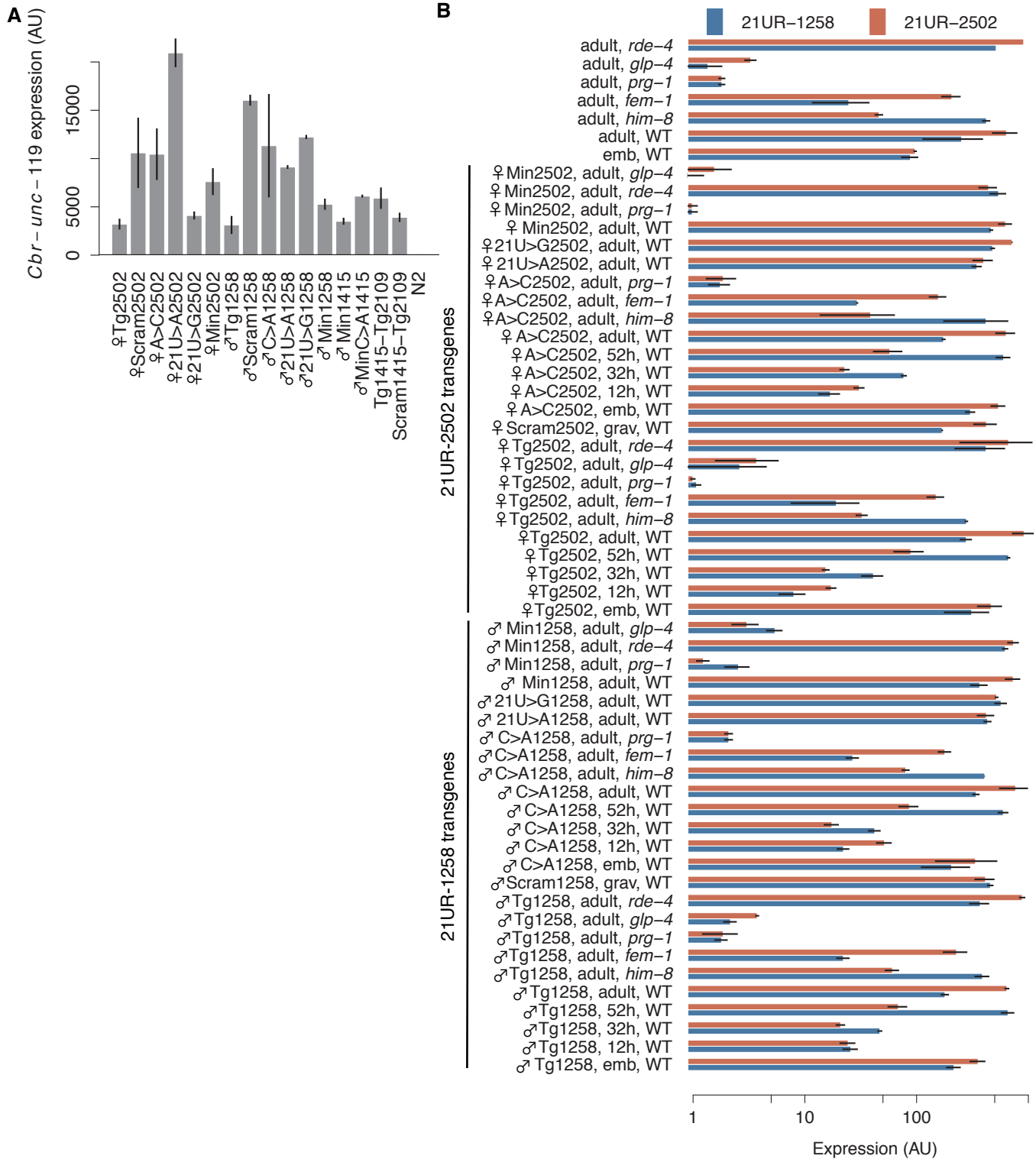


Figure S2.4: (A) Levels of *Cbr-unc-119* mRNA in adult animals were assayed by RT-qPCR for all transgenes and normalized to *act-1* mRNA levels. (B) Expression of transgenic 21UR-synth does not affect expression of endogenous 21U RNA counterparts. Endogenous ♂21UR-1258 and ♀21UR-2502 levels were assayed by Taqman RT-qPCR and normalized to microRNA miR-1 levels in all samples.

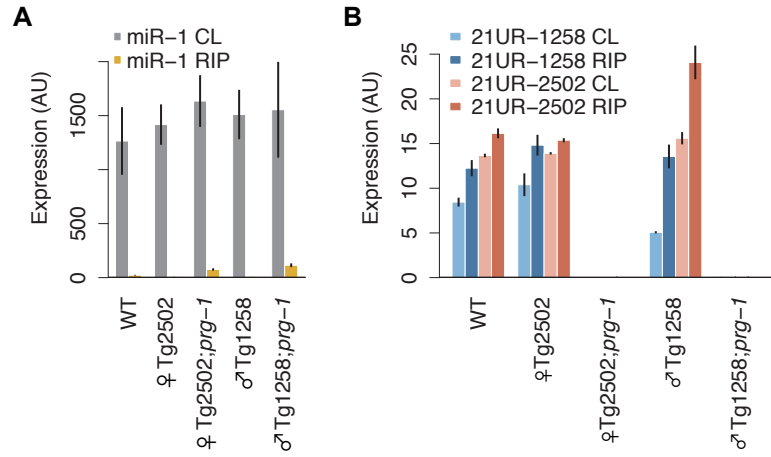


Figure S2.5: (A) anti-PRG-1 antibody does not immunoprecipitate microRNA miR-1. (B) 21UR-synth expression does not interfere with association of endogenous 21U RNAs with PRG-1.

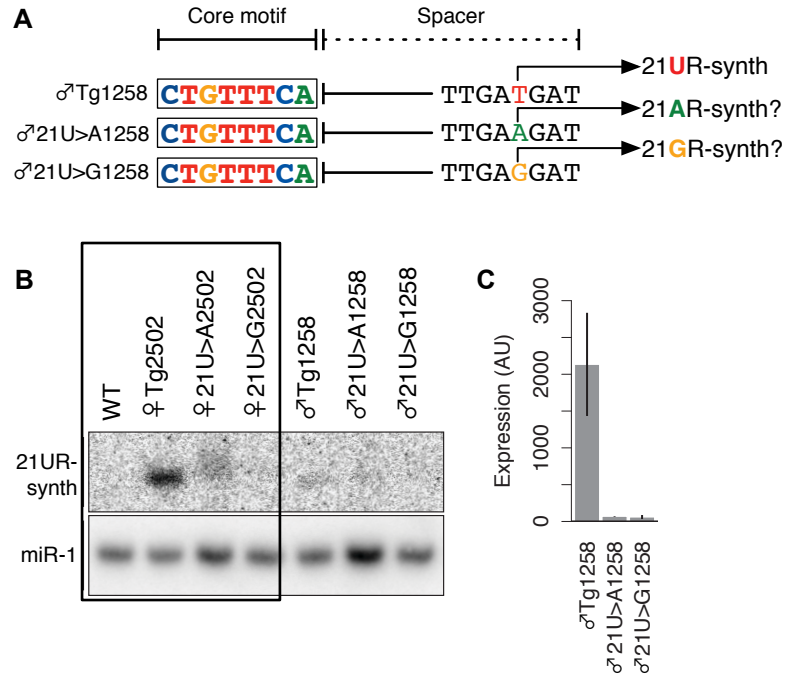


Figure S2.6: (A) Schematic of transgenes. (B-C) Mutation of the 5' genomic thymidine disrupts expression of 21UR-synth by northern blot (B) and Taqman assay (C). WT and ♀Tg2502 lanes in (B) are repeated from Figure 5B for clarity.

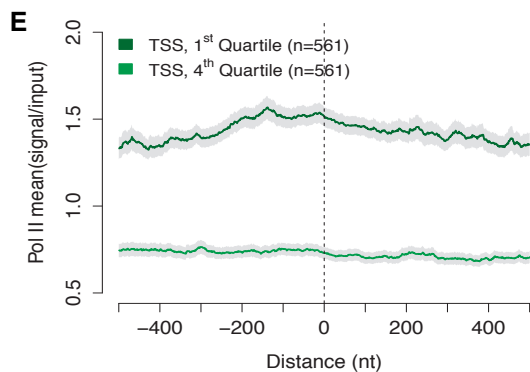
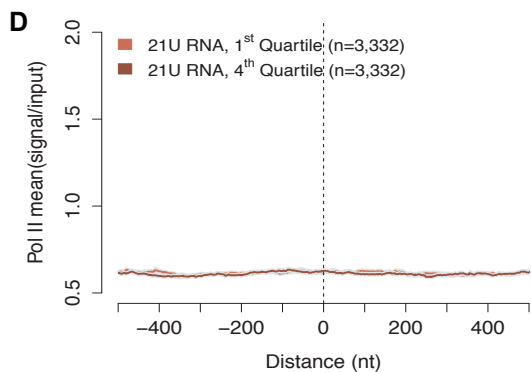
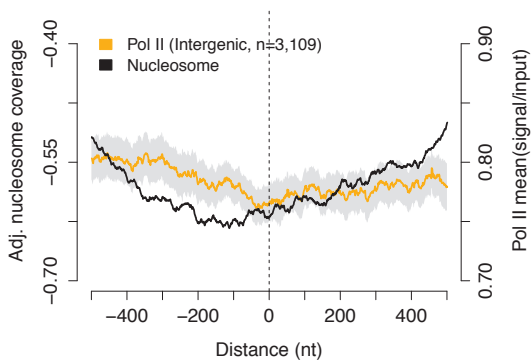
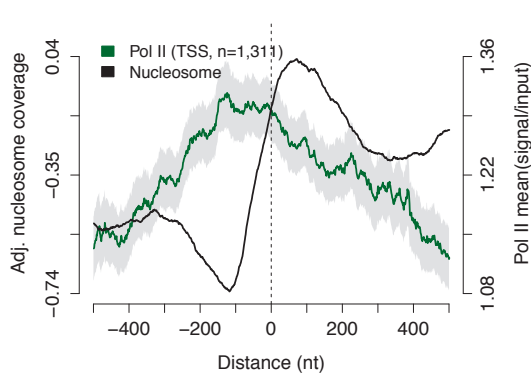
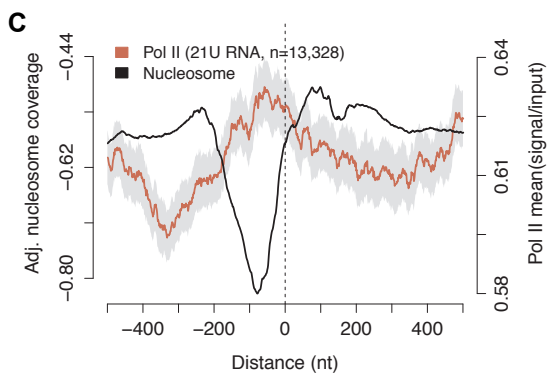
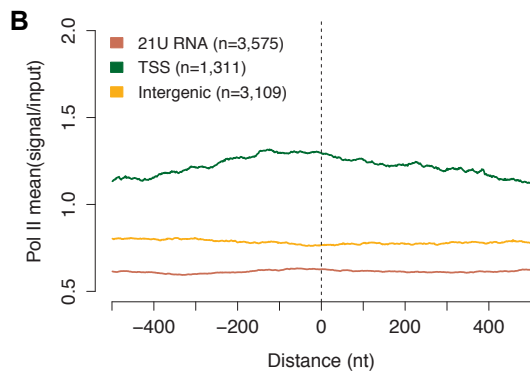
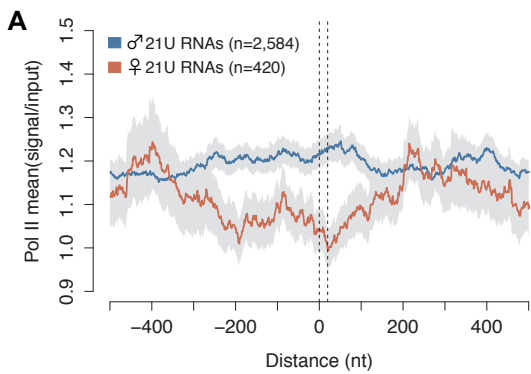


Figure S2.7: (A) Average Pol II occupancy over male and female 21U RNA loci expression 21U RNAs with at least 5 RPM in young adult libraries. Only 21U RNAs on chrIV were assayed. Grey error bands: SEM. (B) Average Pol II occupancy of 21U RNA loci expressing 21U RNAs with at least 5 RPM (red), transcriptional start sites (TSS) expressing transcripts with at least 5 FPKM (green), and randomized intergenic regions (yellow). Only regions on chrIV were assayed. (C) Pol II occupancy as described in (B) but independently scaled for each transcript type and plotted with average nucleosome occupancy (black line). (D) Average Pol II occupancy of 21U RNA loci as (C) but showing the top 25% 21U RNAs by abundance (1st quartile) and the bottom 25% (4th quartile) separately. (E) Same as (D) but showing top and bottom 25% of TSS by transcript abundance.

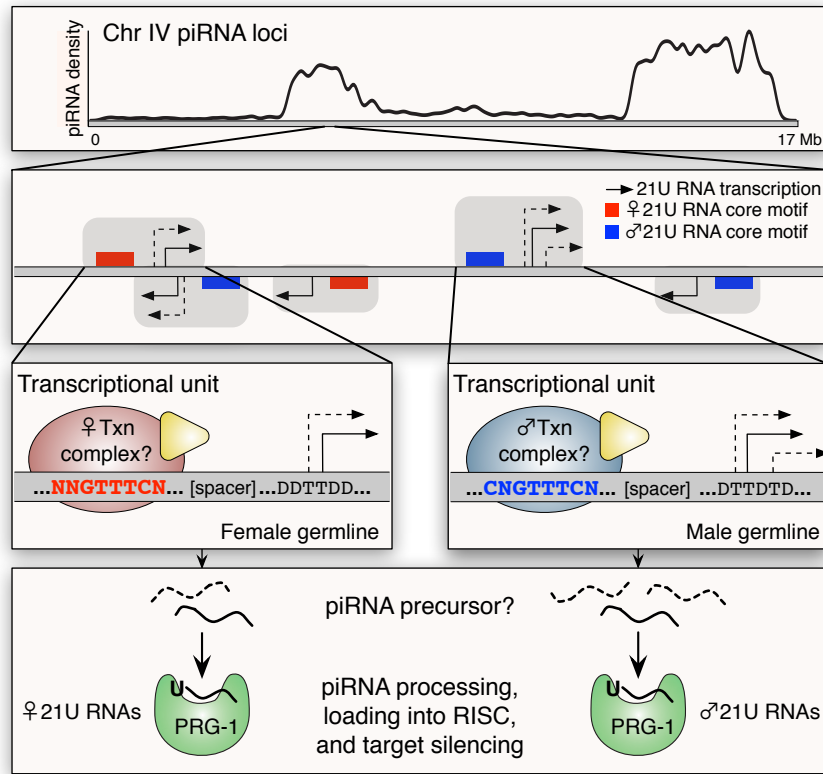


Figure S2.8: Proposed model of 21U RNA germline-enriched expression. Txn: Transcriptional. Solid arrow lines: strong expression. Dotted arrow lines: weak expression.

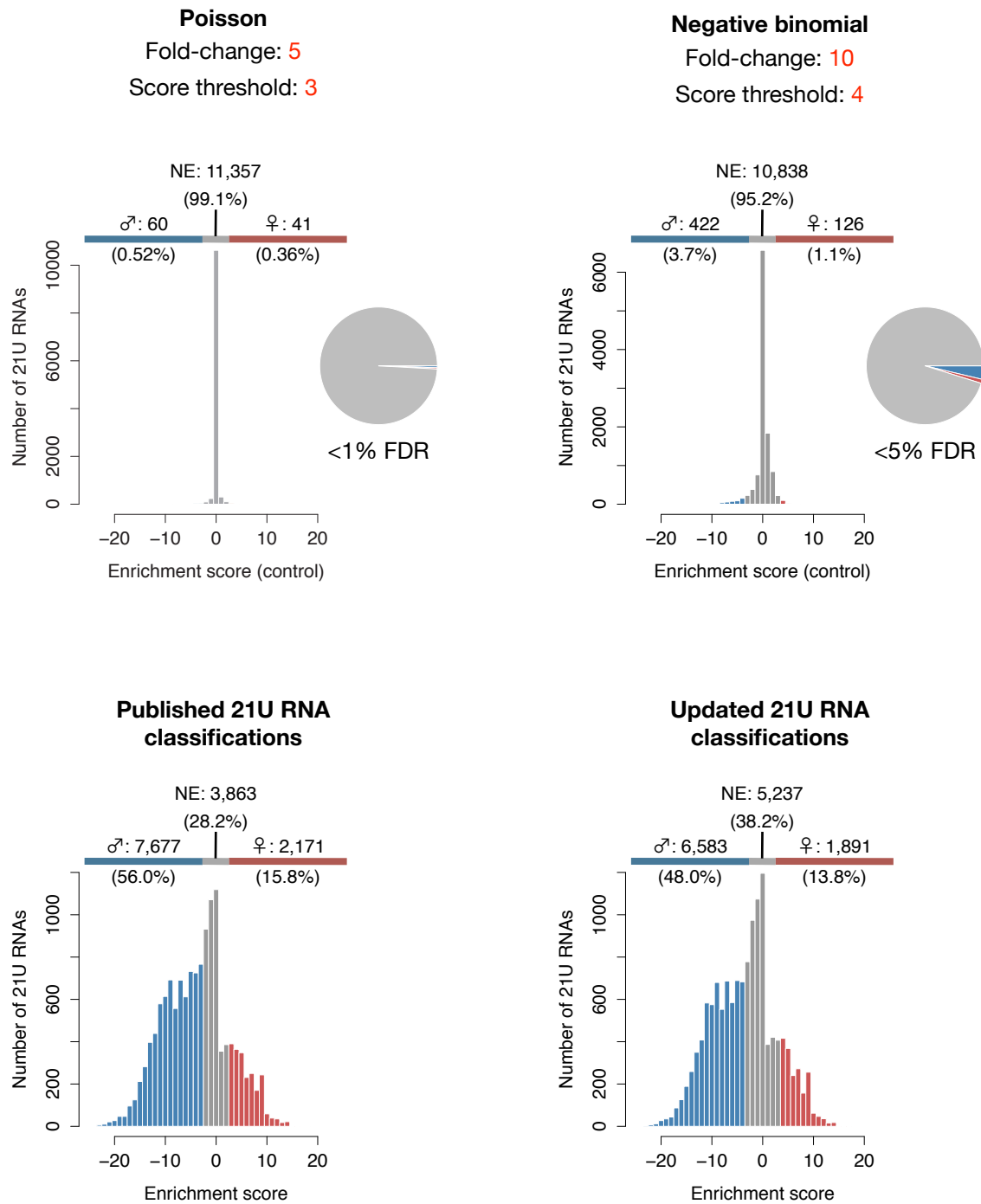


Figure S2.9: 21U RNA Enrichment Score calculations were re-assessed using a negative binomial distribution to model 21U RNA count data (right) instead of a Poisson distribution (left).

CHAPTER III

Global characteristics of PTGR investigated by gPAR-CLIP-seq

3.1 Introduction

A diverse and expanding repertoire of RNA-binding proteins (RBPs) ensures faithful expression and function of substrate mRNAs (Martin and Ephrussi, 2009; Moore and Proudfoot, 2009; Glisovic et al., 2008). Many RNAs are organized by RBPs and other protein co-factors into higher-order ribonucleoprotein (RNP) assemblies that fulfill critical functions in storage, transport, inheritance, and degradation of RNA (Anderson and Kedersha, 2009; Hafidh et al., 2011). For example, over 70% of mRNAs in *Drosophila* embryos are localized to distinct organelles, compartments, and membrane interfaces, providing a means for directing local translation and regulating cellular architectures and functions (Lécuyer et al., 2007). RNA and RBPs can also reversibly aggregate into granules to allow RNA storage and decay in response to stimuli (Han et al., 2012; Kato et al., 2012). These and many other processes are driven by large, complex networks of protein-RNA interactions that provide specificity in gene regulation and fidelity in RNP assembly. Despite important insights regarding the necessity of RNA regulation for cellular functions, the RBP-RNA interactome

Official citation:

Freeberg MA*, Han T*, Moresco JJ, Kong A, Yang Y-C, Lu JL, Yates JR, and Kim JK. (2013) Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*. *Genome Biol* **14**, R13. doi:10.1186/gb-2013-14-2-r1. (*equal contribution)

and its response to changing cellular conditions have yet to be fully elucidated.

Studies of RBP-RNA interactions have historically relied on the identification of target transcripts bound by individual RBPs. *In vitro* selection of RNA sequences that bind RBPs with high affinity (systematic evolution of ligands by exponential enrichment, SELEX) can identify primary sequence recognition elements. For example, Nova proteins, which regulate mRNA splicing in neurons, recognize the RNA consensus sequence YCAY (Jensen et al., 2000b), and Y box-binding protein-1, a member of the cold shock/Y box domain protein family, recognizes a CAYC RNA motif (Wei et al., 2012). Yet these and other primary sequence elements identified *in vitro* are generally short and degenerate and appear too frequently in the transcriptome to be useful for *in silico* target identification. Microarray profiling of transcripts that co-purify with interacting proteins (RIP-Chip) has been widely used to detect transcripts stably associated with RBPs, such as mRNAs bound by translational components HuB, eIF-4E, and PABP in P19 embryonal carcinoma stem cells (Tenenbaum et al., 2000). Similarly, RIP-Chip experiments have identified mRNAs associated with 40 yeast RBPs and uncovered a set of potential RBP recognition motifs (Gerber et al., 2004; Hogan et al., 2008), some of which were validated *in vitro* using SELEX (Riordan et al., 2011). Although capable of identifying mRNA targets for select RBPs, RIP-Chip is prone to artifacts, including RBP-RNA dissociation and re-association after cell lysis (Mili and Steitz, 2004), isolation of non-specific RNAs, and indirect binding through other co-purified RBPs (Darnell et al., 2005). In addition, RIP-Chip cannot detect transient interactions or resolve the exact RBP binding sites on identified transcripts.

To identify transcriptome-wide footprints of RBPs *in vivo*, UV crosslinking has been coupled with immunopurification of RBPs (CLIP) (Licatalosi et al., 2008; Ule et al., 2003). CLIP takes advantage of the photoreactivity of RNA bases, most often pyrimidines, with interacting amino acid side chains upon 254 nm UV irradiation (Ule et al., 2005). The

formation of covalent linkages allows stringent purification of RBP-RNA complexes and subsequent identification of crosslinked RNA fragments via cDNA sequencing. Recently, a modified CLIP technique, PAR-CLIP (photoactivatable-ribonucleoside-enhanced CLIP), has been introduced in which photoactivatable-ribonucleoside analogs are incorporated into the transcriptome in live cells to enable efficient crosslinking using 365 nm UV irradiation (Hafner et al., 2010b). Recent studies employing CLIP in mouse brain (Chi et al., 2009) and *Caenorhabditis elegans* (Zisoulis et al., 2010) and PAR-CLIP in human embryonic kidney cells (Hafner et al., 2010b) have successfully decoded *in vivo* microRNA-mRNA interactions by identifying RNAs bound to Argonaute, a main component of the microRNA-induced silencing complex. PAR-CLIP has also been implemented to elucidate the regulatory mechanisms of human antigen R (HuR) protein, which stabilizes gene expression by binding to AU-rich elements (Lebedeva et al., 2011; Mukherjee et al., 2011), and to identify the transcriptome-wide distribution of non-poly(A) termination factors in yeast (Creamer et al., 2011). In addition to enabling efficient crosslinking, PAR-CLIP generates frequent and non-random nucleotide substitutions at crosslinking sites to reveal specific RBP-RNA contact sites with nucleotide resolution.

Until recently, CLIP and PAR-CLIP have been limited to investigation of individual RBPs. Two recent studies introduced the use of photoactivatable-ribonucleoside-enhanced UV crosslinking with oligo(dT) pull-down of mRNAs followed by tandem mass spectrometry to globally identify mRNA-binding proteins in human cell lines (Baltz et al., 2012; Castello et al., 2012). In addition to identifying known RBPs, these studies identified 315 (Baltz et al., 2012) and 245 (Castello et al., 2012) novel RBPs that lack canonical RNA-binding domains and functional annotation as RNA-binding proteins. Castello et al. (2012) found that RBP amino acid sequences are more disordered than those of non-RBPs and identified potential new classes of RNA-binding domains. Baltz et al. (2012) additionally captured

and sequenced protein-bound mRNAs, providing a transcriptome-wide map of potential *cis*-regulatory elements.

Despite recent advances towards understanding global RBP-RNA interactions, the dynamic nature of these associations *in vivo* and the general principles driving these associations remain unexplored. Here, we adapt the PAR-CLIP technique to map all RBP binding sites across the yeast non-translating mRNAs in different environmental conditions, a method we call global PAR-CLIP (gPAR-CLIP). The comprehensive identification of RBP-RNA crosslinked sites visualized by gPAR-CLIP allows us to derive general properties of RBP-RNA interactions *in vivo*. Additionally, we compared RBP-RNA crosslinked sites in rapidly proliferating versus stress-treated cells and observed large-scale changes in RBP-RNA interactions, providing a starting point for dissecting the network of post-transcriptional gene regulatory mechanisms underlying stress response.

3.2 Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*.

gPAR-CLIP identifies transcriptome-wide RBP crosslinking sites. To construct a global map of RBP binding sites on the transcriptome *in vivo*, we combined PAR-CLIP with high-throughput sequencing (Fig. 3.1A; Materials and methods). Briefly, we metabolically incorporated the photoactivatable-nucleobase analog 4sU in growing yeast and used UV irradiation to crosslink 4sU to juxtaposed proteins, ‘freezing’ protein-RNA interactions *in vivo*. Next, we implemented three biochemical strategies to capture RNA regions bound by the proteome: (i) sucrose gradient centrifugation to reduce ribosome abundance; (ii) oligo(dT) selection to deplete abundant structural non-coding RNAs (for example, rRNAs); and (iii) chemical biotinylation of proteins. We exploited the high-affinity streptavidin-biotin

interaction to purify all biotin-protein-RNA complexes with high efficiency and stringency. After trimming unbound RNA, RBP-protected fragments were ligated to linkers, converted to cDNAs, and subjected to Illumina high-throughput sequencing (Materials and methods). We term this global PAR-CLIP method ‘gPAR-CLIP’. There are two caveats associated with our gPAR-CLIP protocol. First, during crosslinking (approximately 5 minutes), the cells were in nutrient-free buffer and incubated on ice, which could trigger changes in RBP binding. Second, we limited our analysis to mRNAs from the top of the sucrose gradient, which mostly consist of non-translating mRNAs, so our conclusions apply to non-translating mRNAs.

To define the dynamic landscape of RBP-RNA interactions, we constructed duplicate gPAR-CLIP and mRNA-seq libraries, including incorporation of 4sU, for the wild-type *Saccharomyces cerevisiae* strain cultured in complete media or subjected to glucose or nitrogen starvation for 2 hours. An average of 10 million reads were sequenced from each gPAR-CLIP library. Of the 72% of reads that mapped uniquely to the genome, over 70% contained one or two T-to-C conversion events, the signature substitution induced by 4sU crosslinking (Fig. S3.1). From overlapping reads, we derived clusters representing RNA regions crosslinked to proteins (Fig. S3.2 and Fig. S3.3; Materials and methods). Crosslinking site read coverage was normalized to mRNA expression levels calculated as reads per million mapped reads per kilobase of transcript (RPKM). Because our approach captures protein-RNA interactions for potentially all RBPs, we cannot rule out the possibility that some clusters are located proximally to true RBP-binding sites (Sugimoto et al., 2012); therefore, we refer to our gPAR-CLIP read clusters as ‘crosslinking sites’. We empirically assigned a false discovery rate (FDR) to each crosslinking site by deriving clusters from mRNA-seq reads with one or two T-to-C mismatches representing sequencing error and comparing the T-to-C conversion rate of these clusters to those derived from gPAR-CLIP reads (Materials and methods). Us-

ing a 1% FDR threshold, we reproducibly identified 80,883 crosslinking sites that are, on average, 23 nucleotides long (Fig. 3.1B,C): 65,992 in protein-coding sequences (CDSs), 4,508 in 5'UTRs, 8,525 in 3'UTRs, and 818 in introns (Fig. S3.3; [Data file 1 - click to download](#)). Of 6,717 annotated protein-coding transcripts, 6,228 (93%) have at least one crosslinking site. Because CDS crosslinking sites exhibited three-nucleotide periodicity, a hallmark of ribosome binding (Fig. S3.4), we separately analyzed CDS, 5'UTR, and 3'UTR crosslinking sites. We observed high correlation between gPAR-CLIP read coverage of 5'UTRs and CDSs (Pearson correlation coefficient, $R^2 = 0.45$), supporting a prominent role for 5'UTRs in translational regulation (Fig. 3.1D). However, gPAR-CLIP read coverage of 3'UTRs correlated poorly with both 5'UTRs ($R^2 = 0.20$) and CDS ($R^2 = 0.21$), suggesting a greater role in post-transcriptional regulation. As expected, correlation between total mRNA-seq read coverage across all transcript regions was approximately equal (Fig. 3.1D; [Data file 2](#)). We also observed very high correlation of gPAR-CLIP and mRNA-seq read coverage between replicates over each genic region (Fig. 3.1D), reflecting low technical variation between replicates.

gPAR-CLIP captures known and novel crosslinking sites. To assess the performance of gPAR-CLIP in capturing known RBP-RNA interactions, we evaluated its ability to identify binding sites of Puf3p, a Pumilio-family RBP, which we derived from conventional PAR-CLIP using a strain expressing a TAP-tagged Puf3p fusion protein. Of the 1,236 Puf3p binding sites confidently identified by PAR-CLIP, 1,008 (82%) were also captured by gPAR-CLIP (Fig. 3.2A; [Data file 3](#)); for example, the two functionally validated Puf3p binding sites in the *COX17* 3'UTR were identified by both Puf3p PAR-CLIP and gPAR-CLIP (Fig. 3.2B) (Jackson et al., 2004; Olivas and Parker, 2000). It is possible that other Puf proteins with similar RNA recognition motifs are binding at these sites in our gPAR-CLIP libraries (Valley et al., 2012), reflecting that our protocol does not distinguish the RBPs associated

with each crosslinking site. From our Puf3p PAR-CLIP library, we also identified 560 novel Puf3p mRNA targets harboring a binding site containing a Puf3p recognition motif (Fig. S3.5; [Data file 4](#)). Given the high recovery of Puf3p sites by gPAR-CLIP, we conclude that gPAR-CLIP faithfully captures binding sites of a known RBP.

We next examined our data for general RBP-RNA interaction signatures related to mRNA maturation and translational regulation. gPAR-CLIP read coverage of 5'UTRs peaked within 75 nucleotides downstream of annotated transcription start sites but was reduced when yeast were grown in media lacking glucose or nitrogen (Fig. 3.2C). This coverage likely reflects RBP-RNA interactions involved in translation initiation, and the decrease in coverage is consistent with decreased translation initiation that occurs during cellular stress (Ashe et al., 2000; Gallego et al., 1997; Simpson and Ashe, 2012). gPAR-CLIP also effectively captured the spliceosome binding pattern by identifying intronic RBP crosslinking sites clustering 3' of the lariat branch point (BP) bound by the U2 snRNP (Fig. 3.2D). These crosslinking sites contain the canonical BP-binding protein recognition sequence UACUAAC (Berglund et al., 1997; Garrey et al., 2006). Consistent with stress-induced transcriptional repression of ribosomal subunits (Joo et al., 2011), which account for 18% of all protein-coding genes with introns, gPAR-CLIP read coverage at the lariat BPs of ribosome-encoding mRNA introns decreased upon glucose and nitrogen deprivation. Finally, a strong RBP crosslinking signature was identified approximately 20 nucleotides upstream of the most prominent poly(A) junction site identified in each 3'UTR (Fig. 3.2E), consistent with interactions with the polyadenylation complex (Ozsolak et al., 2010). Taken together, these results indicate that gPAR-CLIP faithfully captures diverse RBP-RNA interactions along the discrete anatomy of mRNAs.

RBP crosslinking sites exhibit global conservation in both primary sequences

and secondary structures. Compared to mRNA-seq reads, which were equally distributed among 5' and 3'UTRs and CDSs, gPAR-CLIP reads were 4-fold enriched in 3'UTRs, 2.5-fold enriched in 5'UTRs, and 4-fold depleted in CDSs compared to mRNA-seq reads (Fig. 3.3A). To examine RBP binding activity at nucleotide resolution, we calculated a crosslinking score (CLS) for each T in the genome (U in the transcriptome) as the ratio of gPAR-CLIP reads containing one or two T-to-C conversion events to mRNA-seq reads to normalize for variable mRNA abundance (Data file 5; Materials and methods). There were 378,247 Ts (12.7% of transcriptomic Us) assigned a CLS: high CLS values indicate high crosslinking efficiency and strong RBP-RNA interactions; low CLS values indicate low crosslinking efficiency or weak/transient RBP-RNA interactions. Consistent with the distribution of gPAR-CLIP reads, CLS values were highest in 3'UTRs followed by 5'UTRs and CDSs (Fig. 3.3B; Fig. S3.6). These observations support 3'UTRs as the primary sites for RBP-RNA interactions for non-translating mRNAs. To determine if enrichment of gPAR-CLIP reads on UTRs was biased because of the U-richness of UTRs, we compared the proportion of Us in each crosslinking site to its coverage in gPAR-CLIP and observed only a weak positive correlation, which by itself cannot account for the four-fold enrichment of gPAR-CLIP reads on UTRs (Fig. 3.3C).

A previous comparative analysis of seven *Saccharomyces* genomes revealed that approximately 14% of evolutionarily constrained bases lie outside protein-coding regions, often located in UTRs (Siepel et al., 2005). These conserved regions could represent functional elements interacting with *cis*-acting factors. We found direct evidence of RBPs crosslinking to 35% of conserved sequence blocks in UTRs as defined by phastCons, a score representing the likelihood that a base falls in a conserved element (Fig. 3.3D): 405 of 1,549 5'UTR blocks (26%) and 1,036 of 2,536 3'UTR blocks (41%) completely overlap with at least one RBP crosslinking site, which is significantly higher than randomly defined control blocks (X^2 test,

$P < 10^{-119}$ for 3'UTR and $P < 10^{-22}$ for 5'UTR).

At the gene level, *ATG8*, a key autophagy gene, contains two major crosslinking sites that overlap with conserved sequence blocks in its 3'UTR (Fig. 3.3E, top; Fig. S3.6). Similarly, *TOM40*, which encodes a translocase that mediates import of mitochondria-localized proteins into the mitochondria, contains two major 3'UTR crosslinking sites in regions with high local conservation (Fig. 3.3E, bottom; Fig. S3.6). To further elucidate the connection between RBP binding and conservation, we binned Ts by CLS values and observed that Ts in all 3' and 5'UTR bins, as well as the majority of CDS (78%) bins, were more conserved than randomly binned Ts, suggesting that RBP crosslinking sites are under purifying selection (Fig. 3.3F; Fig. S3.6).

Unexpectedly, 3' and 5'UTR nucleotides in the lowest CLS bins exhibited extremely high conservation. Since a low CLS can indicate inefficient RNA capture, and gPAR-CLIP inefficiently captures highly structured, double-stranded RNA (see Discussion), we hypothesized that low CLS/high conservation bins represent conserved, secondary structure motifs recognized by RBPs. For example, She2p binds a distinct stem-loop structure in several bud-localized mRNAs (Chartrand et al., 1999; Gonzalez et al., 1999), including *ASH1*, for which the She2p 3'UTR recognition element is weakly represented in our gPAR-CLIP dataset. To determine if Ts with low CLS values are located in RNA regions with a high degree of secondary structure, we computed the probability of each T being unpaired using RNAplfold, a local thermodynamic folding algorithm (Bernhart et al., 2006). We observed that Ts with low CLS values exhibited low unpaired probabilities, suggesting they are more likely to exist in double-stranded structures (Fig. 3.4A; Fig. S3.7). Additionally, a strong, positive correlation between unpaired probability and CLS values indicates that unpaired regions crosslink more strongly to RBPs. To probe RNA secondary structures more accurately, we extended the boundaries of each crosslinking site to span 80 nucleotides and calculated

the most thermodynamically stable secondary structure. Consistent with the per nucleotide analysis, crosslinking sites with low CLS values formed predominantly double-stranded RNA structures (Fig. 3.4B; Fig. S3.7).

Secondary structures tolerate substitutions that preserve base pairing in stem regions, a characteristic known as covariance. To identify conserved and thermodynamically stable RNA secondary structures using a covariance model, the seven yeast genomes were scanned with RNAz (Steigele et al., 2007; Washietl et al., 2005), and a small set of potential structural elements was identified: 843 in CDS, 25 in 5'UTRs, and 51 in 3'UTRs. Among Ts assigned a CLS, those with the lowest CLS values in CDS and 3'UTRs were preferentially located in conserved, structural elements compared to control elements (Fig. 3.4C; Fig. S3.7). Taken together, our per nucleotide and per crosslinking site results indicate that high conservation observed for Ts with low CLS values is driven by conserved RNA secondary structures, while Ts with high CLS values are located in exposed, single-stranded RNA regions available for sequence-specific contact with RBPs.

Large-scale changes of RBP crosslinking site occupancy occur upon nutrient deprivation. To explore RBP-RNA interaction dynamics under changing cellular conditions, we compared gPAR-CLIP read coverage of individual 3'UTR crosslinking sites between glucose or nitrogen starvation and log-phase growth conditions. As we selected non-translating mRNAs for gPAR-CLIP analyses, we cannot distinguish whether the changes in binding site coverage reflect changes in RBP binding or changes in RBP distribution in the sucrose gradient (see Discussion). We only examined crosslinking sites with >5 reads per million mapped reads (RPM) in gPAR-CLIP libraries to ensure confident quantification (Fig. S3.8; Materials and methods). The intra-replicate variation of crosslinking site read coverage was quantified as standard deviation $\sigma = 1.3$ -fold (Fig. S3.8); therefore, we consider

crosslinking sites with more than 4-fold (3σ) differences in read coverage between wild type and stress conditions as ‘increased’ or ‘decreased’. We observed >4-fold changes in crosslinking site coverage, also referred to as ‘RBP occupancy’, for 1,129 of 3,803 (30%) 3’UTR sites upon glucose starvation and for 535 of 3,932 (14%) 3’UTR sites upon nitrogen starvation (1,497 of 3,985 3’UTR sites in either condition, 38%) (Fig. 3.5A,B). Similar distributions of changes were observed for crosslinking sites in 5’UTRs (Fig. S3.8). Nineteen percent (116 of 623) of crosslinking sites that exhibited decreased RBP occupancy were affected by both conditions, while only 5% (40 of 885) of crosslinking sites that exhibited increased RBP occupancy were affected by both conditions, suggesting that RBP-RNA interaction changes are largely distinct to glucose or nitrogen deprivation (Fig. 3.5C). Similar to the observation that glucose starvation induced more crosslinking site occupancy changes than nitrogen starvation, comparison of mRNA abundance revealed more changes in gene expression upon glucose than nitrogen starvation (Fig. 3.5D,E). Interestingly, mRNA expression of ribosomal subunits and other known RBPs was significantly down-regulated upon glucose (Welch’s *t*-test, $P < 10^{-27}$) and nitrogen (Welch’s *t*-test, $P < 10^{-36}$) starvation, suggesting that global suppression of post-transcriptional regulation is a general response to nutrient deprivation.

We next examined the overlap of individual genes with 3’UTR crosslinking sites affected by each stress condition (Fig. 3.5F). Genes harboring 3’UTR crosslinking sites with increased RBP occupancy showed little overlap (41 genes, 6%) between the two conditions; genes harboring crosslinking sites with decreased RBP occupancy showed higher overlap (114 genes, 21%). These data suggest that, for the non-translated mRNA transcriptome, loss of RBP occupancy at crosslinking sites of a larger set of common genes is a general response to nutrient limitation while increased RBP occupancy at crosslinking sites of distinct sets of genes is a nutrient-specific response.

We determined if genes exhibiting common or distinct 3'UTR crosslinking site occupancy changes under nitrogen and glucose starvation conditions had shared biological functions or cytological localization using Gene Ontology (GO) enrichment analysis (Fig. 3.5G,H; [Data file 6](#)). When we analyzed the 356 genes with sites decreased in RBP occupancy only during glucose starvation, mitochondrion-related genes and genes associated with cellular respiration were preferentially affected (Fig. 3.5G, top). Analysis of the 77 genes with sites lost only during nitrogen starvation revealed enrichment for ribosomal components and noncoding RNA processing (Fig. 3.5G, middle). The 114 genes harboring 3'UTR crosslinking sites with decreased coverage under both stress conditions were enriched for fatty acid and lipid catabolism (Fig. 3.5G, bottom), consistent with the utilization of stored lipids as energy source in response to nutrient deprivation (Hohmann and Mager, 2003).

Analysis of the 400 genes harboring 3'UTR crosslinking sites with increased occupancy only upon glucose starvation were enriched for terms related to translation (Fig. 3.5G, top). The 254 genes harboring sites with increased RBP occupancy only upon nitrogen starvation were enriched for metabolic processes, including glutamate metabolic processes, which are affected by nitrogen availability (Fig. 3.5H, middle). Of the 41 genes harboring 3'UTR crosslinking sites with increased RBP occupancy under both nitrogen and glucose starvation conditions, 13 (32%) genes represent cellular components of ribosomes or mitochondria (Fig. 3.5H, bottom), consistent with induction of global changes through translational repression and changes in energy metabolism.

In order to determine whether these observations are a result of changes in mRNA abundance, we calculated GO term enrichment of mRNAs up- or down-regulated upon glucose or nitrogen starvation and observed that down-regulated mRNAs are enriched for ribosome- and translation-related genes, while up-regulated mRNAs are enriched for genes related to mitochondrion and metabolic processes (Fig. S3.9). Therefore, the GO term enrichment of

genes with changes in 3'UTR site occupancy cannot be fully explained by GO term enrichment of up- or down-regulated mRNAs. Taken together, these data indicate that general nutrient limitation triggers a remodeling of the post-transcriptional regulatory programs of metabolic pathways, while glucose- and nitrogen-specific stresses affect additional, distinct biological processes.

We further visualized changes in RBP occupancy of 3'UTR crosslinking sites relative to the changes in corresponding mRNA abundance induced by glucose starvation (Fig. 3.6A; Fig. S3.10). Since 3'UTR crosslinking sites with decreased RBP occupancy were enriched for mitochondrion-related genes, we examined sites on a subset of these genes encoding mitochondrial membrane components and observed that the crosslinking sites were significantly depleted of RBP occupancy compared to all 3'UTR crosslinking sites (Welch's t -test, $P < 10^{-21}$), and the mRNAs were significantly up-regulated compared to all genes (Welch's t -test, $P < 10^{-28}$) (Fig. 3.6A, blue dots). This observation suggests that 3'UTR crosslinking sites on mRNAs encoding mitochondrial membrane components are recognized by repressive RBPs, and that upon glucose deprivation, RBP-binding is attenuated, resulting in increased mRNA levels.

Mitochondrial aldehyde dehydrogenase *ALD4* mRNA is regulated transcriptionally under stress conditions (Larochelle et al., 2006; Vyas et al., 2005). In our gPAR-CLIP data, the *ALD4* 3'UTR harbors four highly conserved crosslinking sites displaying 2- to 8-fold decreases in RBP occupancy despite a >7-fold increase in *ALD4* mRNA levels (Fig. 3.6B; Fig. S3.11). These data suggest that post-transcriptional regulation of *ALD4* in response to glucose deprivation also occurs through the release of repressive RBP binding at these 3'UTR sites. *STM1*, which encodes a ribosomal subunit-associated protein required for optimal translation under nutrient stress (Griac and Henry, 1999), has two 3'UTR crosslinking sites, with one exhibiting >25-fold increased RBP occupancy upon glucose starvation (Fig.

3.6C; Fig. S3.11). *STM1* mRNA is conversely down-regulated >3-fold, indicating a potential regulatory role for this site involving mRNA stability and/or decay. Interestingly, *STM1* mRNA expression is also down-regulated upon nitrogen starvation despite no change in RBP occupancy of this site, pointing to non-overlapping regulatory mechanisms that contribute to *STM1* regulation in glucose and nitrogen starvation conditions.

Next we explored changes in RBP occupancy of 3'UTR crosslinking sites relative to changes in corresponding mRNA abundance upon nitrogen starvation. 3'UTR crosslinking sites on mRNAs associated with ribosome biogenesis showed significantly greater decrease in RBP occupancy compared to all 3'UTR crosslinking sites (Welch's *t*-test, $P < 10^{-12}$) (Fig. 3.7A, red dots; Fig. S3.12). Inositol-3-phosphate synthase *INO1* is transcriptionally regulated under stress (Griac and Henry, 1999). *INO1* 3'UTR has four conserved crosslinking sites, one of which exhibits a >50-fold increase in RBP occupancy upon nitrogen starvation despite a >10-fold decrease in *INO1* mRNA levels (Fig. 3.7B; Fig. S3.13). These data suggest post-transcriptional regulation of *INO1* mRNA by a specific RBP-RNA interaction in the 3'UTR. We also identified three crosslinking sites under normal growth conditions in the 3'UTR of *AGP3*, an amino acid permease capable of supplying amino acids as an alternative nitrogen source in nitrogen-poor conditions (Schreve and Garrett, 2004). RBP occupancy at these sites was completely lost upon nitrogen starvation while two additional sites emerged (Fig. 3.7C; Fig. S3.13). *AGP3* mRNA levels moderately increased approximately two-fold (Fig. 3.7C), suggesting complex, combinatorial post-transcriptional regulation of *AGP3* expression in nitrogen-poor conditions.

3.3 Materials and methods

3.3.1 Specific to this study

Strains, media and growth conditions

The following strains were used in this study: wild-type BY4742 (*MAT α his3 Δ 1 leu2 Δ 0 lys2 Δ 0 ura3 Δ 0*), TAP-tagged strains picked from TAP-tagged yeast strain collection (*MAT α his3 Δ 1 leu2 Δ 0 met15 Δ 0 ura3 Δ 0 YFG-TAP::HIS5*) (Ghaemmaghami et al., 2003). Strains were grown at 30°C with vigorous shaking (250 rpm) in synthetic defined media, supplemented with 200 μ M 4sU (Sigma-Aldrich 440736; St. Louis, MO, USA), to OD₆₀₀ = 0.7 to 0.8. Starvation was performed by pelleting cells for 5 minutes at 3,000Xg at room temperature, discarding all media, rinsing once with H₂O, and resuspending cells in an equal volume of synthetic defined media without glucose or nitrogen (supplemented with 200 μ M 4sU). Cells were returned to 30°C with shaking for 2 h. Strains used are defective in uracil synthesis (*ura3 Δ*) and readily take up 4sU from the media. Inside the cell, 4sU is converted by Fur1p (uracil phosphoribosyltransferase) to 4-thiouridine monophosphate that can be incorporated during RNA synthesis.

Estimation of 4sU incorporation rates

4sU incorporation rates were measured as described (Favre et al., 1993). Briefly, RNA samples isolated from cells grown in the presence or absence of 4sU were dissolved in 100 μ l of 12 mM Tris buffer, pH 7, and their A₂₆₀ absorption was adjusted to the same value. A₃₃₀ was measured for both samples using a Q6 quartz cuvette with 1 mm light path in a Thermo Scientific BioMate 3 UV-Vis spectrophotometer. 4sU incorporation rates per kilobase RNA were calculated as $500 \times [(A_{330}(+4sU)) - (A_{330}(-4sU))]/A_{260}$. 4sU was incorporated at roughly four 4sU per kilobase of transcript, with little interference with cell growth and only minor

changes in gene expression (Fig. S3.3).

gPAR-CLIP procedures

UV crosslinking

Mid-log-phase cultures (50 ml; OD₆₀₀ of 0.7 to 0.8) were pelleted for 5 minutes at 3,000Xg at room temperature, resuspended in 2 ml of 1x HBSS (Invitrogen 14025; Grand Island, NY, USA) and transferred to a 60 mm cell culture dish (BD Biosciences 353002; San Jose, CA, USA), placed on ice, and irradiated with 365 nm UV at 150 mJ/cm² four times using a UVP CL-1000L UV crosslinker. The cells were then pelleted for 2 minutes at 5,000Xg at 4°C. After removing 1x HBSS, the cells were frozen in liquid nitrogen.

Extract preparation

Crosslinked cells were resuspended in polysome lysis buffer (20 mM HEPES pH 7.5, 140 mM KCl, 1.5 mM MgCl₂, 1% Triton X-100, 1x Complete Mini Protease Inhibitor EDTA-free (Roche Applied Science 1 836 170; Indianapolis, IN, USA), 0.2 U/μl SUPERaseIn (Invitrogen AM2696)), mixed with half a volume of acid-washed glass beads, and lysed by vortexing four times at 4°C, 1 minute each with a 1 minute incubation on ice in between. Cell debris was removed by centrifugation for 5 minutes at 1,300Xg at 4°C. The supernatant was cleared by 20,000Xg spin for 10 minutes at 4°C.

Ribosome depletion using sucrose density gradients

Sucrose density gradients (15 to 50% (w/v)) were prepared in Beckman polycarbonate centrifugation tubes (11 x 34 mm) by sequentially layering and freezing 0.24 ml of 50%, 41.25%, 32.5%, 23.75% and 15% sucrose dissolved in polysome gradient buffer (20 mM HEPES pH 7.5, 140 mM KCl, 5 mM MgCl₂). Gradients were thawed overnight at 4°C before use. 100 μl of clarified lysate was loaded on top of a gradient, centrifuged for 1 h at 54,000 rpm at 4°C using a TLS-55 rotor in an Optima MAX-E ultracentrifuge (Beckman Coulter; Palo Alto,

CA, USA). The top 600 μl of the gradient were recovered and supplemented with 2 μl of SUPERaseIn (20 U/ μl).

Chemical biotinylation and polyA selection

Sixty microliters of freshly prepared 10 mM EZ-Link NHS-SS-Biotin (Pierce 21441; Rockford, IL, USA) dissolved in dimethylformamide was added to the recovered lysate and incubated on a Nutator for 2 h at 4°C; 50 μl of 5 M NaCl was added to increase the total salt concentration to 0.5 M. Biotinylated lysate was mixed with 1 mg of oligo(dT)₂₅ magnetic beads (NEB S1419S; Ipswich, MA, USA), then incubated on a Nutator for 30 minutes at 4°C. The beads were washed four times with ice-cold hybridization buffer (10 mM HEPES pH 7.5, 0.5 M NaCl, 1 mM EDTA) and the RNAs were eluted by incubating beads with 500 μl of elution buffer (10 mM HEPES pH 7.5, 1 mM EDTA) and heating at 65°C for 3 minutes. The eluted sample was transferred to a new tube and mixed with 55 μl of 10x phosphate-buffered saline (PBS).

Streptavidin binding and RNase T₁ digestion

PolyA-selected samples were mixed with 1 mg of streptavidin M280 Dynabeads (Invitrogen 112-05D) and incubated on a Nutator for 30 minutes at 4°C. The beads were washed three times with 1x PBS, then incubated with 20 μl of 50 U/ μl RNase T₁ (Fermentas EN0541; Waltham, MA, USA, 1:20 dilution in 1x PBS) at 22°C for 15 minutes on an Eppendorf Thermomixer (15 s shaking at 1,000 rpm followed by a 2 minute rest interval), followed by a 5 minute incubation on ice. Beads were washed twice with wash buffer (1x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40), twice with high-salt wash buffer (5x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40) and twice with 1x PNK buffer (50 mM Tris pH 7.4, 10 mM MgCl₂, 0.5% NP-40).

On-bead CIP treatment

Beads were incubated with 20 μl of CIP mix (50 mM Tris pH 7.9, 100 mM NaCl, 10 mM

MgCl₂, 0.5 U/ μ l calf intestinal alkaline phosphatase (CIP); NEB M0290S) at 37°C for 15 minutes, with 15 s shaking at 1,000 rpm followed by a 2 minute rest interval on a Thermomixer. After CIP treatment, beads were washed twice with 1x PNK+EGTA buffer (50 mM Tris pH 7.4, 20 mM EGTA, 0.5% NP-40) and twice with 1x PNK buffer.

On-bead 3' DNA linker ligation

Beads were incubated with 20 μ l of ligation mix (50 mM Tris pH 7.4, 10 mM MgCl₂, 0.5 mM DTT, 2 μ M pre-adenylated 3' DNA linker, 25% PEG-8000, 10 U/ μ l T4 RNA ligase 2, truncated K227Q; NEB M0351S) at 16°C overnight (\geq 16 h), with 15 s shaking at 1,000 rpm followed by a 2 minute interval on a Thermomixer. After linker ligation, beads were washed three times with 1x PNK+EGTA buffer.

SDS-PAGE and nitrocellulose transfer

Beads were mixed with 12 μ l of 1x PNK+EGTA buffer, 3 μ l of freshly made 1 M DTT and 15 μ l of 4x NuPAGE LDS sample buffer (Invitrogen NP0007), and incubated at 70°C for 10 minutes. Beads were removed, and the supernatant was loaded onto NuPAGE 4-12% Bis-Tris gel (Invitrogen NP0335BOX) and run at 150 V for 35 minutes. The gel was transferred to Protran BA 85 nitrocellulose membrane (pore size 0.45 μ m, Whatman 10402594; St. Louis, MO, USA) using Novex wet transfer at 30 V for 1 h. A broad band from 31 kDa up to the top of the gel was excised, cut into small pieces, and transferred into a microfuge tube.

RNA isolation and purification

Excised membranes were incubated with 500 μ l of 4 mg/ml Proteinase K prepared in 1x PK buffer (100 mM Tris pH 7.5, 50 mM NaCl, 10 mM EDTA) for 20 minutes at 37°C on a Thermomixer. We added 500 μ l of 7 M urea prepared in 1x PK buffer to the tube followed by another 20 minute incubation at 37°C. The Proteinase K digestion reaction was mixed with 1 ml of phenol:chloroform:isoamyl alcohol 25:24:1 (Sigma-Aldrich P2069) by vortexing and spun for 5 minutes at 20,000Xg. The liquid phase was transferred into a new tube, mixed

with 125 μ l of 3 M NaOAc, 2.5 ml of 100% ethanol and 1 μ l of 15 mg/ml glycoblue (Invitrogen AM9516), and precipitated for 2 h at -80°C . RNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature followed by two washes with cold 75% ethanol.

RNA 5' end phosphorylation

RNA pellets were air-dried briefly, resuspended in 10 μ l of PNK mix (70 mM Tris pH 7.6, 10 mM MgCl_2 , 5 mM DTT, 1 mM ATP, 1 U/ μ l T4 polynucleotide kinase (NEB M0201S), 1 U/ μ l SUPERaseIn) and incubated at 37°C for 30 minutes. The reaction was combined with 90 μ l of H_2O and 100 μ l of phenol:chloroform:isoamyl alcohol 25:24:1, mixed well and spun for 5 minutes at 20,000Xg. The liquid phase was mixed with 12.5 μ l of 3 M NaOAc, 250 μ l of 100% ethanol, 1 μ l of 15 mg/ml glycoblue and precipitated for 2 h at -80°C . RNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol.

5' RNA linker ligation

RNA pellets were resuspended in 10 μ l of ligation mix (50 mM Tris pH 7.5, 10 mM MgCl_2 , 10 mM DTT, 1 mM ATP, 0.1 mg/ml bovine serum albumin, 2 μ M 5' RNA linker, 1 U/ μ l T4 RNA ligase (Fermentas EL0021), 1 U/ μ l SUPERaseIn, 10% DMSO) and incubated at 15°C for 2 h.

RNA size selection

Ligation reaction was terminated by adding 10 μ l of 2x formamide gel loading buffer (Invitrogen AM8546G), heated for 2 minutes at 70°C and then quickly chilled on ice. Samples were loaded onto a 6% TBE UREA gel (Invitrogen EC6865BOX) and run at 150 V for 45 minutes. After staining with 1x Sybr Gold Stain (Invitrogen S-11494), a gel piece corresponding to a 70 to 90 nucleotide RNA (80 to 100 nucleotide single-stranded DNA) was excised, crushed, and soaked in 400 μ l of 0.3 M NaOAc overnight at room temperature. After removing gel pieces, the solution was combined with 1 ml of 100% ethanol and 1 μ l of 15 mg/ml glycoblue

and precipitated for 2 h at -80°C . RNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol. After brief drying, RNAs were resuspended in 15 μl of H_2O .

RT-PCR

The ligated RNA (10 μl) was combined with 2 μl of 5 μM RT primer, heated at 65°C for 5 minutes, and then quickly chilled on ice, and followed by the addition of 1 μl of 10 mM dNTP, 1 μl of 0.1 M DTT, 4 μl of 5x First strand buffer, 1 μl of SUPERaseIn (20 U/ μl) and 1 μl of SuperScript III Reverse transcriptase (Invitrogen 18080-093, 200 U/ μl). The RT reaction was kept at 50°C for 45 minutes, 55°C for 15 minutes and 90°C for 5 minutes. A test PCR was performed with 2.5 μl of RT product in 50 μl PCR mix: 1x AccuPrime PCR buffer I, 0.5 μM P5 long primer, 0.5 μM P7 primer, 0.2 μl AccuPrime Taq High Fidelity (Invitrogen 12346-086, 5 U/ μl). PCR was carried out with an initial 3 minute denaturation at 98°C , followed by 14 to 22 cycles of 80 s denaturation at 98°C , 90 s annealing and extension at 65°C , and termination with a final 5 minute extension at 65°C . PCR product (15 μl) was collected after 14, 18, and 22 cycles and analyzed on a 10% TBE gel (Invitrogen EC6275BOX) at 150 V for 1 h to determine the optimal amplification cycles (the lowest cycle number required to generate 96 to 116 bp amplicons detected by Sybr Gold staining).

Preparation of sequencing libraries

A 50 μl PCR reaction was carried out with the determined cycle number. Amplicons were purified using DNA clean and concentrator-5 (Zymo D4013; Irvine, CA, USA), run on 10% TBE gels at 150 V for 1 h and stained with Sybr Gold. A gel piece corresponding to 96 to 116 bp DNA was excised, crushed, and soaked overnight in 400 μl 0.3 M NaOAc at room temperature. After removing gel pieces, the solution was combined with 1 ml of 100% ethanol and 1 μl of 15 mg/ml glycoblue and precipitated for 2 h at -80°C . DNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature, followed by two

washes with cold 75% ethanol. After brief drying, amplicons were resuspended in 20 μ l of H₂O. Purified amplicons (5 μ l) were used to seed a second round of PCR in 50 μ l: 1x AccuPrime PCR buffer I, 0.5 μ M Illumina Primer A, 0.5 μ M Illumina Primer B, 0.2 μ l AccuPrime Taq High Fidelity for 6 to 12 cycles. Second PCR amplicons were purified with DNA clean and concentrator-5 (Zymo D4013) and sequenced on an Illumina HiSeq 2000 sequencer.

Puf3p PAR-CLIP procedures

Puf3p PAR-CLIP was performed similarly to gPAR-CLIP with the following modifications. The *PUF3-TAP::HIS5* strain was cultured and UV-crosslinked as in gPAR-CLIP. Cells were lysed in 1x PBS, 0.5% NP-40, 1x Complete Mini Protease Inhibitor, EDTA-free and cleared by sequential spins at 1,300Xg for 5 minutes and 20,000Xg for 10 minutes at 4°C. The clarified lysate was passed through a Costar Spin-X filter (Corning CLS8160; St. Louis, MO, USA), mixed with RNase T₁ (Fermentas EN0541) to 1 U/ μ l, and incubated at 22°C for 15 minutes followed by a 5 minute incubation on ice. The lysate was then directly mixed with IgG magnetic beads (prepared by coupling rabbit IgG (Sigma-Aldrich I5006) to Dynabeads M-270 Epoxy (Invitrogen 143-01)) to pull down Puf3p::TAP. RNase T₁ digestion, CIP treatment, and 3' DNA linker ligation were performed as described in gPAR-CLIP. Afterwards, 5' end phosphorylation was performed on-bead in 20 μ l of PNK mix (70 mM Tris pH 7.6, 10 mM MgCl₂, 5 mM DT, 1 μ l P32 rATP (6000 Ci/mmol 10 mCi/ml Perkin Elmer BLU502Z500UC; Waltham, MA, USA), 1 U/ μ l T4 polynucleotide kinase) and incubated at 37°C for 15 minutes. ATP (2 μ l of 10 mM) was added to the mix and the reaction was incubated for 10 minutes. After SDS-PAGE and transfer, crosslinked RNAs were visualized by autoradiography and the corresponding Puf3p band was excised. The remaining steps were carried out as described in gPAR-CLIP procedures, omitting the 5' end phosphorylation step.

mRNA-seq procedures

Yeast strains were grown under normal and starvation conditions described above in the presence of 4sU. Additional replicate mRNA-seq libraries were prepared with yeast strains grown under normal conditions without the addition of 4sU.

Total RNAs were extracted with acid-phenol:chloroform, pH 4.5 with isoamyl alcohol, 25:24:1 (Ambion; Grand Island, NY, USA). Replicate, strand-specific total mRNA-seq libraries were prepared in parallel using the two linker ligation protocol as described (Levin et al., 2010).

For preparation of ribo- mRNA-seq libraries, extract preparation, and ribosome depletion using sucrose density gradients were carried out as described in the gPAR-CLIP procedure (avoiding UV crosslinking). PolyA+ mRNAs were enriched using oligo(dT)₂₅ beads and converted into sequencing libraries as described (Levin et al., 2010).

Oligos for constructing gPAR-CLIP, PAR-CLIP, and mRNA-seq libraries

Oligos used in this study were synthesized by Integrated DNA Technologies, except the 5' RNA linker, which was synthesized by Dharmacon (Waltham, MA, USA).

Barcodes

3' DNA linker oligos (5' phosphorylated, and 3' block with inverted deoxythymidine):

Index 1: 5' pATCACGTCGTATGCCGTCTTCTGCTTGidT 3'

Index 2: 5' pCGATGTTCGTATGCCGTCTTCTGCTTGidT 3'

Index 3: 5' pTTAGGCTTCGTATGCCGTCTTCTGCTTGidT 3'

Index 4: 5' pTGACCATTCGTATGCCGTCTTCTGCTTGidT 3'

Index 5: 5' pACAGTGTCGTATGCCGTCTTCTGCTTGidT 3'

Index 6: 5' pGCCAATTCGTATGCCGTCTTCTGCTTGidT 3'

Index 7: 5' pCAGATCTTCGTATGCCGTCTTCTGCTTGidT 3'

Index 8: 5' pACTTGATCGTATGCCGTCTTCTGCTTGidT 3'

Pre-adenylation of 3' DNA linker oligos was performed with Mth RNA ligase (5' DNA adenylation kit, NEB E2610S) following the vendor's instructions.

5' RNA linker

5' GUUCAGAGUUCUACAGUCCGACGAUC 3'

Barcoded RT primers

Index 1: 5' CAAGCAGAAGACGGCATAACGACCGTGAT 3'

Index 2: 5' CAAGCAGAAGACGGCATAACGAACATCG 3'

Index 3: 5' CAAGCAGAAGACGGCATAACGAGCCCTAA 3'

Index 4: 5' CAAGCAGAAGACGGCATAACGATTGGTCA 3'

Index 5: 5' CAAGCAGAAGACGGCATAACGACACTGT 3'

Index 6: 5' CAAGCAGAAGACGGCATAACGAATTGGC 3'

Index 7: 5' CAAGCAGAAGACGGCATAACGAGATCTG 3'

Index 8: 5' CAAGCAGAAGACGGCATAACGATCAAGT 3'

P7 primer

5' CAAGCAGAAGACGGCATAACGA 3'

P5 long primer

5' AATGATACGGCGACCACCGACAGGTTTCAGAGTTCTACAGTCCGA 3'

Illumina primer A

5' AATGATACGGCGACCACCGA 3'

Illumina primer B

5' CAAGCAGAAGACGGCATAACGA 3'

Data processing of Illumina HiSeq sequencing reads

gPAR-CLIP, PAR-CLIP, and mRNA-seq reads were processed to remove linkers, sorted into libraries based on six-nucleotide barcodes (underlined residues above), and removed if they were low quality (Fig. S3.2). First, reads with a perfect match to a barcode were successfully sorted, followed by reads with one mismatch to a barcode. If both barcodes were perfectly matched in a read or both barcodes were found with one mismatch, the 3'-most barcode was chosen; 99.97% of reads were successfully sorted into libraries under these rules. Next, reads were removed if they met any of the following criteria: <18 nucleotides, only homopolymer As, missing 3' adapter, 5'-3' adapter ligation products, 5'-5' adapter ligation products, low quality (more than 4 bases with quality scores below 10 or more than 6 bases with a quality score below 13); 97.0% of gPAR-CLIP, 80.0% of PAR-CLIP, and 99.7% of mRNA-seq reads passed these filters. High quality reads were mapped to the *S. cerevisiae* genome version S288C with Bowtie (Langmead et al., 2009) using the following parameters: -v 3 (map with up to 3 mismatches), -k 275 (map at up to 275 loci), -best, and -strata. Mapped reads were annotated using custom scripts to known genomic elements in the S288C genome (sacCer3, April 2011) including external UTR annotations (Nagalakshmi et al., 2008; Yassour et al., 2009). Fig. S3.1 provides read counts at each processing step.

Assessing data reproducibility

To determine mRNA-seq and gPAR-CLIP replicate library reproducibility, we calculated replicate correlation using normalized read counts (RPM) of each gene. Pearson correlation coefficients for mRNA-seq libraries ranged from 0.984 to 0.994, while coefficients for gPAR-CLIP libraries ranged from 0.967 to 0.971. Due to high reproducibility, subsequent measures of read coverage represent averages of two biological replicate libraries. To determine if the addition of 4sU to growth media substantially alters transcription, biological replicate mRNA-seq libraries were generated from wild-type yeast grown under normal conditions

without the addition of 4sU. These replicate libraries had a Pearson correlation coefficient of 0.988, indicating high reproducibility, and a Pearson correlation coefficient of 0.982 when compared to wild-type yeast grown in the presence of 4sU.

Calculation of per-nucleotide crosslinking scores

To measure RBP crosslinking strength, we calculated a crosslinking score for each genomic T position as the RPM coverage from reads with a T-to-C at that position. Because transcript abundance varies greatly, from zero to tens of thousands of copies, T-to-C coverage of crosslinking sites on highly expressed genes would be preferentially higher than T-to-C coverage of crosslinking sites of lowly expressed genes. To avoid this bias, we normalized T-to-C RPMs to length-normalized transcript abundances (RPKM) from our mRNA-seq libraries. Two percent of Ts with T-to-C RPM coverage in gPAR-CLIP libraries were located on genes that lacked mRNA-seq coverage and were thus removed from further analysis. To adjust for the additional kilobase normalization factor used in RPKM, ratios of gPAR-CLIP RPM:mRNA-seq RPKM were multiplied by a factor of 1,000.

Calculation of RBP crosslinking sites

Generation of read clusters from gPAR-CLIP libraries

All six gPAR-CLIP libraries were aggregated into one large dataset to generate read clusters. A read cluster was defined as a continuous stretch of nucleotides covered by at least one gPAR-CLIP read harboring one or two T-to-C conversion events. This step resulted in 84,136 gPAR-CLIP clusters and 1,915 Puf3p PAR-CLIP clusters.

Defining crosslinking site boundaries

Manual inspection of read clusters revealed long (>100 nucleotide) regions covered by gPAR-CLIP reads containing one or more distinct peaks indicative of distinct crosslinking sites.

To distinguish between read peaks within long read clusters and trim low read coverage surrounding strong single peaks, we fit a Gaussian smoothed curve (normal kernel function, bandwidth 21) to each read cluster and used the inflection points of this curve to define the boundaries of individual crosslinking sites. This step resulted in 91,290 gPAR-CLIP crosslinking sites and 1,915 Puf3p PAR-CLIP crosslinking sites.

Calculating read coverage of crosslinking sites

From the set of RBP crosslinking sites derived from all gPAR-CLIP libraries, we determined read coverage for each site from each individual library by calculating the average RPM covering each nucleotide in the crosslinking site. This coverage was divided by the RPKM of the associated gene and multiplied by 1,000 to enable direct comparison of RBP occupancy of crosslinking sites between growth conditions.

Assigning FDR to each crosslinking site

A small fraction of T-to-C mismatches in gPAR-CLIP reads likely represent sequencing error instead of crosslinking events, so crosslinking sites derived from this error were removed. We repeated the crosslinking site generation steps using mRNA-seq reads with one or two T-to-C mismatches, which represent the rate of T-to-C sequencing error for the Illumina HiSeq platform. For each gPAR-CLIP and mRNA-seq crosslinking site, we calculated the T-to-C conversion rate as the number of reads with T-to-C conversion events divided by the number of total reads covering Ts. gPAR-CLIP and mRNA-seq crosslinking sites were binned into groups based on total read coverage. For each gPAR-CLIP crosslinking site in each bin, we determined the proportion of mRNA-seq crosslinking sites with a higher T-to-C conversion rate than the gPAR-CLIP crosslinking site. This proportion represents the FDR for that gPAR-CLIP crosslinking site. Using a strict 1% FDR threshold, we identify 80,883 gPAR-CLIP crosslinking sites.

Effect of counting statistics on error in crosslinking site coverage measurement

Read coverage of replicate gPAR-CLIP crosslinking sites was analyzed to measure reproducibility. For each site, we compared the number of reads coming from one replicate library to the total number of reads from both libraries. Perfect reproducibility would result in a ratio of 1:2. We binned crosslinking sites based on total RPM and calculated the standard deviation of these ratios for each bin. We predicted the standard deviation for counting statistics by binomial partitioning of total reads for each crosslinking site in each bin between the two replicates. When the total number of reads was below 5 RPM, binomial partitioning predominantly contributed to replicate variation (Fig. S3.8). Above 5 RPM, replicate variation stabilized, and counting statistics error contributed little to replicate error.

Conservation analysis

phastCons conservation scores for each genomic nucleotide were downloaded from Siepel et al. (2005). Ts with CLSs were grouped into 5'UTR, CDS, and 3'UTR regions and then ranked and binned by CLS so each bin overlapped adjacent bins by 50%. phastCons scores in each bin were averaged. As controls, Ts with no CLS were grouped in 5'UTR, CDS, and 3'UTR regions, randomly ranked, and binned as described. phastCons scores in each bin were averaged. Controls were calculated ten times for each region.

Unpaired probability analysis

The unpaired probability of each genomic position was calculated using RNAplfold (Bernhart et al., 2006) from the ViennaRNA package version 1.8.5 using a span of 40 nucleotides and an averaging window of 80 nucleotides. Ts with CLSs were grouped into 5'UTR, CDS, and 3'UTR regions and then ranked and binned by CLS so each bin overlapped adjacent

bins by 50%. Unpaired probabilities in each bin were averaged. As controls, Ts with no CLS were grouped into 5'UTR, CDS, and 3'UTR regions, randomly ranked, and binned. The unpaired probabilities in each bin were averaged.

Crosslinking site pairedness analysis

Genomic regions corresponding to crosslinking sites were extended to 80 nucleotides centered on the original crosslinking site. These sequences were subjected to folding using RNAfold (Hofacker et al., 1994) from the ViennaRNA package version 1.8.5, and the minimum free energy structures were extracted. Predicted structures were aligned and ranked by average crosslinking site CLS and divided into 100 equally sized, non-overlapping bins. The percentage of nucleotides predicted to be unpaired at each position in each bin was computed. Selected structures from low, middle, and high CLS bins were visualized using VARNA (Darty et al., 2009).

Enriched motif analysis

gPAR-CLIP crosslinking sites passing a 5% FDR threshold from genes identified as RBP targets by RIP-Chip experiments (Gerber et al., 2004; Hogan et al., 2008) were analyzed by MEME (Bailey and Elkan, 1994). 5'UTR, CDS, and 3'UTR crosslinking sites were analyzed separately, and third-order Markov models based on all 5'UTR, CDS, or 3'UTR regions were used to model background nucleotide compositions. Because gPAR-CLIP crosslinking sites on each target might represent a combination of RBP recognition sites, we implemented MEME using the `-mods zoops` parameter to allow zero or one motif to be found in each site. The following parameters were also used: `-evt 20`, `-minw 6`, and `-maxw 15`.

Gene Ontology enrichment analysis

GO analysis was performed on genes harboring 3'UTR crosslinking sites that were four-fold up- or down-regulated upon glucose or nitrogen starvation or both. The topGO R Bioconductor package was implemented using Fisher's exact test for enrichment and Bonferroni correction of p -values to adjust for multiple testing (Alexa et al., 2006). Up to 20 GO terms were reported with a p -value <0.01 .

3.3.2 Generalizable to any system

Summary

Protein:RNA interactions are integral components of post-transcriptional gene regulatory processes including mRNA processing and assembly of cellular architectures. Dysregulation of RNA-binding protein (RBP) expression or disruptions in RBP:RNA interactions underlie a variety of human pathologies and genetic diseases including cancer and neurodegenerative diseases (reviewed in Cooper et al., 2009; Darnell, 2010b; Lukong et al., 2008). Recent studies have uncovered only a small proportion of the extensive RBP:RNA interactome in any organism (Baltz et al., 2012; Castello et al., 2012; Freeberg et al., 2013; Hogan et al., 2008; Mitchell et al., 2013; Schueler et al., 2014; Silverman et al., 2014; Tsvetanova et al., 2010). To expand our understanding of how RBP:RNA interactions govern RNA-related processes, we developed gPAR-CLIP-seq (global photoactivatable-ribonucleoside-enhanced crosslinking and precipitation followed by deep sequencing) for capturing and sequencing all regions of the *Saccharomyces cerevisiae* transcriptome bound by RBPs (Freeberg et al., 2013). This chapter describes a pipeline for bioinformatic analysis of gPAR-CLIP-seq data. The first half of this pipeline can be implemented by running locally installed programs or by running the

Official citation:

Freeberg MA and Kim JK. (2015) Mapping the transcriptome-wide landscape of RBP binding sites using gPAR-CLIP-seq: Bioinformatic Analysis. *Methods in Mol Biol* (in press).

programs using the Galaxy platform (Goecks et al., 2010; Giardine et al., 2005; Blankenberg et al., 2010). The second half of this pipeline can be implemented by user-generated code in any language using the pseudocode provided as a template.

1 Introduction

RNA-binding proteins (RBPs) are responsible for regulating a variety of processes including storage, transport, inheritance, and degradation of RNAs. The identification of both RBP-specific and general RBP interactions with RNA is necessary for understanding the mechanisms underlying these key biological processes. Recently, techniques utilizing UV light to induce covalent bond formation between directly interacting nucleotides and amino acid side chains followed by purification of a protein of interest and deep sequencing of the bound RNAs (CLIP-seq) have been successfully implemented to identify the precise sites on target mRNAs bound by RBPs (Darnell, 2010a). In one of the first studies to pioneer this approach, distinct mRNA binding sites were identified for neuron-specific RNA-binding Nova proteins (Ule et al., 2003), which are associated with paraneoplastic neurologic degenerations (Licatalosi and Darnell, 2006) and involved in regulating alternative splicing through direct binding of targets (Jensen et al., 2000a). A modified CLIP-seq technique, PAR-CLIP-seq, was subsequently developed that incorporated photoactivatable ribonucleoside analogs into nascent transcripts to improve crosslinking efficiency and create a mismatch signature in resulting deep sequencing reads to more accurately pinpoint RBP:RNA contact sites (Hafner et al., 2010b).

Expanding upon these recent studies, we developed an approach to identify a comprehensive set of sites on the *Saccharomyces cerevisiae* transcriptome that interact with any RBP under normal or environmentally stressed conditions. Similar to traditional PAR-CIP-seq, our global PAR-CLIP-seq (gPAR-CLIP-seq) approach, described in a companion chapter,

utilizes 4-thiouridine (4sU)-enhanced UV crosslinking to promote covalent bond formation *in vivo* between closely interacting nucleotides and amino acid side chains. Instead of immunopurifying a protein of interest, we biochemically biotinylated all proteins, purified RBP:RNA complexes, and sequenced the RBP-bound RNA fragments. Our protocol captures binding patterns of all RBPs, so a novel bioinformatic analysis approach is required that was different from published methods for analyzing single-RBP PAR-CLIP-seq data. Similar to these published methods, we take advantage of the nucleotide mismatch signature resulting from crosslinking (Chou et al., 2013; Corcoran et al., 2011; Erhard et al., 2013); however, unlike these published methods, we incorporate sequencing error from mRNA-seq libraries to assign a false-discovery rate to our identified RBP binding sites, thus enabling an accurate measure of confidence that we are identifying biologically relevant RBP binding sites from background noise. We describe below our novel pipeline developed for bioinformatic analysis of high-throughput sequencing data derived from the gPAR-CLIP-seq protocol. The basic steps of the protocol, outlined in [Fig. 3.8](#), include: processing sequencing reads ([Section 3.1](#)), mapping reads to a reference genome ([Section 3.2](#)), generating binding sites and per-nucleotide crosslinking scores from mapped reads ([Sections 3.3](#) and [3.5](#)), assessing binding site quality ([Section 3.4](#)), and functionally characterizing binding sites ([Section 3.6](#)).

In parallel to performing the gPAR-CLIP-seq protocol to identify RBP-bound sites on mRNAs, we recommend performing traditional mRNA-seq to quantify transcript abundance. This allows comparisons of the relative strength of RBP binding across different transcripts using read coverage as a proxy for binding strength ([Sections 3.3](#) and [3.5](#)). mRNA-seq reads are also used in [Section 3.4](#) to calculate a false-discovery rate for each gPAR-CLIP-seq-derived binding site. If transcript abundance data are not available, global binding sites can still be calculated, but users must be careful when comparing read coverage of binding sites located on different transcripts as more gPAR-CLIP-seq reads will be recovered and

sequenced from the most highly abundant transcripts (Wang et al., 2009b; Mortazavi et al., 2008).

2 Materials

The methods presented here can be run on any operating system (Max OS X, Windows, or Linux) depending on user preference and algorithm dependencies. We implemented our methods using custom Perl (v5.10.1) scripts and code developed in R (v2.15.2) (R Core Team, 2013) or using downloadable programs, which are indicated at each step in the Methods section. Our work was executed on a single RedHat Enterprise Linux 6 machine with 256GB of RAM and two Intel Xeon E5-2680v2 10-core processors capable of hyperthreading. The computer was attached to the network and 3TB of fast NFS-based storage via 10Gbit Ethernet. Manual parallelization can speed up performance at individual steps of the computational pipeline depending on the exact programs used. For example, read mapping (**Section 3.3**) can be parallelized by breaking up a raw sequencing read file into multiple input files for the Bowtie algorithm.

3 Methods

3.1 *Processing sequencing reads*

Prior to mapping, sequencing reads must be sorted into their respective samples, if libraries were multiplexed, and processed to remove undesirable sequences. Sequencing centers may offer to perform these steps before returning files of sequencing reads. If not, users should perform the following steps.

1. De-multiplex libraries. In the case of multiplexed libraries, reads need to be sorted into their respective samples based on barcode sequences added during cDNA library preparation. This can be accomplished using the *FASTX Barcode splitter* algorithm from

the FASTX-Toolkit available through the Galaxy platform or for download at http://hannonlab.cshl.edu/fastx_toolkit/download.html. Recommended parameters for *FASTX Barcode splitter*: [-mismatches 1] for 6-nt barcodes (see **Note 1**).

2. Trim adapters/linkers. Sequencing reads need to be trimmed of artificial adapter sequences added during cDNA library preparation. This can be accomplished using the *FASTQ/A Clipper* algorithm from the FASTX-Toolkit. Recommended parameters for *FASTQ/A Clipper*: [-l 15] [-C] (see **Note 2**).
3. Remove low-quality sequences. To increase mapping efficiency, low-quality reads and low-quality nucleotides from 3' ends of reads should be removed (see **Note 3**). This can be accomplished using the *FASTQ Quality Filter* (for removing low-quality reads) and *FASTQ Quality Trimmer* (for removing low-quality nucleotides from 3' ends of reads) algorithms from the FASTX-Toolkit. Recommended parameters for *FASTQ Quality Filter*: [-q 30]. Recommended parameters for *FASTQ Quality Trimmer*: [-t 30] [-l 15].

3.2 Mapping reads to the genome

Many programs are available for mapping sequencing reads to reference genomes. Users are encouraged to use a mapping program with which he or she is most familiar.

1. Map gPAR-CLIP-seq reads to a reference genome. Use Bowtie (Langmead et al., 2009), or an alternate mapping algorithm, to map gPAR-CLIP-seq reads to a reference genome. Bowtie is available through Galaxy or for download at <https://github.com/BenLangmead/bowtie>. Mapping output can be saved in BAM and SAM formats, which are commonly used as input to a variety of downstream analysis programs. Recommended Bowtie parameters: [-v 3] [-best] [-strata] (see **Note 4**).

2. Map mRNA-seq reads to a reference genome. Use Bowtie, or an alternate mapping algorithm, to map mRNA-seq reads to the same reference genome used above. Reads mapping with 0 mismatches will be used for transcript quantification (**Section 3.3**); reads mapping with 0-2 T-to-C mismatches will be used for FDR calculations (**Section 3.4**).
3. Assess mapping quality. Results of mapping should be assessed for quality and efficiency. Users can choose from a variety of programs including: *flagstat* (reports total number of reads, number of duplicate reads, percentage of reads mapped, etc.) from the SAMtools (Li et al., 2009b) package (available through Galaxy or for download at <http://sourceforge.net/projects/samtools/files/samtools/>) and *DepthOfCoverage* (reports read coverage per interval, gene, etc.) from the Genome Analysis Toolkit (McKenna et al., 2010; DePristo et al., 2011) software package (available through Galaxy or for download at <http://www.broadinstitute.org/gatk/download>). If mapping quality or efficiency is unacceptable (this will depend on the organism and sample being analyzed), refer to the manual for the algorithm used to improve mapping results.
4. Remove duplicate reads. To eliminate amplification bias introduced during PCR in the library preparation step, duplicated read artifacts should be removed (DePristo et al., 2011; Li et al., 2009b). This can be accomplished using *rmdup* from the SAMtools package (available through Galaxy or for download). Alternatively, users can use the *MarkDuplicates* algorithm from Picard, which is available for download at [http:](http://)

Note on removing duplicate reads:

There is much debate in the field as to whether duplicate reads should be removed or retained. We removed duplicates based on the read sequence in addition to the start and end coordinates, which results in removal of a limited number of reads due to the distribution of T-to-C conversion events. Users are urged to determine for themselves whether duplicate removal is appropriate based on T-to-C conversion rates and the number of PCR cycles used to generate cDNA libraries.

[//sourceforge.net/projects/picard/files/picard-tools/](https://sourceforge.net/projects/picard/files/picard-tools/). Recommended parameters for MarkDuplicates: [REMOVE_DUPLICATES=true] to prevent duplicate reads from being written to a new file.

3.3 *Generating binding sites*

For this step, code was generated in-house and is not available through Galaxy or for download as a stand-alone program. The original code was written in R, but pseudocode is provided (Fig. 3.9 and 3.11) so that users may implement the algorithm in any language.

1. Generate read clusters. Read clusters are defined as continuous stretches of nucleotides covered by at least one read harboring 0, 1, or 2 T-to-C conversion events only (see **Note 4**; Fig. 3.9). To differentiate between true RBP binding sites and noise in the data, read clusters that do not contain any T-to-C conversions (suggesting these RNA regions are not actually bound by an RBP) are treated as “low-confidence” and removed in a final filtering step (**Section 3.4**).
2. Refine binding sites. Some read clusters span hundreds of nucleotides and contain one or more distinct peaks indicative of unique RBP:RNA binding events (Fig. 3.10). To isolate distinct peaks within long read clusters, we fit a Gaussian curve (normal kernel function) to each read cluster and used the inflection points of this curve to define the boundaries of individual binding sites (Fig. 3.11). The bandwidth parameter for the normal kernel function was chosen to reflect the expected size of an RBP binding site (*e.g.* 21nt); however, this parameter can be empirically determined by the user and adjusted as needed (see **Note 5**). From this new set of refined binding sites, read coverage is determined by averaging the reads per million mapped read (RPM) values at each position across each refined binding site.

3. Calculate transcript abundance. Using perfectly mapped mRNA-seq reads obtained in **Section 3.2**, calculate transcript reads per million mapped reads per kilobase of transcript (RPKM; also called FPKM) using an established method such as Cufflinks (Trapnell et al., 2010) available through Galaxy or for download at <http://cufflinks.cbc.umd.edu/downloads/>. Additional methods and documentation describing best practices for quantifying transcript levels are readily available (Garber et al., 2011), so details will not be described here. Alternatively, published RPKM values of transcripts can be obtained and used in **Section 3.3**; however, (1) the sample conditions must be similar so that the published RPKM values are an accurate proxy for transcript levels in the samples used to generate gPAR-CLIP-seq libraries, and (2) published mRNA-seq libraries cannot be used for FDR calculation as the rate of sequencing error varies from machine to machine.

4. Normalize binding site RPM to transcript abundance. To allow comparison of RBP binding sites on different transcripts, binding site read coverage must be normalized by transcript abundance by dividing binding site RPM by the RPKM of the associated transcript and multiplying by 1,000 to account for the kilobase normalization of RPKM values. Some transcripts with no mRNA-seq reads contain gPAR-CLIP-seq binding site (typically very few; often with no T-to-C conversion events); these site are treated as “low-confidence” and removed in a final filtering step (**Section 3.4**).

3.4 Defining high-quality RBP binding sites from gPAR-CLIP-seq data

A small fraction of T-to-C mismatches in gPAR-CLIP-seq reads likely represent sequencing error instead of true RBP:RNA interaction events, so binding sites derived from this error need to be identified and removed. The general approach is to calculate an FDR for each

gPAR-CLIP-seq-derived binding site by performing binding site generation (**Sections 3.1-3.3**) using mRNA-seq reads with 0-2 T-to-C mismatches, which is a proxy for the rate of T-to-C sequencing error. This is followed by comparison of T-to-C conversion rates between gPAR-CLIP-seq-derived and mRNA-seq-derived binding sites, removal of low-confidence binding sites, and assessment of data reproducibility.

1. Generate mRNA-seq binding sites. Repeat steps in **Sections 3.1-3.3** (through step 2) using mRNA-seq reads and the same reference genome and algorithm parameters that were used for generating binding sites from gPAR-CLIP-seq data.
2. Calculate T-to-C conversion rates. For each gPAR-CLIP-seq- and mRNA-seq-derived binding site, calculate the T-to-C conversion rate as the number of reads with at least 1 T-to-C conversion event divided by the total number of reads covering at least 1 thymine.
3. Assign FDRs. Bin gPAR-CLIP-seq- and mRNA-seq-derived binding sites separately into groups based on total read coverage. Because total read coverage values are approximated by a negative binomial distribution, we recommend binning such that roughly equal numbers of binding sites are in each bin (**Fig. 3.12**). For each gPAR-CLIP-seq-derived binding site within each bin, calculate the proportion of mRNA-seq binding sites in that bin with a higher T-to-C conversion rate. This proportion represents the FDR for that gPAR-CLIP-seq binding site.
4. Filter off “low-confidence” binding sites. To determine a final set of “high-confidence” gPAR-CLIP-seq-derived binding sites, we removed binding sites that met any of the following criteria: (1) contain no T-to-C conversion events, (2) map to transcripts with no mRNA-seq reads, (3) have low RPM coverage, or (4) have an FDR above 1% (strict) or 5% (conservative).

5. Assess data reproducibility. If replicate gPAR-CLIP-seq libraries are generated, reproducibility can be assessed by calculating a correlation coefficient for binding site RPM values.

3.5 Calculating per-nucleotide crosslinking scores (CLSs)

In addition to identifying discrete, transcriptome-wide RBP binding sites, users can calculate a measure of crosslinking, or binding, strength on a per-nucleotide level (*see Note 6*). The general approach is similar to how FDR values are assigned to binding sites in **Section 3.4**.

1. A CLS for each transcriptomic uracil (represented by thymines in our cDNA sequencing libraries) is calculated as the number of reads covering that position that contain a T-to-C conversion event divided by the transcript RPKM and multiplied by a factor of 1,000 to account for the kilobase normalization in RPKM values.

3.6 Functional annotation of binding sites and their targets

After obtaining a set of “high-confidence” RBP binding sites across the transcriptome, binding sites and their mRNA targets can be further analyzed to obtain biologically functional information. Below are some common analysis tools. Specific information about how to run these tools and interpret the results is beyond the scope of this paper; these tools generally have helpful documentation available online.

1. Gene ontology term enrichment. Discover sets of terms describing the molecular functions, biological processes, and cellular compartments associated with mRNAs harboring RBP-binding sites. Tool is available online or for download at <http://www.geneontology.org> (Ashburner et al., 2000).

2. Gene set/pathway enrichment. Further characterize mRNAs harboring RBP-binding sites using comprehensive functional annotation tools such as DAVID, available through Galaxy or online at <http://david.abcc.ncifcrf.gov/> (Huang et al., 2009a,b), and g:Profiler, available through Galaxy or online at <http://biit.cs.ut.ee/gprofiler/> (Reimand et al., 2011, 2007), to identify enriched pathways, discover gene-disease associations, and identify enriched gene groups.
3. Primary sequence motif analysis. Identify putative functional sequence motifs using the MEME suite of sequence analysis tools, available through Galaxy, online, or for download at <http://meme.nbcr.net/meme/> (Bailey et al., 2009).
4. Secondary RNA structure analysis. Identify potential RNA secondary structure of interest using the ViennaRNA package of tools (*e.g.* RNAfold, RNAplfold) available for download at <http://www.tbi.univie.ac.at/RNA/> (Lorenz et al., 2011).
5. Conservation/homology analysis. Primary sequence conservation scores can be downloaded from the UCSC genome browser (Siepel et al., 2005; Pollard et al., 2010) and used to explore evolutionary conservation of calculated binding sites.

4 Notes

1. We recommend using the `[-eol]` parameter to ensure that the barcode is matched at the 3'-most end of the read, which is typically where barcodes are added. This ensures that matches to barcode sequences that occur randomly throughout the genome, and therefore might appear elsewhere in a read, are not mistaken for the true barcode.
2. A read length threshold of 15nt (`[-l 15]`) was chosen for reads being mapped to the *S.*

cerevisiae genome, but can be optimized depending on the size of the genome being used. Users should also check their sequences for 5' "N" nucleotides, which can be trimmed before mapping.

3. Users should also remove read artifacts if they are: homopolymers, missing 3' adapter, 5'-3' adapter ligation products, or 5'-5' adapter ligation products. Also note that only reads in FASTQ format may be analyzed using the Quality Filter and Quality Trimmer algorithms, as FASTA-formatted sequences do not contain quality information. The [-q 30] and [-t 30] parameters both correspond to minimum Phred quality scores to keep reads/nucleotides. A Phred quality score of 30 indicates a base call accuracy of 99.9%. Users may choose other Phred quality score thresholds if they desire.
4. Low-frequency incorporation of 4sU into nascent mRNAs induces a mis-pairing of guanine to 4sU during reverse transcription that manifests as T-to-C mismatches to the genome (*i.e.* a cytosine is sequenced where there should be a thymine). As of this publication, there are no published mapping algorithms that allow for differential treatment of different types of mismatches to the genome. Therefore, mapping algorithm parameters must be set to allow for multiple mismatches. Only reads with 0 or 1-2 T-to-C mismatches will be used in subsequent steps.
5. Some secondary analyses performed on binding sites after Gaussian curve fitting will be affected by the choice of bandwidth parameter. For example, calculating the average binding site length after Gaussian curve fitting will result in an average length close to the bandwidth parameter chosen. For most secondary analyses, however, the choice of bandwidth parameter will not have an effect on results.
6. Calculating a crosslinking score is possible because gPAR-CLIP-seq T-to-C conversions only occur when a 4sU is within a few angstroms of an amino acid side chain (Morrison,

1990). Because a single RBP:RNA interaction site spans many nucleotides, calculating a score for each thymine within the binding site could give insight into which thymines are the most biologically important for RBP:RNA interactions. For this calculation, we assume that rate of incorporation of the ribonucleoside analog into nascent transcripts during transcription is uniform across the length of the transcript.

3.4 Discussion and concluding remarks

3.4.1 Discussion of findings

RNP complexes exhibit dynamic properties that are sensitive to environmental conditions. For example, granules containing stalled translation pre-initiation complexes are formed under stress but rapidly dissociate when the cell returns to favorable conditions (Grousl et al., 2009). Despite insight into how particular RNP complexes are affected by stress, global effects of stress on all RBP-RNA interactions have until now remained unexplored. We detect reproducible changes in occupancy for 38% of 3'UTR crosslinking sites on non-translating mRNAs under glucose or nitrogen starvation conditions: loss of RBP occupancy at RBP crosslinking sites was a phenomenon common to both glucose and nitrogen stress conditions, while more distinct sets of crosslinking sites increased RBP occupancy (Fig. 3.5C).

In our current work, we limited our gPAR-CLIP analyses to protein-RNA interactions residing in non-translated RNPs (Fig. 3.1A; Materials and methods), which mediate important functions for mRNA translation, localization, and degradation (Dreyfuss et al., 2002). Because we have no information on the identities of the RBPs or their distribution in the sucrose gradient, we cannot distinguish whether the changes in RBP coverage represent changes in RBP binding and/or distribution. This is particularly relevant in glucose or ni-

trogen starvation, as many RBPs redistribute under these conditions (Balagopal and Parker, 2009). Future comparative gPAR-CLIP analyses on both non-translating RNP and translating RNPs in stress conditions will distinguish changes in RBP binding versus changes in RBP localization.

RNAs are capable of forming complex two- and three-dimensional structures, and some RBPs are known to recognize such structural motifs. For example, She2p mediates the localization of several bud-localized transcripts during cell division by recognizing and binding to specific stem-loop structures in mRNAs (Chartrand et al., 1999; Gonzalez et al., 1999). Examination of the structural properties of our global RBP crosslinking sites revealed a preference for single-stranded regions, which agrees with previous reports of crosslinking sites of the RNA-binding protein FUS occurring at single-stranded regions directly adjacent to the FUS RNA recognition motif (Rogelj et al., 2012). Unpaired loop and bulge regions can be unstructured or form tertiary structural modules, both of which can be readily recognized by RBPs. In contrast, double-stranded RNAs, in general, do not provide good platforms for RBP binding: structured RNA regions captured by gPAR-CLIP generally had low CLS values (Fig. 3.4), likely resulting from crosslinking and/or RNase T₁ cleavage inefficiency. In structured regions, 4-thiouridines are more likely to be locked in U:A or U:G pairing, preventing crosslinking to proteins. In addition, structured regions are less accessible to RNase attack during sequencing fragment preparation, resulting in under-representation in gPAR-CLIP libraries. Nevertheless, despite their low crosslinking efficiencies, Ts in double-stranded, paired RNA regions show extremely high conservation compared to Ts with no crosslinking evidence. These data indicate that RNAs with high secondary structure are evolutionarily conserved and can serve as functional, secondary structure motifs recognized by select RBPs.

RBP binding sites functioning as *cis*-regulatory elements are expected to be under pu-

rifying selection. We identified a substantial fraction (35%) of conserved elements in UTRs overlapping RBP crosslinking sites. This represents an underestimation because RBPs and RNAs that are not expressed under our experimental conditions or that fail to crosslink will not be captured. Although crosslinking sites in general are more highly conserved than non-crosslinking sites in UTRs, many sites are not well conserved and might represent species-specific *cis*-regulatory elements that allow adaptation to different environments and stressors.

A preference of RBP binding to 3'UTRs observed in this study and others (Hogan et al., 2008; Riordan et al., 2011) is consistent with the function and evolution of 3'UTRs as major sites for post-transcriptional regulation. Unlike protein-coding regions, 3'UTRs do not directly engage ribosomes during translation and therefore provide accessible platforms for RBP binding and RNP assembly. One important aspect of gene regulation is combinatorial control, which allows a single gene to be controlled by more than one regulator. In our study, 23% of all nucleotides in annotated 3'UTRs were located within RBP crosslinking sites, corresponding to an average of 1 crosslinking site, on average 23 nucleotides long, in every 100 nucleotides. For a median-sized yeast 3'UTR that is 166 nucleotides long (Ozsolak et al., 2010), there are, on average, 2 RBP crosslinking sites, suggesting that most yeast genes are subject to combinatorial post-transcriptional regulation. Since *S. cerevisiae* lacks post-transcriptional regulation by the highly conserved and pervasive microRNA regulatory pathway, combinatorial regulation by RBPs may play a more prominent role than in organisms with small RNA-mediated post-transcriptional gene regulation.

Unlike focused interrogation of individual RBPs, gPAR-CLIP does not directly identify the RBP that recognizes each crosslinking site. To enable identification of primary sequence motifs recognized by individual RBPs, we searched gPAR-CLIP crosslinking sites located on target mRNAs identified *in vitro* by RIP-Chip for 29 RBPs (Gerber et al., 2004; Hogan et al.,

2008) and identified 39 motifs for 15 RBPs ([Data file 7](#); Materials and methods). Notably, 35 of the sequence motifs derived by gPAR-CLIP differed significantly from previous motif predictions, which were based on scanning whole transcript sequences for enriched k-mers. This discrepancy between primary sequence motifs identified by our gPAR-CLIP data and previous predictions illustrates the potential utility of deriving motifs based on direct *in vivo* evidence of RBP-RNA interactions, which narrows the search space to enhance the signal of *bona fide* primary sequence recognition elements.

Our study provides a comprehensive map of RBP crosslinking sites across the budding yeast non-translating mRNA transcriptome and for the first time describes the dynamics of mRNA-RBP binding under normal and nutrient-limited growth conditions. Delineating *in vivo* sites of RBP binding will aid in directing future studies for identification of sites responsive to environmental or genetic perturbations, refinement of primary sequence and secondary structural elements recognized by specific RBPs, and elucidation of the complex network of regulatory processes that contribute to regulation of expression of each individual mRNA. gPAR-CLIP is readily applicable to other organisms for profiling global RNA-protein interactions underlying post-transcriptional regulation and the effects of environmental perturbations upon these interactions.

3.4.2 Investigating PTGR using gPAR-CLIP-seq

To catalog footprints of yeast RBPs *in vivo*, we designed a novel methodology to isolate all mRNA fragments bound by RBPs for deep sequencing. From computational analysis of these data, I confirmed that RBPs preferentially recognize 3'UTRs. Prior to our study, it was only hypothesized that 3UTRs served as the major site of PTGR since many studies of individual RBPs or small RNAs showed interaction with *cis*-regulatory elements in 3UTRs.

Our detection of multiple RBP binding sites on each mRNA suggests that many yeast genes are under combinatorial control by multiple RBPs, despite having 3UTRs that are only a few hundred bases long. In more complex eukaryotes, where 3UTRs can easily be thousands of nts in length, the opportunity for combinatorial control increases greatly. In yeast, we show that 35% of conserved UTR blocks are bound by RBPs, indicating that these sites are under positive selection because they are important platforms for RBP:RNA interactions. Included in these conserved elements are regions of predicted RNA double-strandedness, suggesting that RNA secondary structures are evolutionarily conserved and can serve as functional RBP binding motifs. Until our study, RNA secondary structure was known to be required for interaction of target mRNAs with a small subset of RBPs that have been studied. Now we have empirical evidence that RNA secondary structure is a much more pervasive feature of RBP:RNA interactions than originally thought, and that these features are also under positive selection as much as primary sequence motifs directing RBP:RNA interactions.

One of the advantages of performing global PAR-CLIP-seq is that we now have a foundational database of *cis*-regulatory regions in a single system (budding yeast) that can easily be manipulated with a range of genetic and environmental perturbations, allowing us to assess changes in multiple PTGR pathways at once. In addition, other investigators studying a single RBP of interest can harness the gPAR-CLIP-seq datasets to interrogate *in silico* how binding of their RBP might change under a variety of conditions without actually performing time-consuming experiments.

3.4.3 Bioinformatic challenges

One of the key challenges in analyzing PAR-CLIP-seq data is accounting for the T-to-C mismatches that appear in the sequenced reads. We know that spurious sequencing

errors occur, especially towards the 3' ends of reads as the fidelity of the ligation reactions decreases (Kircher et al., 2009; Minoche et al., 2011). Thus, there is a low frequency of T-to-C error that occurs in addition to the expected UV crosslinking-induced T-to-C transition events. How can we distinguish between the two types of events? The most direct way is by using the frequency of reads covering a particular U nucleotide position that contain a U or a C. When the nucleoside analogs are incorporated into nascent transcripts, the reverse transcription and amplification steps only further increase the number of reads containing evidence of the crosslinking event. After the reads have been amplified, the sequencing by ligation reactions occur which then introduce error, most often at the 3 ends of reads, on a read-by-read basis. Thus, the rate of T-to-C mismatches in reads due to 4sU incorporation is higher than “background” sequencing error T-to-C rates. According to a study by Nakamura et al. (2011), mismatch rates across Illumina reads mapping to genomes from organisms with varied GC% (33-68%) show no significant enrichment for T-to-C errors compared to T-to-A or T-to-G; these results are confirmed in additional studies (Meacham et al., 2011; van Gurp et al., 2013). Therefore, I chose to estimate background T-to-C rates for the gPAR-CLIP-seq data from the mRNA-seq data, for which nearly all T-to-C mismatches should be a result of sequencing error. **Fig. S3.3C** confirms that the higher read coverage of an mRNA-seq-derived cluster, the higher the T-to-C error rate becomes; therefore, it was also necessary to bin gPAR-CLIP-seq- and mRNA-seq-derived read clusters based on total read coverage in order to assign an FDR value.

Integration of mRNA-seq data with g/PAR-CLIP-seq data is highly encouraged. In our work and others work (Uren et al., 2012; Comoglio et al., 2015), it is noted that comparing binding site read coverage across different transcripts is meaningless unless the abundances of the transcripts are known. For example, if a binding site for RBP X on Transcript A has 100 RPM coverage, and a site on Transcript B has 20 RPM coverage, one might suspect that

RBP X binding at Transcript A is 5-fold greater than the site on Transcript B. If Transcript A expression is actually 50-fold greater than that of Transcript B, then RBP X is actually binding to the site on Transcript B more efficiently (10-fold greater site RPM normalized to transcript abundance). Coincidentally, mRNA-seq data also is required for the FDR analysis, which makes it all the more important that both CLIP-seq and mRNA-seq libraries are sequenced to explore mechanisms of RBP:RNA interactions.

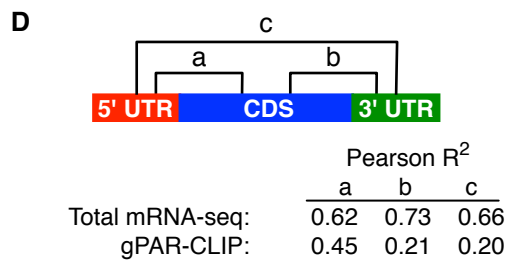
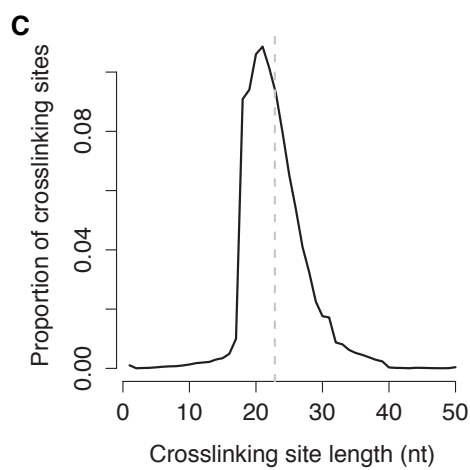
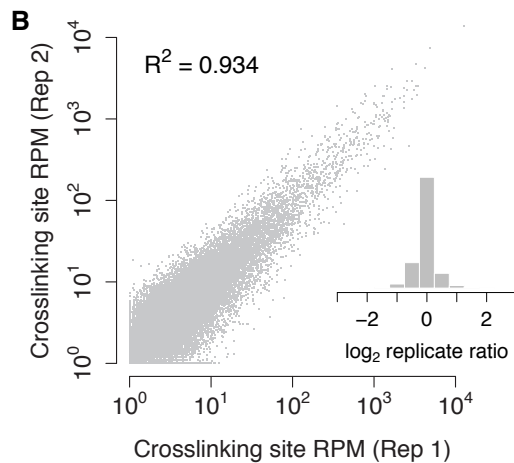
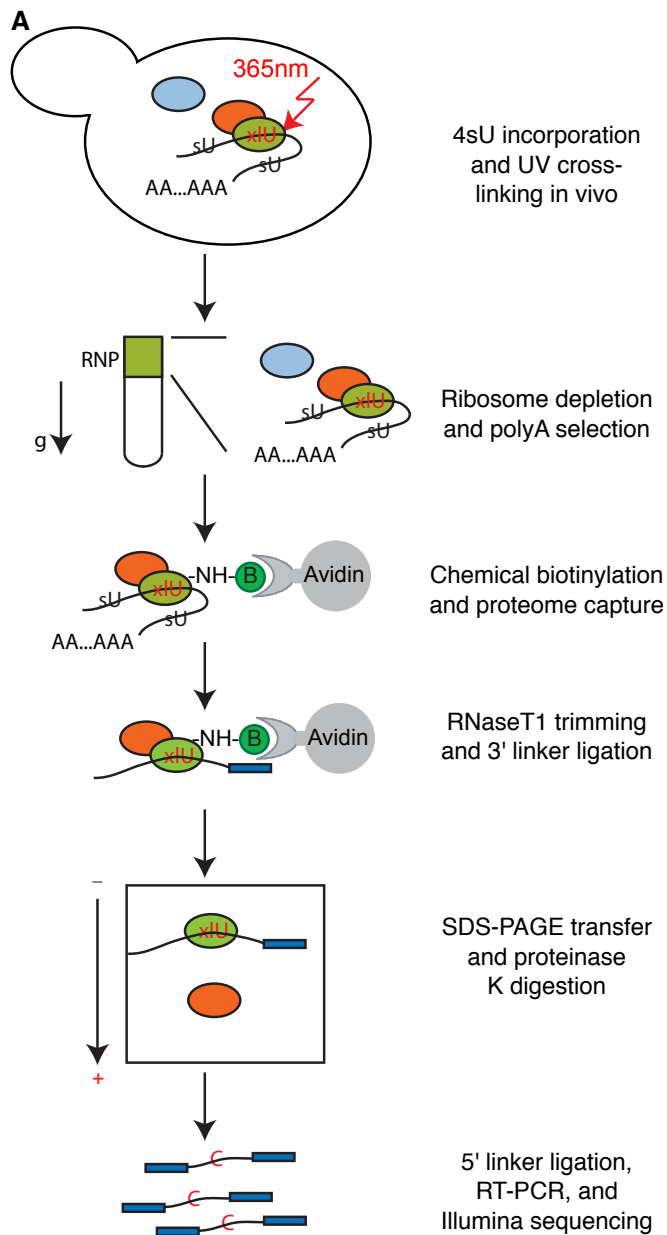
3.5 Data availability

gPAR-CLIP-seq and mRNA-seq data are available through the NCBI Gene Expression Omnibus using series entry GSE43747.

3.6 Acknowledgements

Ting Han performed a majority of the molecular biology experiments associated with this study and developed the gPAR-CLIP-seq experimental methods. Daniel Klionsky and Lois Weisman generously shared yeast strains. Andy Kong, Zhi John Lu, and Yu-Cheng Yang contributed to RNA structure analysis. These manuscripts benefitted from thoughtful discussions with Allison Billi, Kris Gunsalus, Daniel Klionsky, James Taylor, Natasha Weiser, Lois Weisman, and Danny Yang. Manuscript was written by MAF, TH, and JKK.

This work was supported by the National Science Foundation Open Data IGERT grant 0903629 (MAF), the National Institutes of Health grant GM088565 (JKK), and the Pew Charitable Trusts (JKK).



Replicates	5' UTR	CDS	3' UTR
-ribosome mRNA-seq:	0.990	0.994	0.988
gPAR-CLIP:	0.934	0.935	0.968

Figure 3.1: (A) Schematic of the gPAR-CLIP protocol. (B) Reproducibility of crosslinking sites generated from replicate gPAR-CLIP libraries prepared from yeast grown in synthetic defined media (abbreviated as WT gPAR-CLIP hereafter). Pearson correlation coefficient is indicated. Inset: distribution of \log_2 crosslinking site RPM ratios between replicates. Replicate error $\sigma=1.3$ -fold. (C) Length distribution of crosslinking sites in WT gPAR-CLIP libraries. Dotted line: average crosslinking site length of 23 nt. (D) Pearson correlation coefficients of total mRNA-seq and gPAR-CLIP read coverage between 5'UTR, CDS, and 3'UTR regions as well as correlation coefficients of ribosome depleted (-ribosome) mRNA-seq and gPAR-CLIP read coverage between replicate WT libraries.

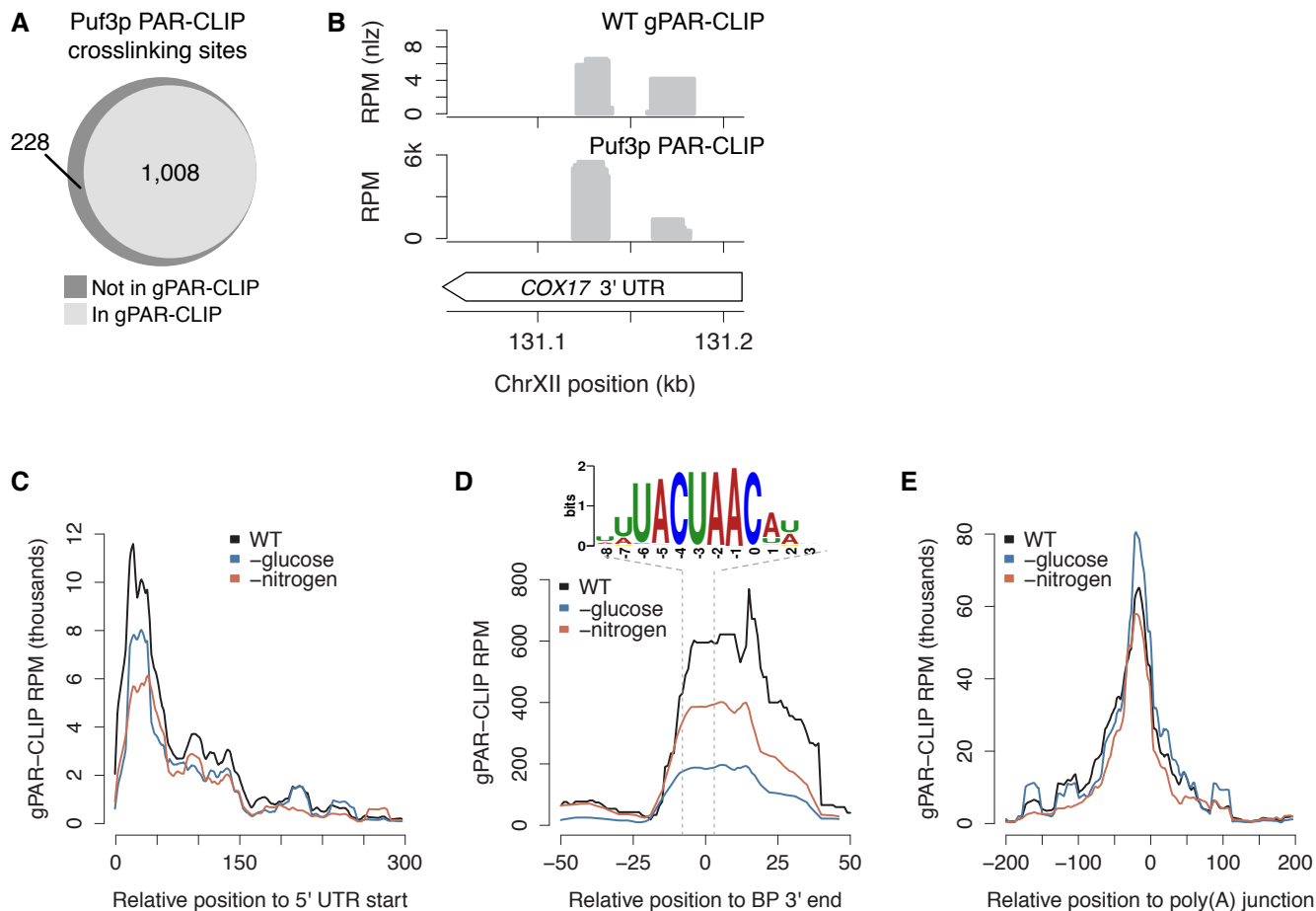


Figure 3.2: (A) Overlap of crosslinking sites identified in Puf3p PAR-CLIP and WT gPAR-CLIP. Puf3p PAR-CLIP crosslinking sites with $>1\%$ T-to-C conversion rate (see Additional file 8) were considered captured by gPAR-CLIP if at least 50% of their nts overlapped with a WT gPAR-CLIP crosslinking site with $FDR < 1\%$. (B) Identification of known Puf3p binding sites on *COX17* mRNA in WT gPAR-CLIP and Puf3p PAR-CLIP. (C-E) Aggregate gPAR-CLIP crosslinking site coverage of the first 300 nt of 2,626 annotated 5'UTRs (C), 51 annotated ribosomal gene introns centered at the branch point (BP) 3' end (D), and 4,241 3'UTRs centered on the poly(A) junction (E).

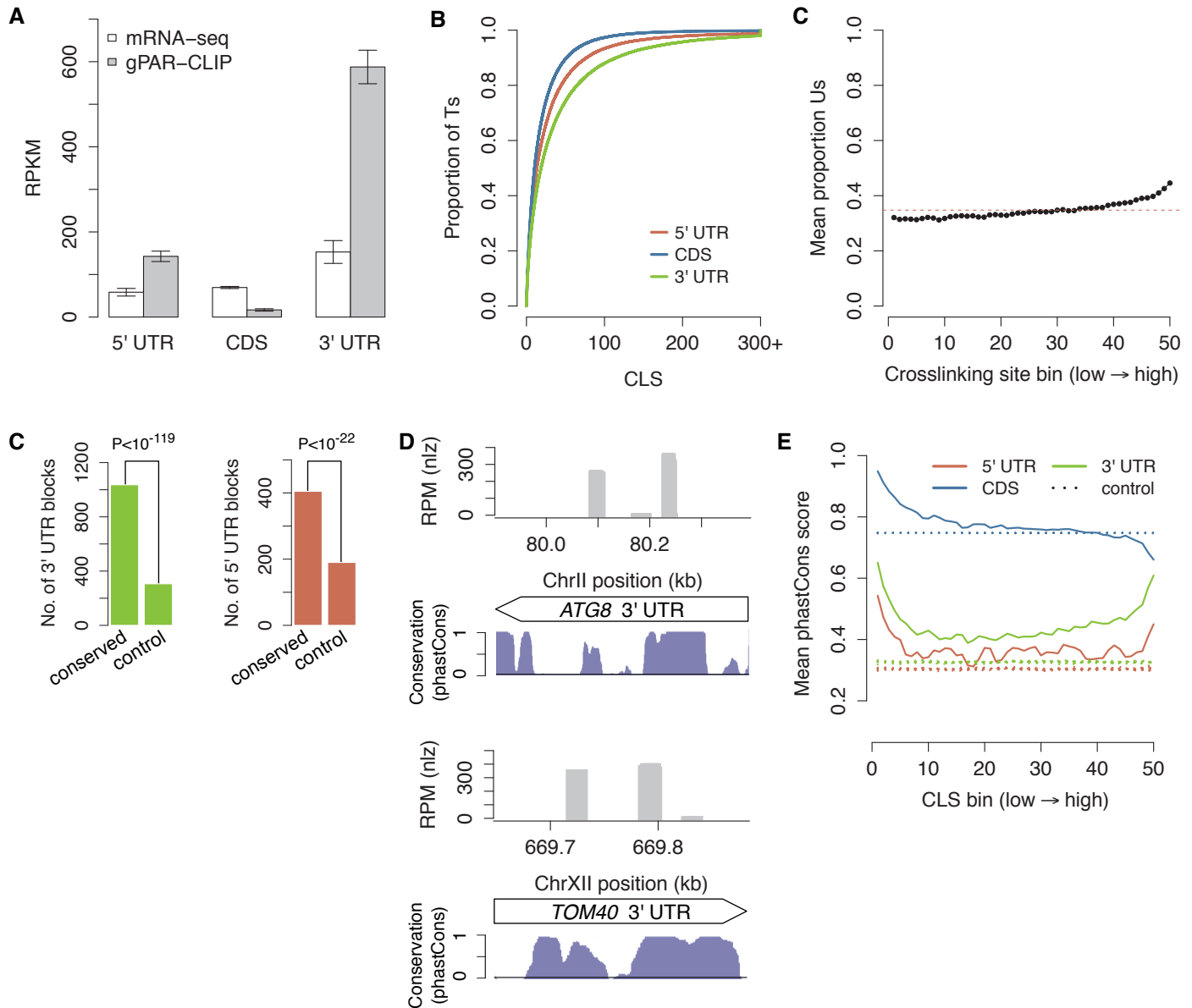


Figure 3.3: (A) Average ribosome depleted mRNA-seq and gPAR-CLIP read distributions across 5'UTR, CDS, and 3'UTR regions for all libraries. Error bar: ± 1 standard deviation. RPKM: reads per million mapped reads per kilobase. (B) Cumulative distribution of CLS values from WT libraries. (C) Proportion of Ts in crosslinking site binned by crosslinking site coverage (RPM). Dotted red line indicates average T content of all crosslinking sites. (D) Number of conserved blocks in 3' and 5'UTRs overlapping 100% with WT gPAR-CLIP crosslinking sites (χ^2 p -values indicated). Control blocks were randomly generated within 3' and 5'UTRs to match the number and size of conserved blocks. (E) Two major gPAR-CLIP crosslinking sites in *ATG8* 3'UTR (top) and *TOM40* 3'UTR (bottom) overlapping conserved blocks. (F) Mean phastCons scores for Ts ranked and binned by CLS. Control lines represent mean phastCons scores of randomly ranked and binned Ts with no CLS, repeated 10 times.

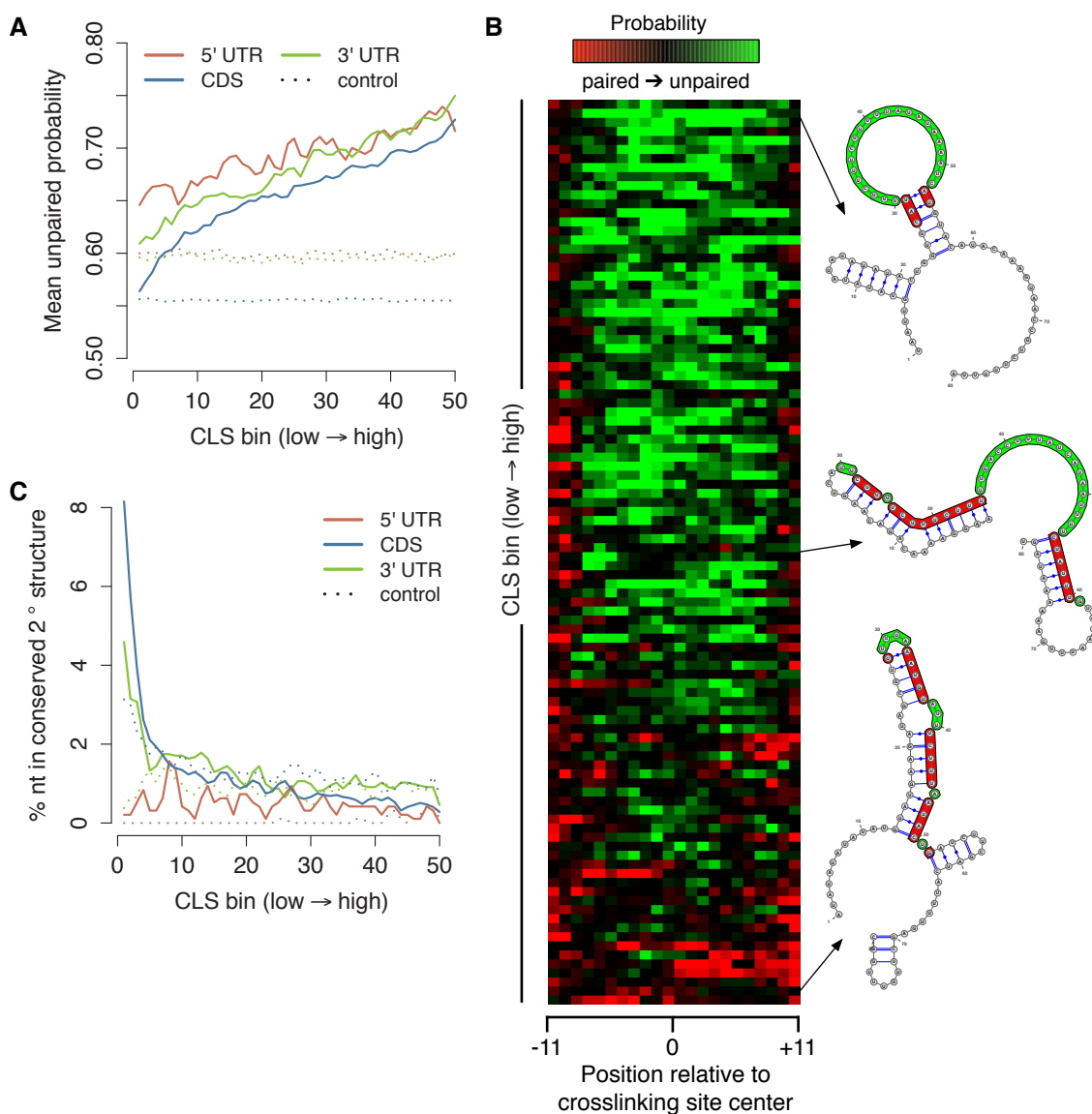
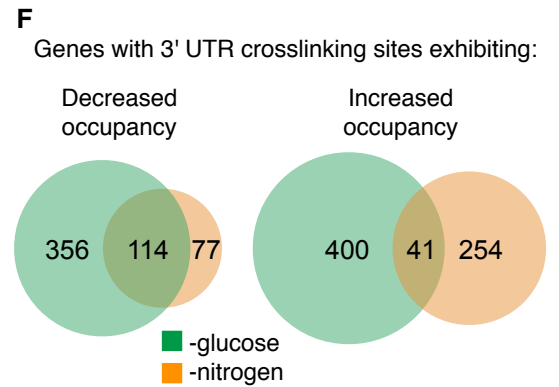
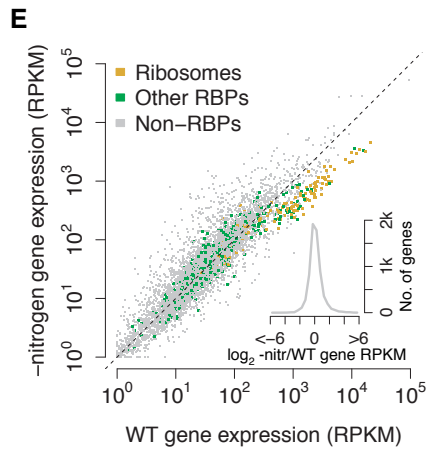
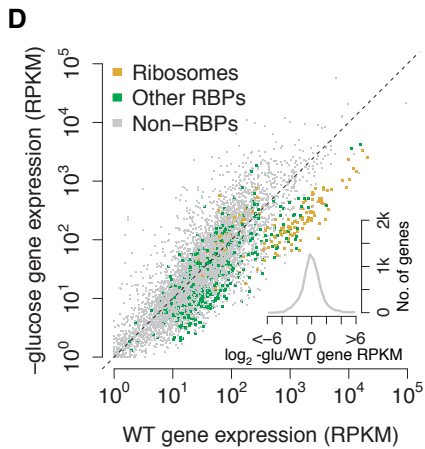
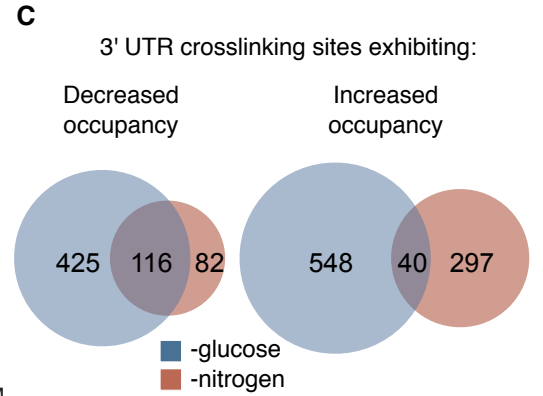
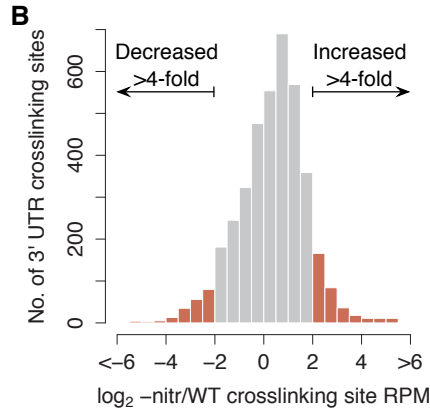
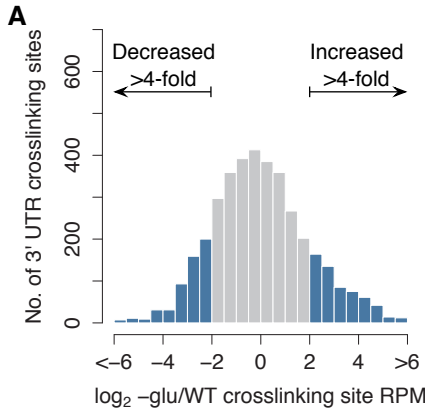
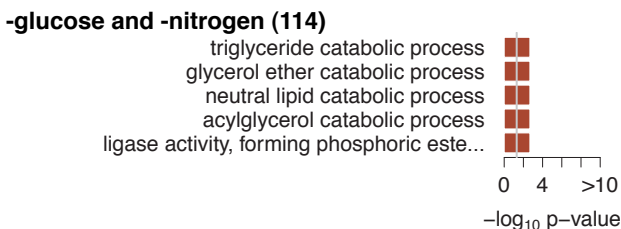
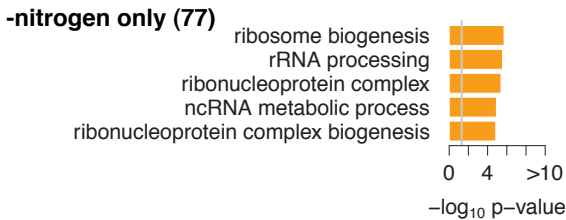
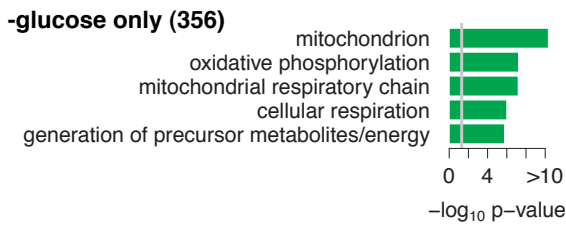


Figure 3.4: (A) Mean unpaired probability scores for Ts ranked and binned by CLSs. Control lines represent mean unpaired probability of randomly ranked and binned Ts with no CLS. Pearson correlation coefficients: 5'UTR $R^2=0.933$, CDS $R^2=0.976$, 3'UTR $R^2=0.986$. (B) Crosslinking site pairedness visualized as a heatmap. Columns represent nucleotide positions within crosslinking sites. Rows represent average unpaired probability for 100 crosslinking sites in that bin. Select secondary structure predictions from low, middle, and high CLS regions are indicated with the crosslinking site colored. (C) Percentage of Ts ranked and binned by CLSs in conserved secondary structural elements as defined by RNAz. Control lines represent percentage of randomly ranked and binned Ts with no CLS in conserved secondary structural elements.



G Genes with 3' UTR crosslinking sites exhibiting **decreased** occupancy in:



H Genes with 3' UTR crosslinking sites exhibiting **increased** occupancy in:

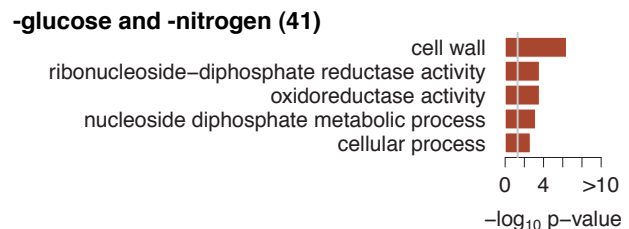
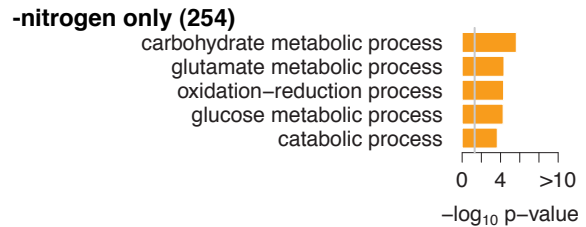
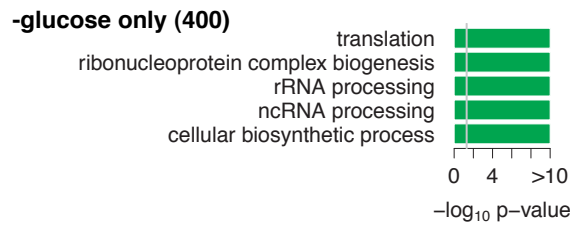


Figure 3.5: (A-B) Changes in 3'UTR crosslinking site coverage upon glucose (A) or nitrogen (B) starvation. Standard deviations of intra-replicate variation: WT 1.31-fold; glucose starvation 1.24-fold; nitrogen starvation 1.15-fold. (C) Overlap of 3'UTR crosslinking site changes affected by glucose or nitrogen starvation conditions. (D-E) Global changes in mRNA abundance upon glucose (D) or nitrogen (E) starvation. (F) Overlap of mRNAs with 3'UTR crosslinking site changes affected by glucose or nitrogen starvation conditions. (G-H) Enriched GO terms for mRNAs with 3'UTR crosslinking sites with decreased (G) or increased (H) RBP occupancy upon glucose or nitrogen starvation or both. Grey lines indicate p -value of 0.05.

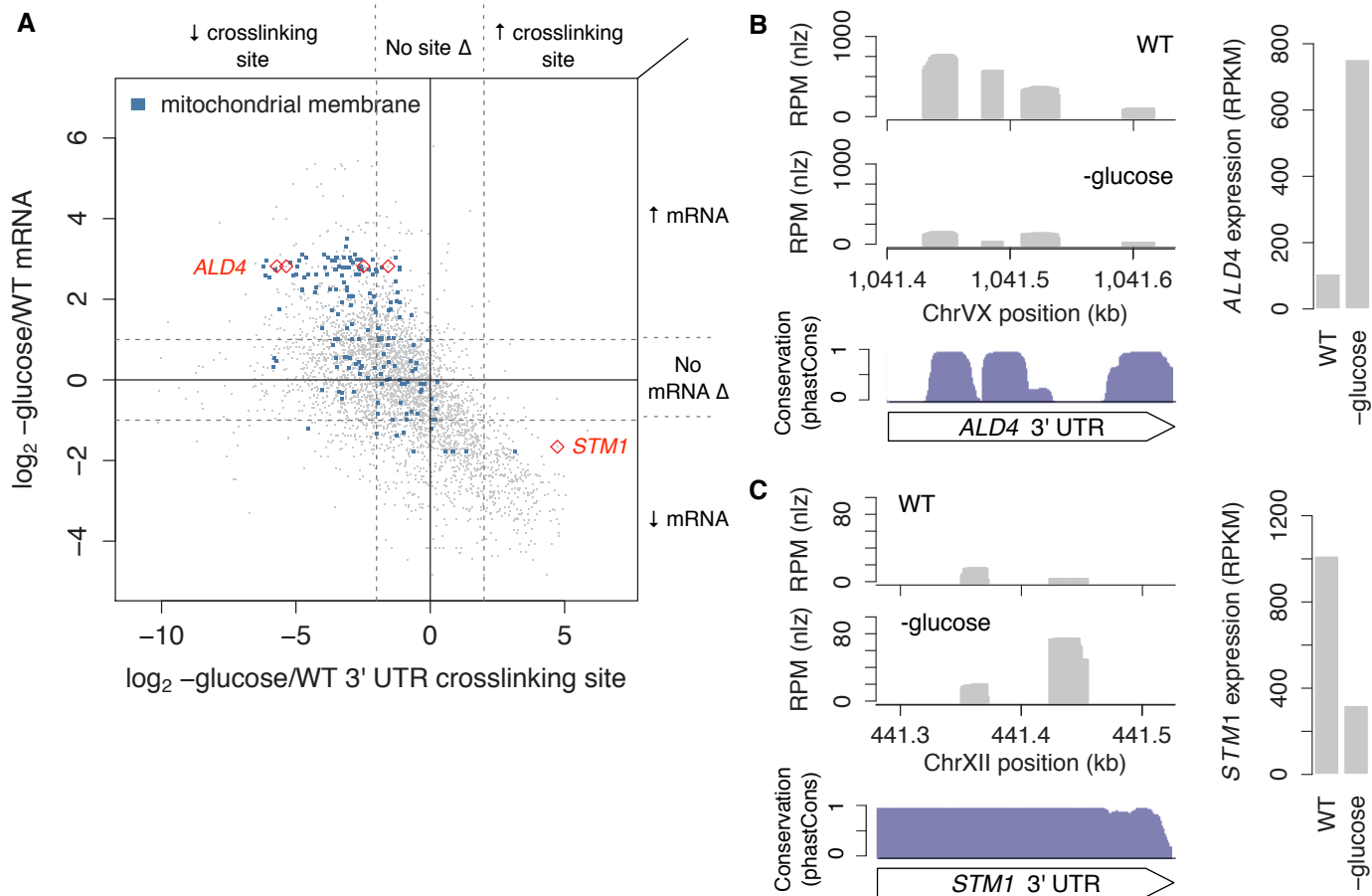


Figure 3.6: (A) Changes in 3'UTR crosslinking site coverage versus changes in the corresponding mRNA upon glucose starvation. Crosslinking sites on genes annotated with "mitochondrial membrane" GO term are colored blue. Dotted lines indicate ≥ 4 -fold changes in crosslinking site coverage (vertical) or ≥ 2 -fold change in mRNA expression (horizontal). (B) *ALD4* 3'UTR contains four crosslinking sites that decrease 2- to 8-fold in RBP occupancy upon glucose starvation and overlap with conserved blocks (red diamonds in (A)). *ALD4* mRNA expression is up-regulated upon glucose starvation. (C) *STM1* 3'UTR contains one crosslinking site that increases in coverage upon glucose starvation (red diamond in (A)). *STM1* mRNA expression is down-regulated upon glucose starvation.

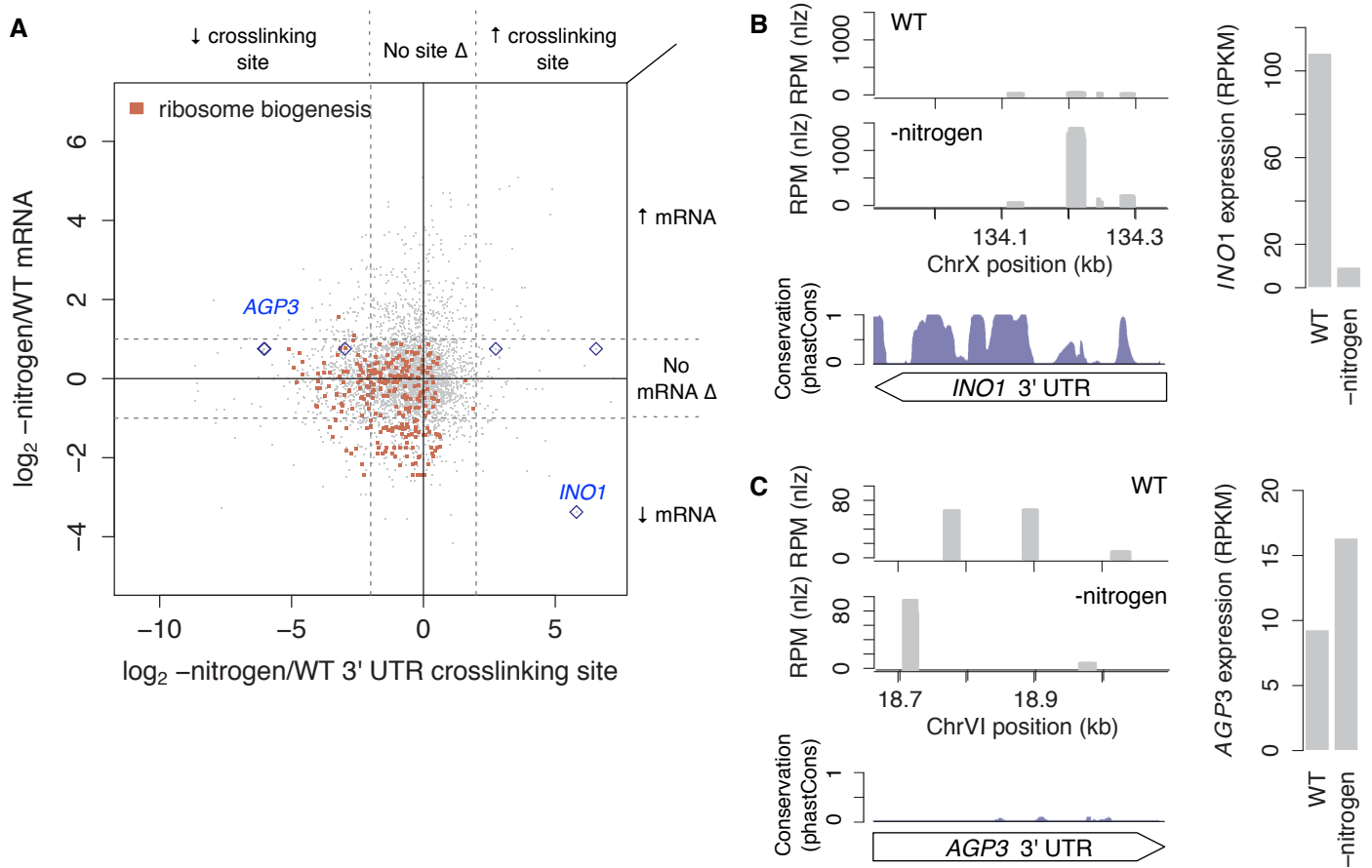


Figure 3.7: (A) Global changes in 3'UTR crosslinking site coverage versus changes in the corresponding mRNA upon nitrogen starvation. Crosslinking sites on genes annotated with “ribosome biogenesis” GO term are colored red. Dotted lines indicate ≥ 4 -fold changes in crosslinking site coverage (vertical) or ≥ 2 -fold change in mRNA expression (horizontal). (B) *INO1* 3'UTR contains one crosslinking site that increases in coverage upon nitrogen starvation and falls within a conserved block (blue diamond in (A)). *INO1* mRNA expression is down-regulated upon nitrogen starvation. (C) *AGP3* 3'UTR contains three crosslinking sites that are lost and two crosslinking sites that appear upon nitrogen starvation (blue diamonds in (A)). *AGP3* mRNA expression is up-regulated upon nitrogen starvation.

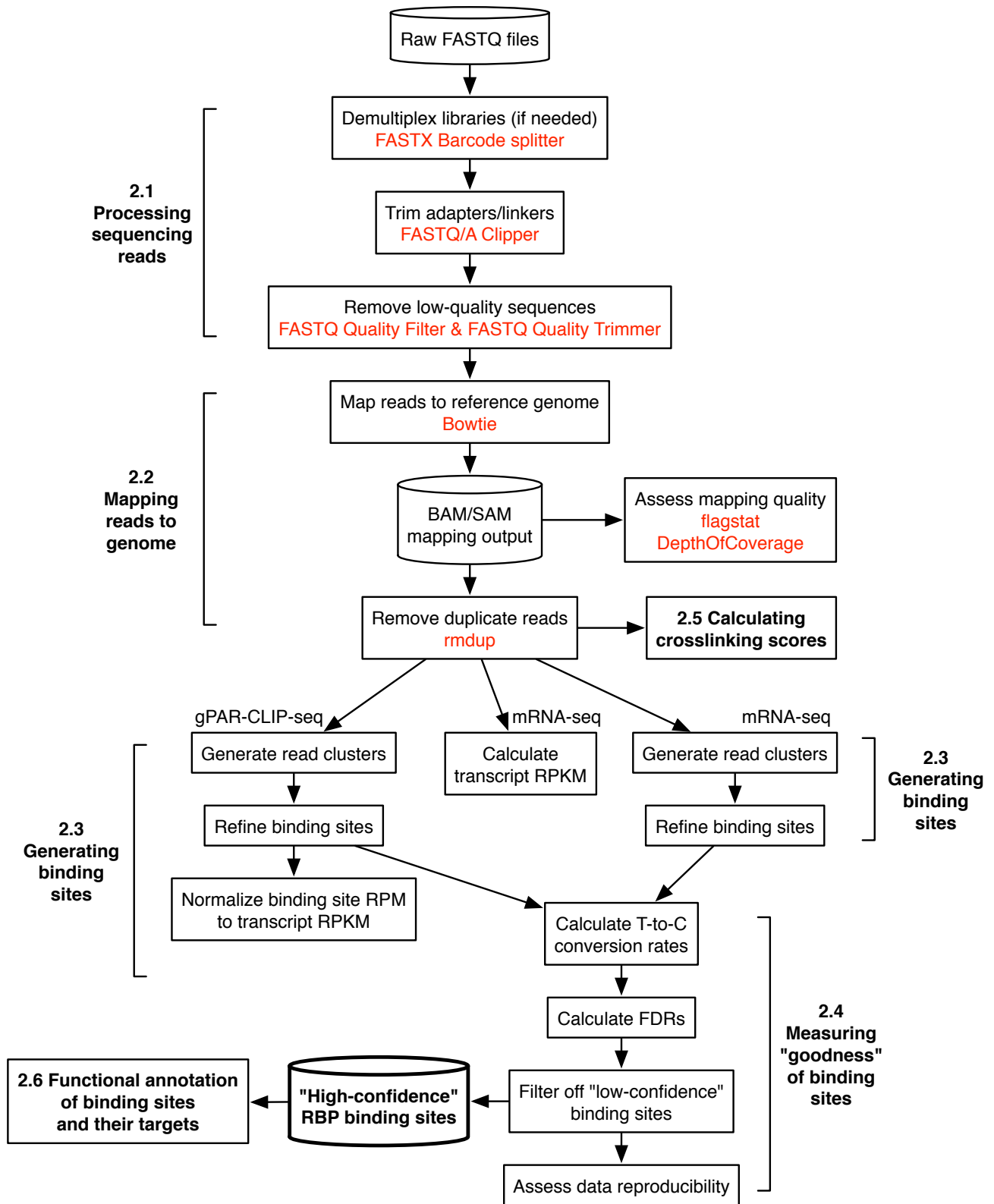


Figure 3.8: Steps corresponding to the bioinformatic analysis pipeline are displayed as a flowchart. Programs available for download or through Galaxy are highlighted in red.

Algorithm 1 Generate read clusters

```

1: for each chromosome chr do
2:   for each read rd on chromosome chr do
3:     for each genomic position pos in rd do
4:       readcount(pos) ← readcount(pos) + 1
5:     end for
6:   end for
7:   i ← 1                                     ▷ Initialize clusterID to 1
8:   for each genomic position pos on chromosome chr do
9:     if readcount(pos) > 0 then
10:      clusterID(pos) ← i
11:      while readcount(nextpos) > 0 do
12:        clusterID(nextpos) ← i
13:      end while
14:      i ← i + 1
15:    end if
16:  end for
17: end for
end

```

Figure 3.9: A read cluster is defined as a continuous stretch of nucleotides covered by at least 1 gPAR-CLIP-seq read with 0-2 T-to-C conversion events. Input: chromosome, start position, end position, and strand information for reads mapping to the genome with 0-2 T-to-C mismatches. Output: a list of every chromosomal position, how many unique reads map to each position, and a cluster ID denoting the distinct read cluster to which each position belongs. Users can obtain genomic start and end coordinates of each read cluster by calculating the minimum and maximum position for each cluster ID. In the case of stranded sequencing libraries, separately analyze reads mapping to the plus and minus strands of the reference.

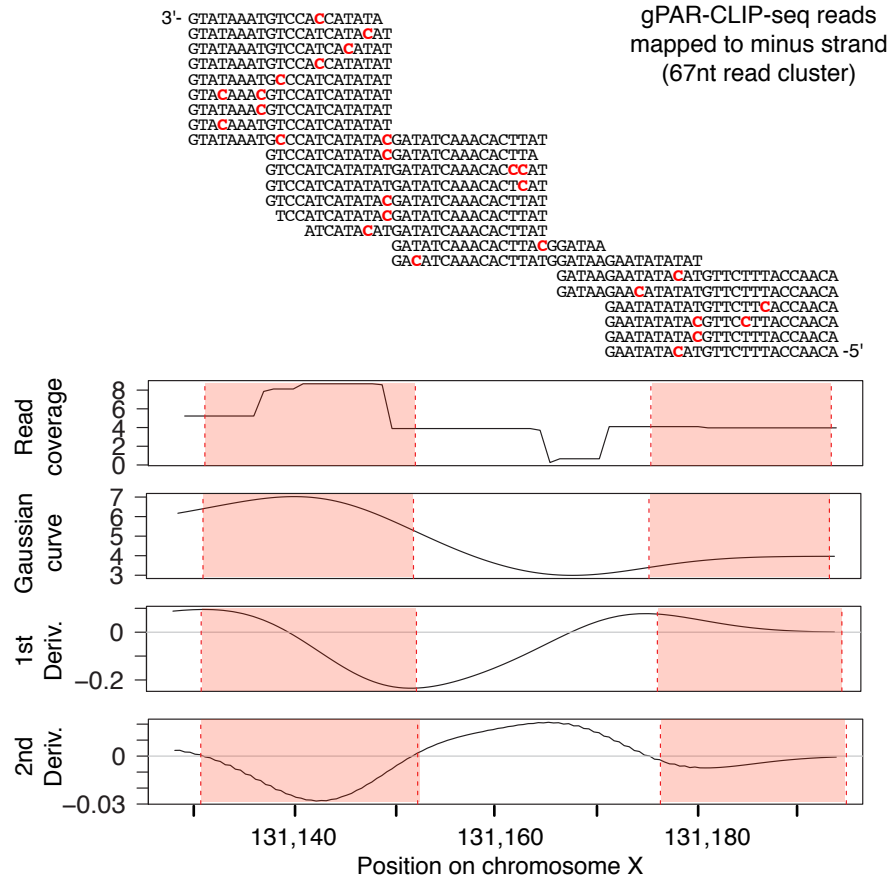


Figure 3.10: Shown are sequencing reads mapping to the minus strand of chromosome X with T-to-C conversion events highlighted in red. These read form a read cluster 67nt long. Two distinct binding events are determined by identifying the inflection points of the second derivative of the fitted Gaussian curve. Red shaded blocks indicate the new, refined binding sites.

Algorithm 2 Refine binding sites

```

1: for each read cluster  $rc$  do
2:    $f(rc) \leftarrow$  Gaussian curve fit to read count data across  $rc$ 
3:   Calculate  $f'(rc)$  (first derivative)
4:   Calculate  $f''(rc)$  (second derivative)
5:   for each position  $pos$  in  $rc$  do
6:     if  $f''(pos) = 0$  and slope of  $f''(pos) < 0$  then
7:       append  $pos$  to startcoordinates
8:     else if  $f''(pos) = 0$  and slope of  $f''(pos) > 0$  then
9:       append  $pos$  to endcoordinates
10:    end if
11:  end for
12: end for
end

```

Figure 3.11: Read cluster start and end coordinates are refined to break long read clusters into smaller, distinct peaks. The general approach is to fit a Gaussian curve to the read counts across each binding site and define start and end coordinates as the inflections points of this curve. Input: output from Algorithm 1. Output: refined binding site boundary start and end coordinates. While read count data are discrete, the data are treated as continuous for this analysis. In lines 6 and 8, the points at which $f''(pos)=0$ will likely not be integers, so users should round to the nearest integer to get chromosomal coordinates.

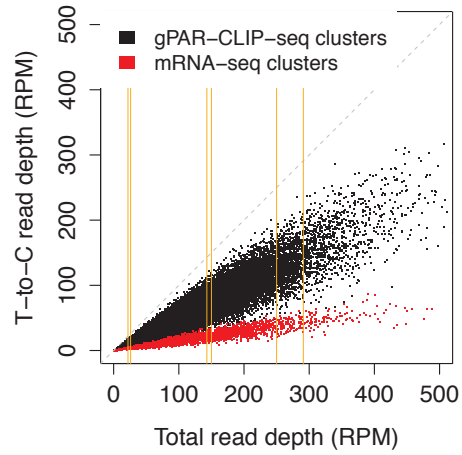


Figure 3.12: Plotted are total read coverage versus T-to-C read coverage of 50,000 random gPAR-CLIP-seq (black) and 50,000 random mRNA-seq (red) read clusters. gPAR-CLIP-seq read clusters were grouped into 50 bins with $\sim 1,000$ clusters in each bin. The 5th, 40th, and 49th bins are demarcated by orange lines and show that although the bins contain roughly the same number of gPAR-CLIP-seq binding sites, they cover a varied range of total read depth RPMs.

condition	protocol	4sU added?	Ribosome depletion?	total raw reads		total high quality reads			total mapped reads			total mapped reads (1 T>C)		% of high quality reads	
				reads	quality reads	0 mm	1 mm	2 mm	reads	reads	reads	reads	reads	%	%
WT	gPAR-CLIP	Yes	Yes	10,345,660	9,892,996	855,723	5,652,529	1,080,794	5,396,767	692,138	54.6%	692,138	7.0%		
WT	gPAR-CLIP	Yes	Yes	13,354,315	12,354,153	1,020,368	7,138,708	1,184,089	6,848,424	728,086	55.4%	728,086	5.9%		
-glucose	gPAR-CLIP	Yes	Yes	8,698,657	8,571,862	1,186,644	4,130,861	683,064	3,808,408	275,737	44.4%	275,737	3.2%		
-glucose	gPAR-CLIP	Yes	Yes	11,545,105	11,415,553	1,587,380	5,253,564	973,283	4,828,045	400,089	42.3%	400,089	3.5%		
-nitrogen	gPAR-CLIP	Yes	Yes	12,565,700	12,315,329	1,846,036	6,128,110	837,766	5,654,249	274,345	45.9%	274,345	2.2%		
-nitrogen	gPAR-CLIP	Yes	Yes	7,524,557	7,393,201	1,009,996	3,616,734	508,681	3,334,758	174,574	45.1%	174,574	2.4%		
WT	mRNA-seq	Yes	No	11,164,331	11,126,176	8,074,185	823,549	92,132	341,800	9,700	3.1%	9,700	0.1%		
WT	mRNA-seq	Yes	No	11,626,588	11,596,323	8,495,022	888,394	105,960	345,756	9,539	3.0%	9,539	0.1%		
WT	mRNA-seq	Yes	Yes	20,881,199	20,881,199	15,528,464	161,166	119,230	47,856	2,241	0.2%	2,241	0.0%		
WT	mRNA-seq	Yes	Yes	19,254,100	19,254,100	14,955,841	145,716	110,676	47,400	3,927	0.2%	3,927	0.0%		
WT	mRNA-seq	No	Yes	18,570,981	18,570,981	14,911,218	206,401	157,157	34,727	2,567	0.2%	2,567	0.0%		
WT	mRNA-seq	No	Yes	18,618,595	18,618,595	14,654,128	140,281	112,501	22,707	2,185	0.1%	2,185	0.0%		
-glucose	mRNA-seq	Yes	Yes	20,764,203	20,764,203	17,367,447	143,677	114,825	26,350	2,251	0.1%	2,251	0.0%		
-glucose	mRNA-seq	Yes	Yes	16,396,102	16,396,102	13,415,811	115,996	92,410	22,998	2,518	0.1%	2,518	0.0%		
-nitrogen	mRNA-seq	Yes	Yes	17,424,942	17,424,942	14,094,672	145,566	102,798	31,929	2,246	0.2%	2,246	0.0%		
-nitrogen	mRNA-seq	Yes	Yes	24,514,654	24,514,654	19,641,672	220,397	157,438	42,438	3,057	0.2%	3,057	0.0%		
WT (Puf3p IP)	PAR-CLIP	Yes	No	10,498,429	8,394,537	1,450,253	3,431,211	470,044	3,219,759	181,296	38.4%	181,296	2.2%		

Figure S3.1: Read counts and T-to-C conversion rates for all gPAR-CLIP, mRNA-seq, and PAR-CLIP libraries.

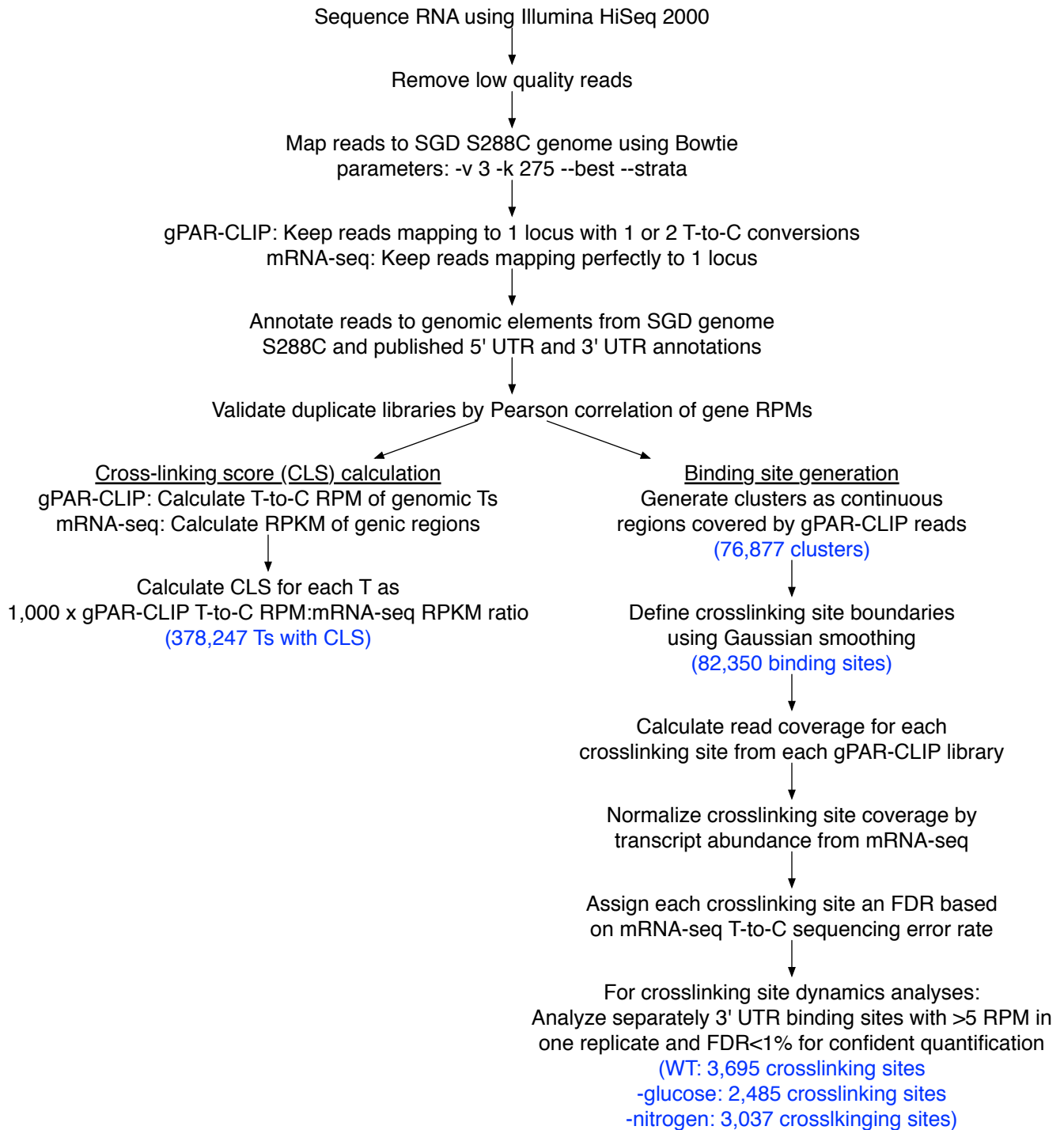


Figure S3.2: Processing steps used to generate crosslinking scores and crosslinking sites from gPAR-CLIP and mRNA-seq data.

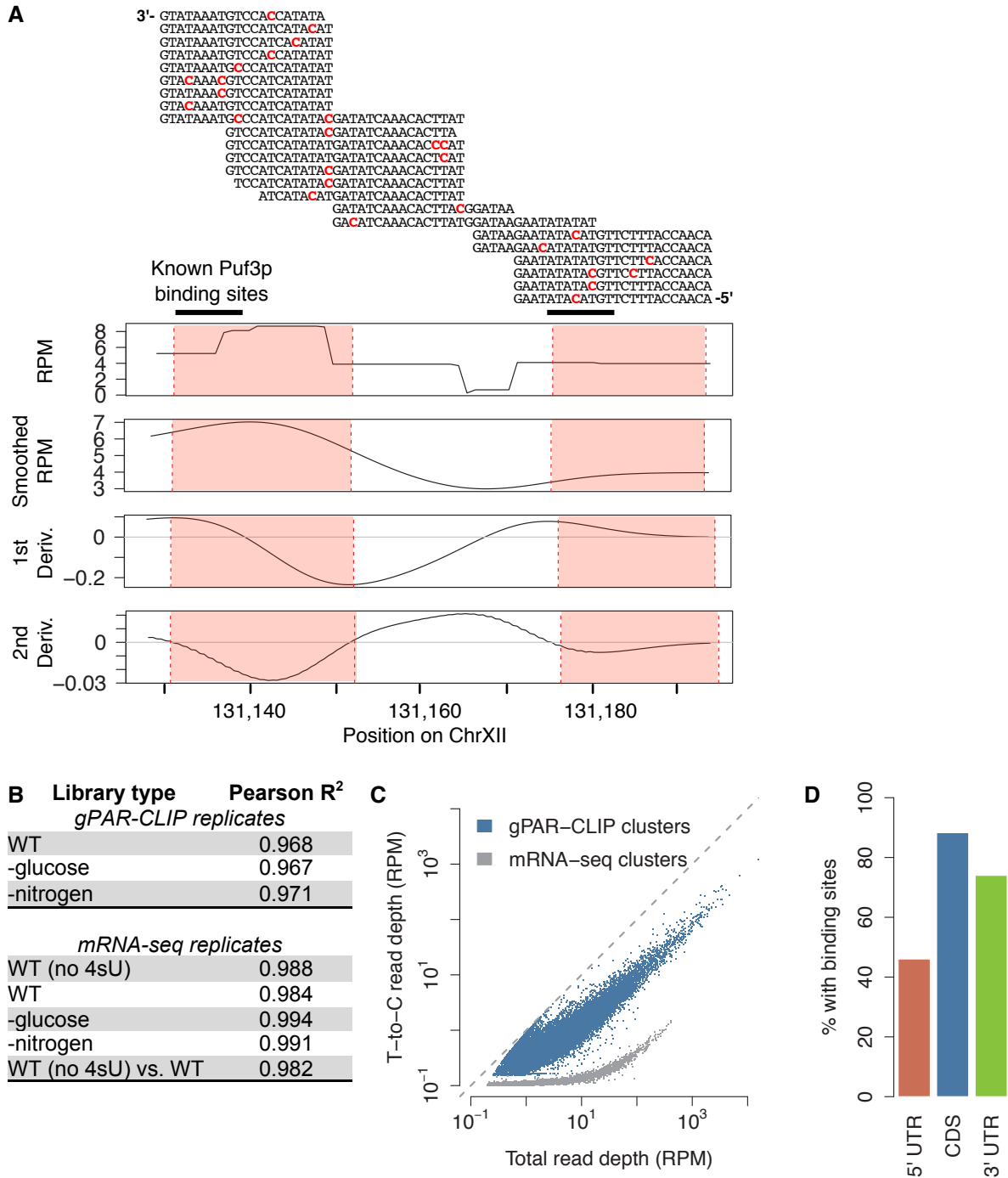


Figure S3.3: (A) Illustration of sequence block generation, Gaussian distribution fitting, and cluster segmentation to identify individual crosslinking sites. (B) Pearson correlation coefficients for all gPAR-CLIP and mRNA-seq replicate libraries based on gene RPM values. (C) Separation of T-to-C sequencing errors from crosslinking-induced mismatches. Plotted for each cluster is T-to-C RPM coverage versus total RPM coverage from gPAR-CLIP or mRNA-seq libraries. (D) Percent of annotated 5'UTR, CDS, and 3'UTR regions with at least 1 crosslinking site with >5 RPM.

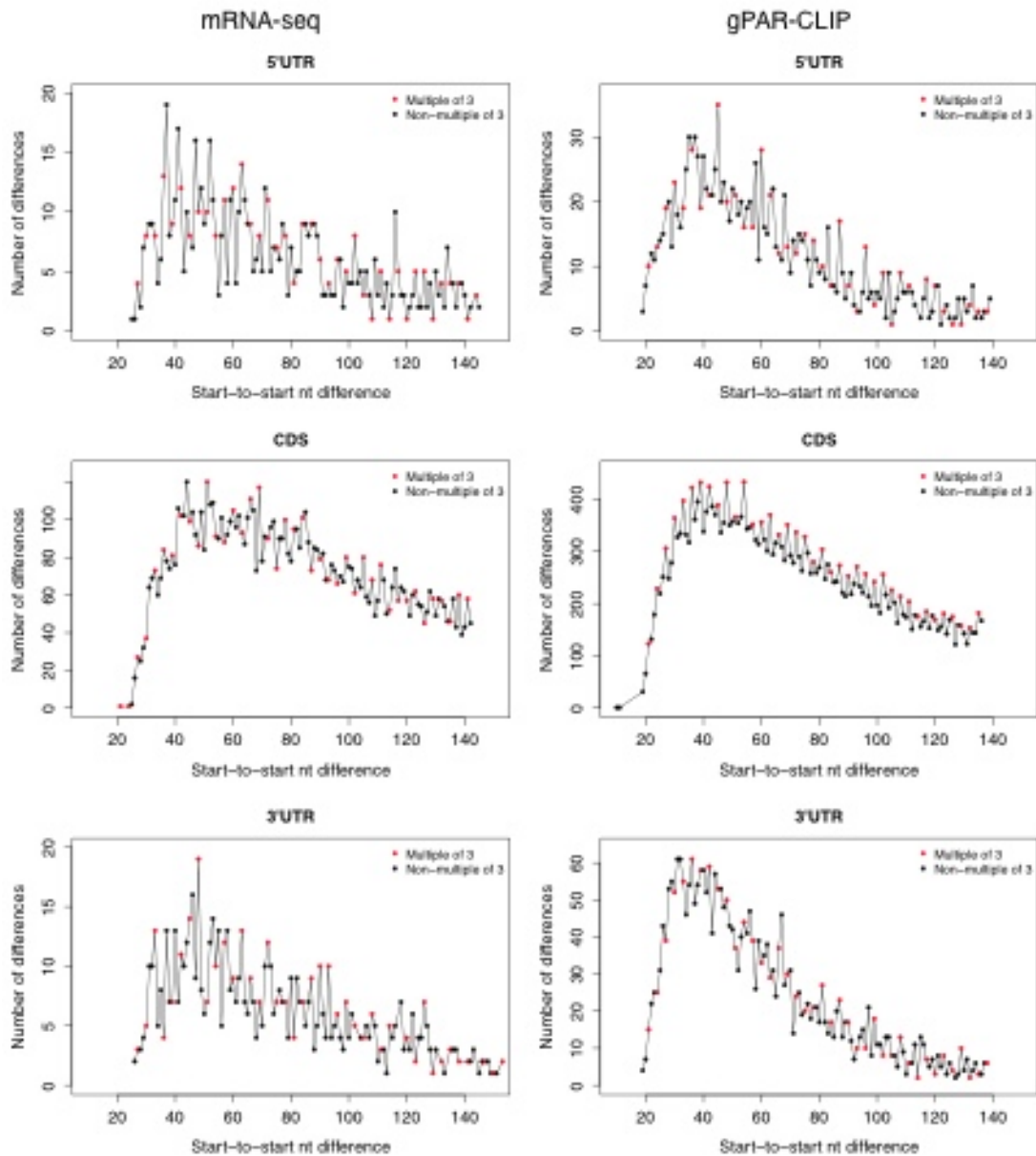


Figure S3.4: Distribution of start-to-start nucleotide distances between 5'UTR, CDS, and 3' UTR read clusters from gPAR-CLIP and mRNA-seq libraries. Only distances from gPAR-CLIP CDS read clusters were enriched for multiples of 3 (red dots).

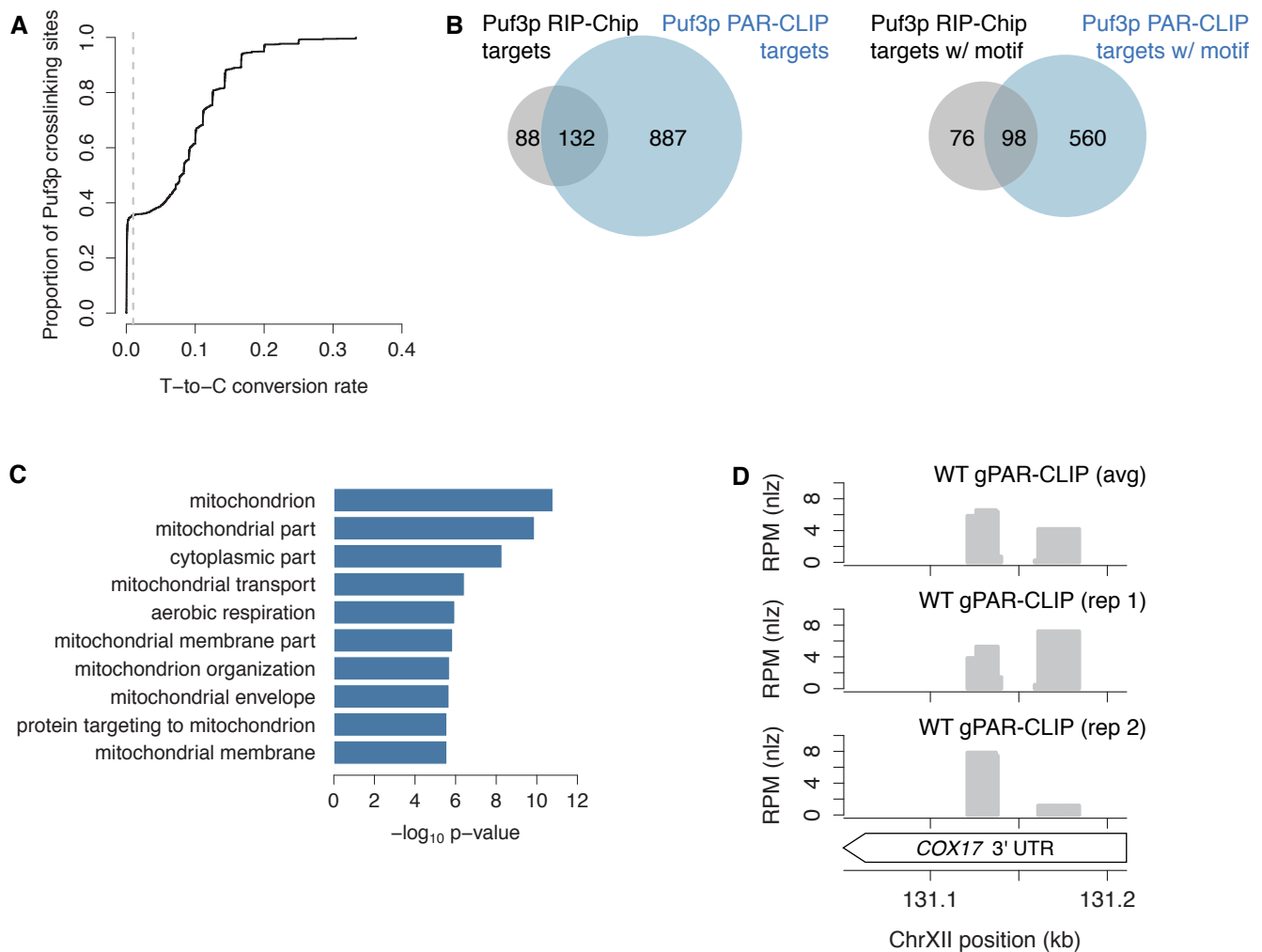


Figure S3.5: (A) Puf3p PAR-CLIP identified crosslinking sites in 147 (67%) of the 220 Puf3p target mRNA identified by RIP-Chip. 174 Puf3p RIP-Chip-identified target mRNAs contain the Puf3p recognition motif UGUAAAUA (Gerber et al., 2004; Hogan et al., 2008). Puf3p PAR-CLIP identified motif-containing crosslinking sites in 76 (44%) of these mRNAs and in 265 additional mRNAs, suggesting post-transcriptional regulation by Puf3p for these 265 novel targets. (B) GO enrichment analysis of 265 PAR-CLIP-identified, motif-containing Puf3p targets. Results are consistent with Puf3p's role in localization, deadenylation, and repression of mRNAs encoding proteins destined for the mitochondria (Olivas and Parker, 2000). (C) Individual replicate coverage of *COX17* 3'UTR in gPAR-CLIP with average coverage as shown in Figure 3.2B.

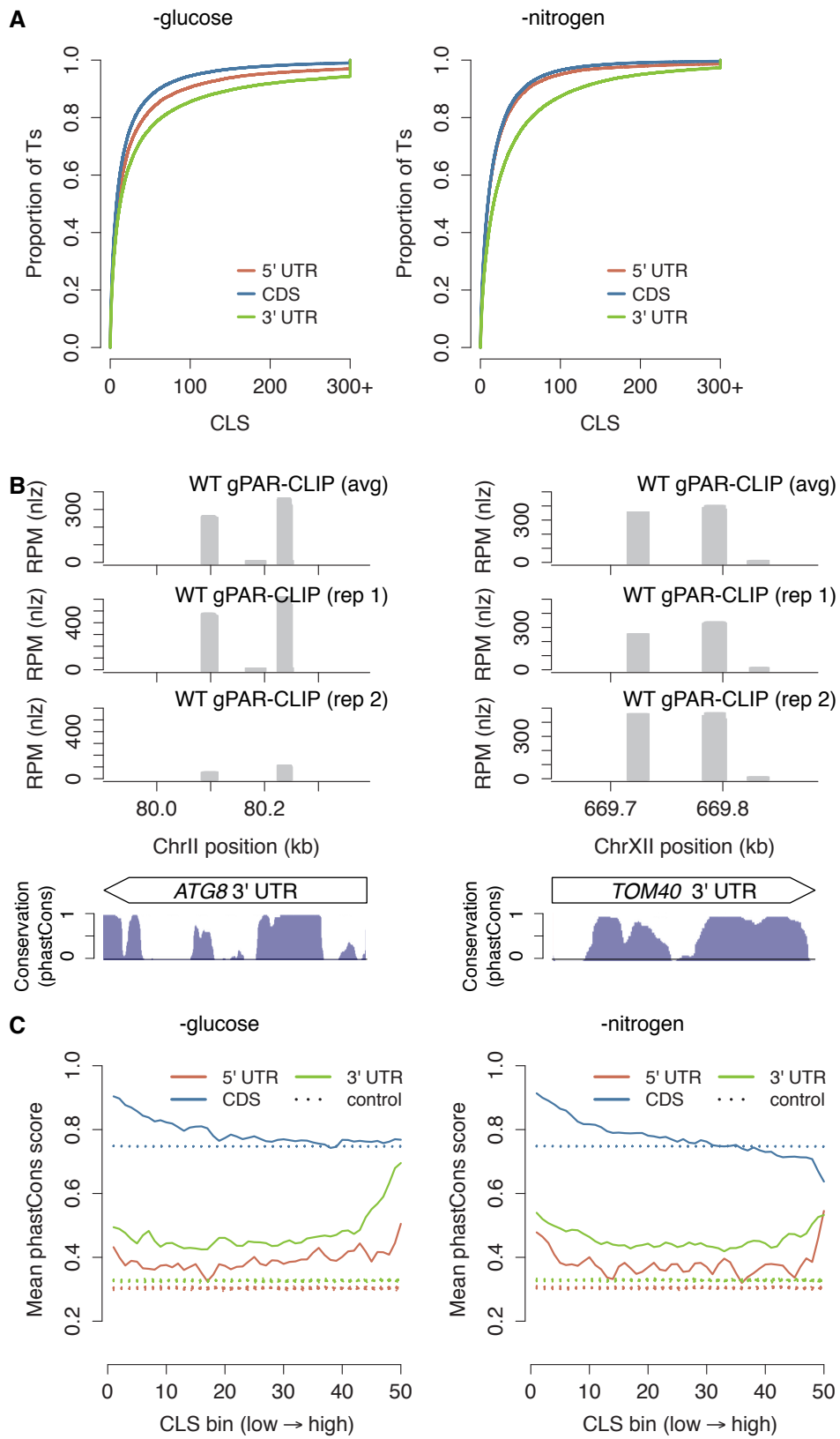


Figure S3.6: (A) Cumulative distribution of CLSs from 5'UTR, CDS, and 3'UTR regions. (B) Individual replicate coverage of *ATG8* and *TOM40* 3'UTRs in gPAR-CLIP with average coverage as shown in Figure 3.3D. (C) Mean phastCons scores for Ts ranked and binned by CLSs. Control lines represent mean phastCons scores of randomly ranked and binned Ts with no CLS, repeated 10 times.

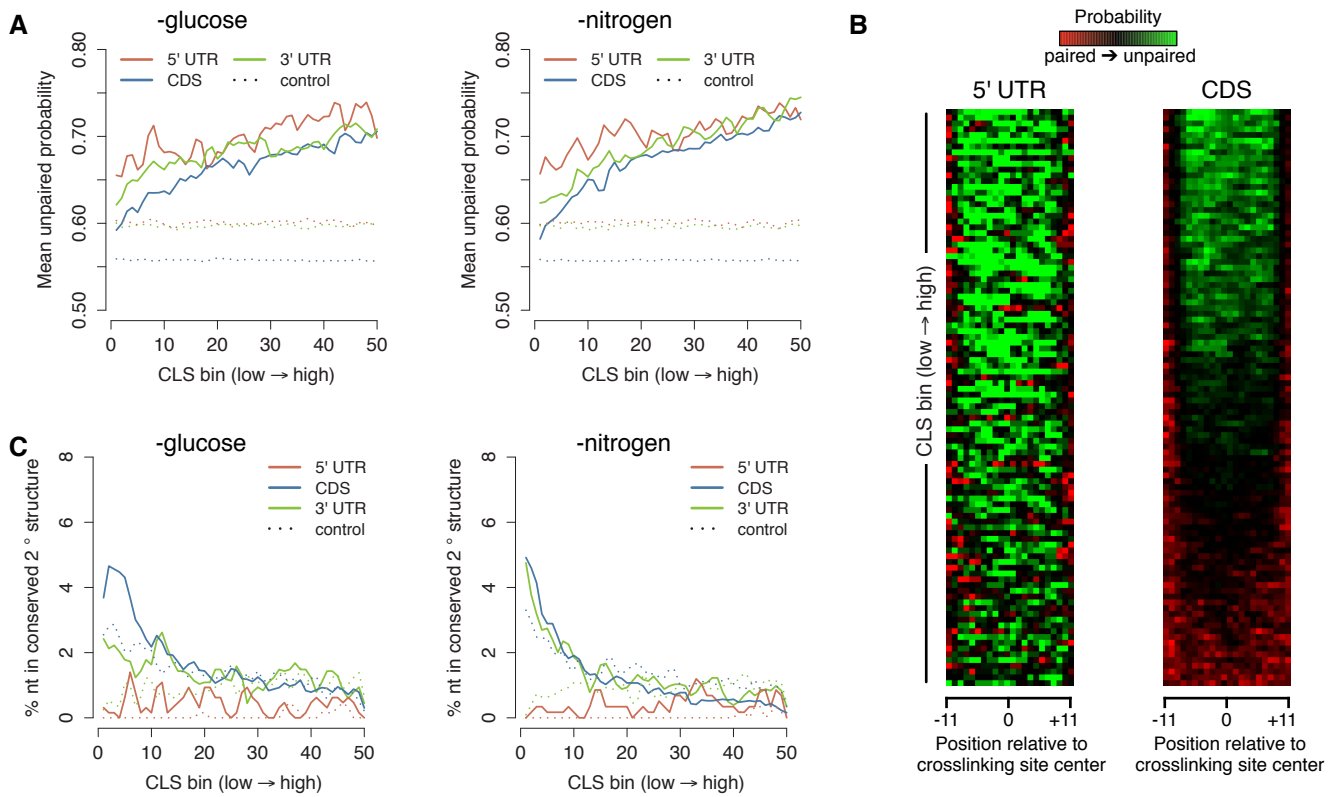


Figure S3.7: (A) Mean unpaired probability scores for Ts ranked and binned by CLSs. Control lines represent mean unpaired probability of randomly ranked and binned Ts with no CLS, repeated 10 times. (B) Heatmaps of pairedness of 5'UTR and CDS crosslinking sites ranked by average crosslinking site CLS. (C) Percentage of Ts ranked and binned by CLSs in conserved secondary structural elements as defined by RNAz. Control lines represent percentage of randomly ranked and binned Ts with no CLS in conserved secondary structural elements, repeated 10 times.

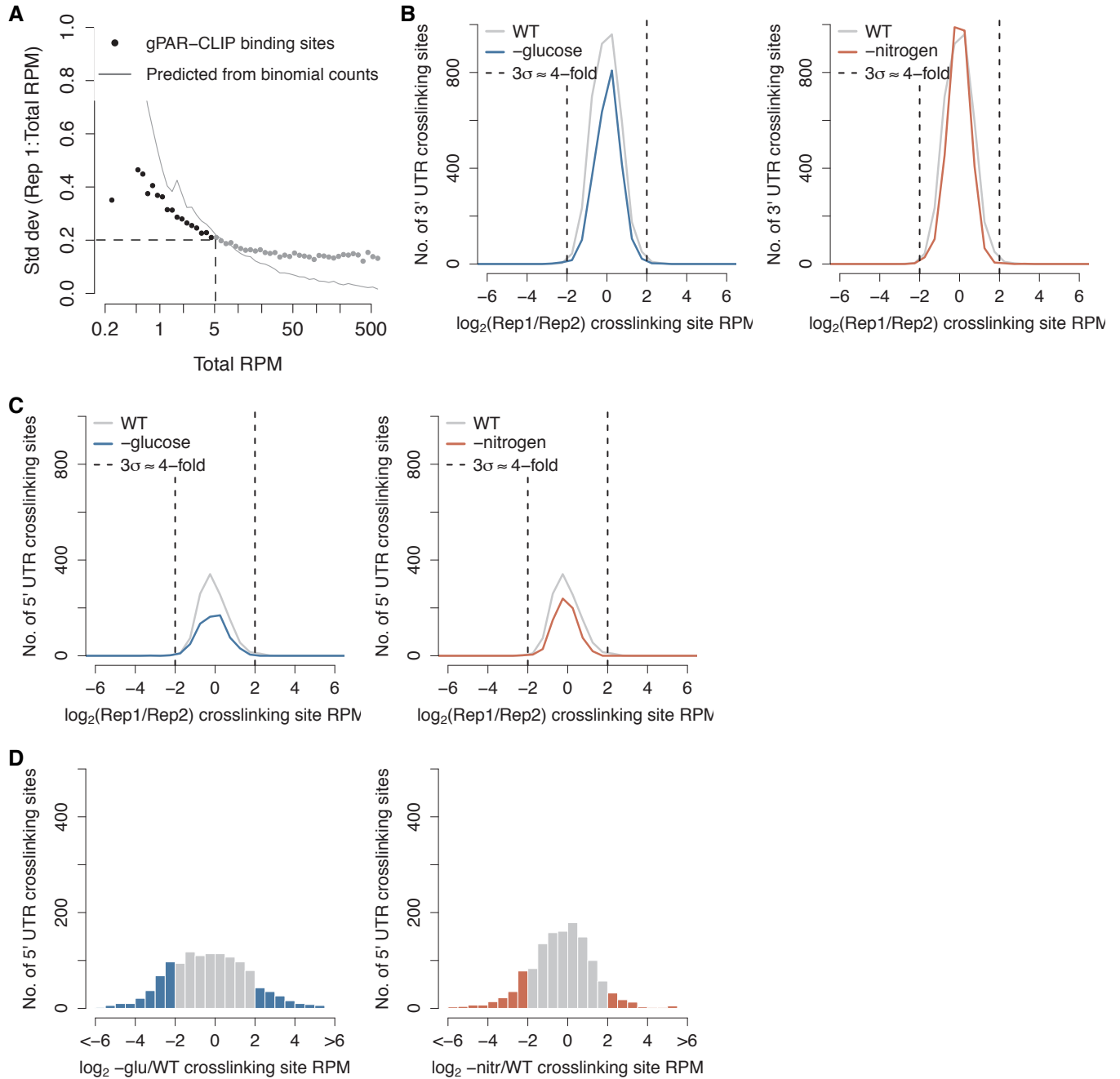


Figure S3.8: (A) Determination of minimum crosslinking site coverage required for comparison of sites across environmental conditions. 5 RPM was chosen as the minimum crosslinking site coverage needed for confident quantification since, at this coverage, the standard deviation of the fraction of crosslinking site reads coming from one replicate library stabilized at <0.2 . Shown is data from WT replicate libraries; similar results were obtained for all library types. (B) Intra-replicate variation of 3'UTR crosslinking sites in WT and glucose (left) or nitrogen (right) starvation conditions. Dotted lines represent 3 standard deviations from the mean and correspond to ~ 4 -fold change between WT replicates. (C) Same as (A) but for 5'UTR crosslinking sites. (D) Global changes in 5'UTR crosslinking site coverage upon glucose (left) or nitrogen (right) starvation.

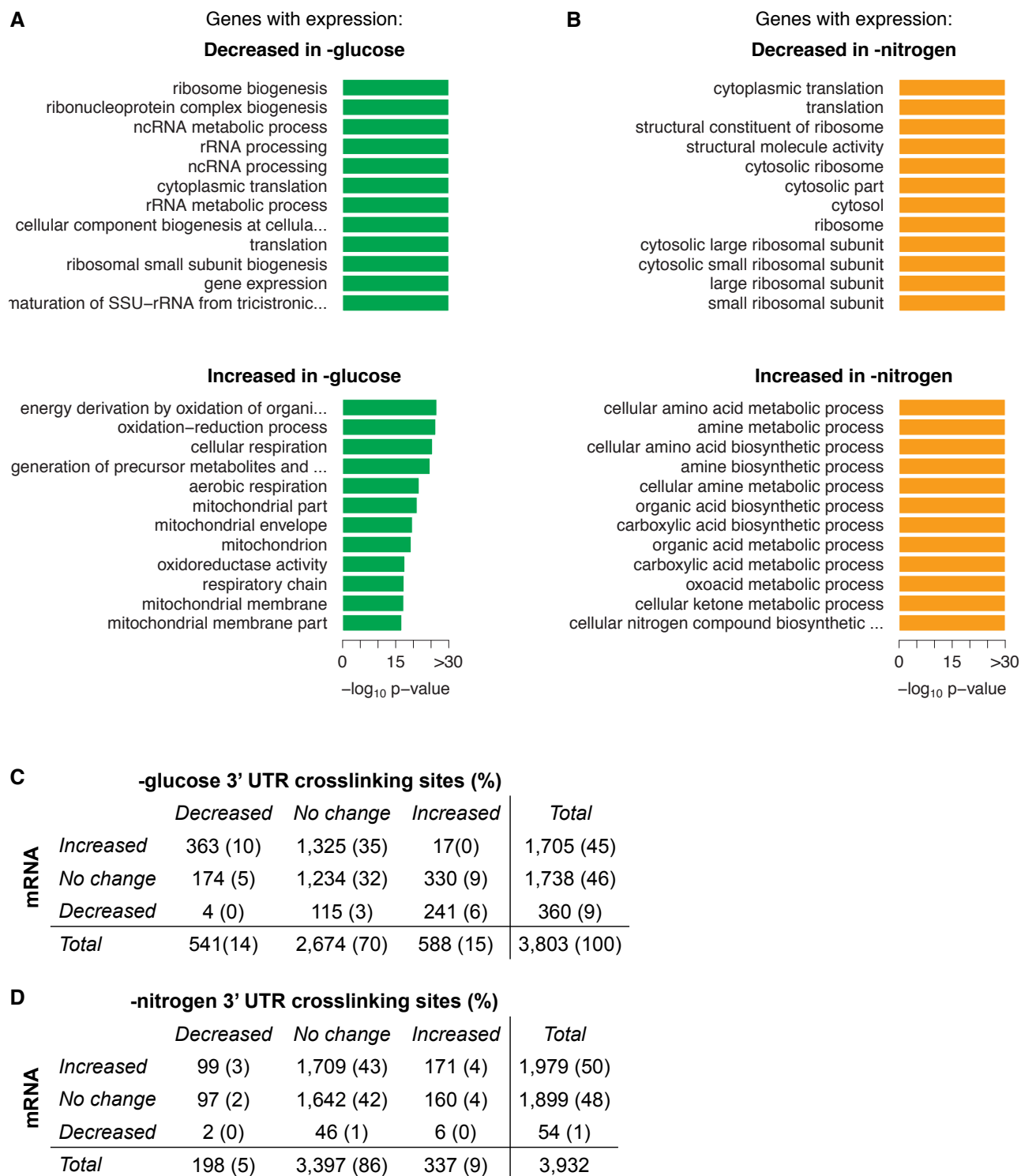


Figure S3.9: (A-B) Enriched GO terms of genes up- and down-regulated upon glucose (A) or nitrogen (B) starvation. (C) The number and percentage of 3'UTR crosslinking sites with indicated changes in crosslinking site coverage and corresponding mRNA expression upon glucose starvation. (D) The number and percentage of 3'UTR crosslinking sites with indicated changes in crosslinking site coverage and corresponding mRNA expression upon nitrogen starvation.

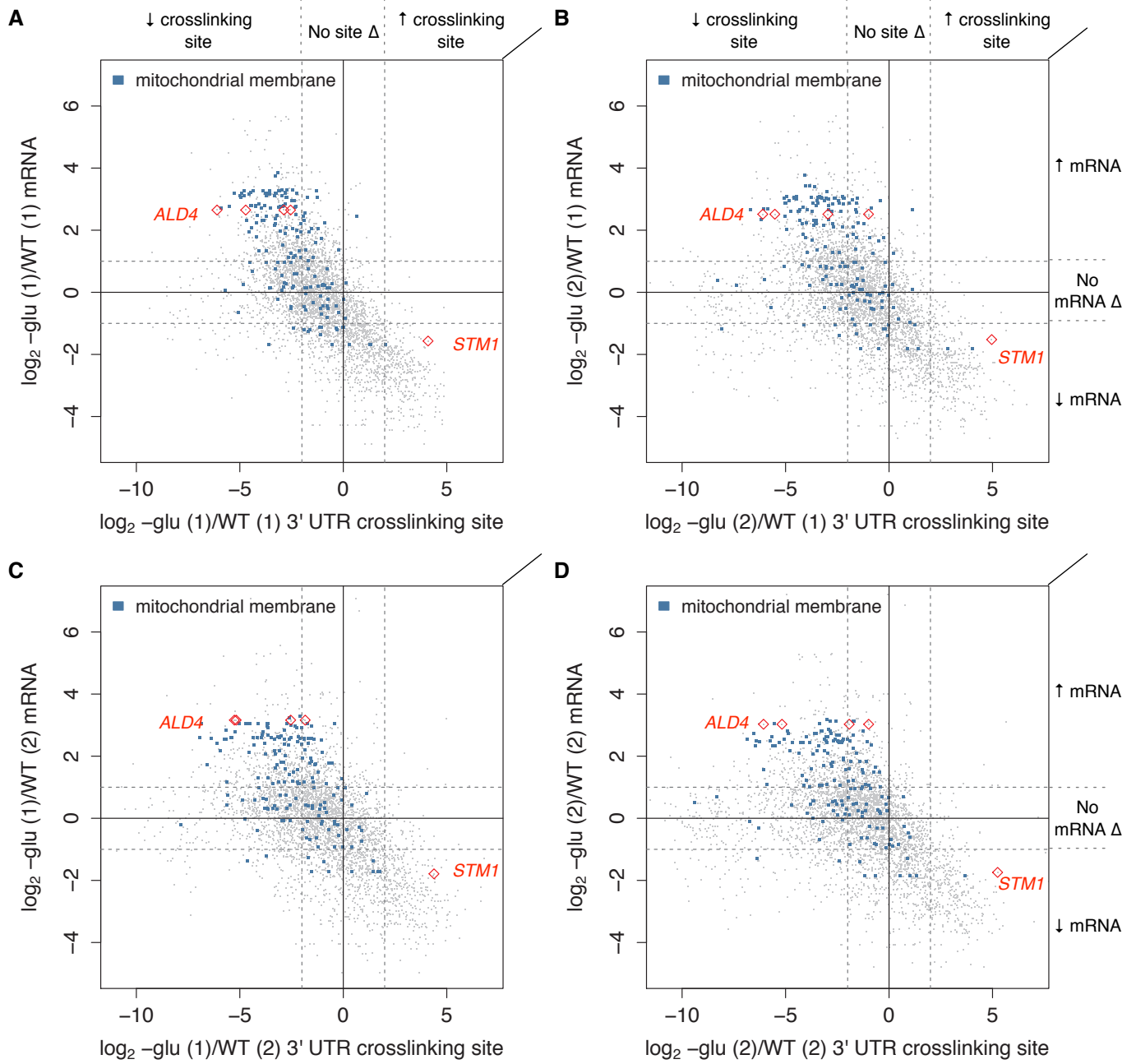


Figure S3.10: Changes in crosslinking site coverage from one replicate library each of WT and glucose starvation conditions are plotted versus changes in the corresponding mRNA from one replicate library each of WT and glucose starvation conditions. Dotted lines and colors are as in Figure 3.6A.

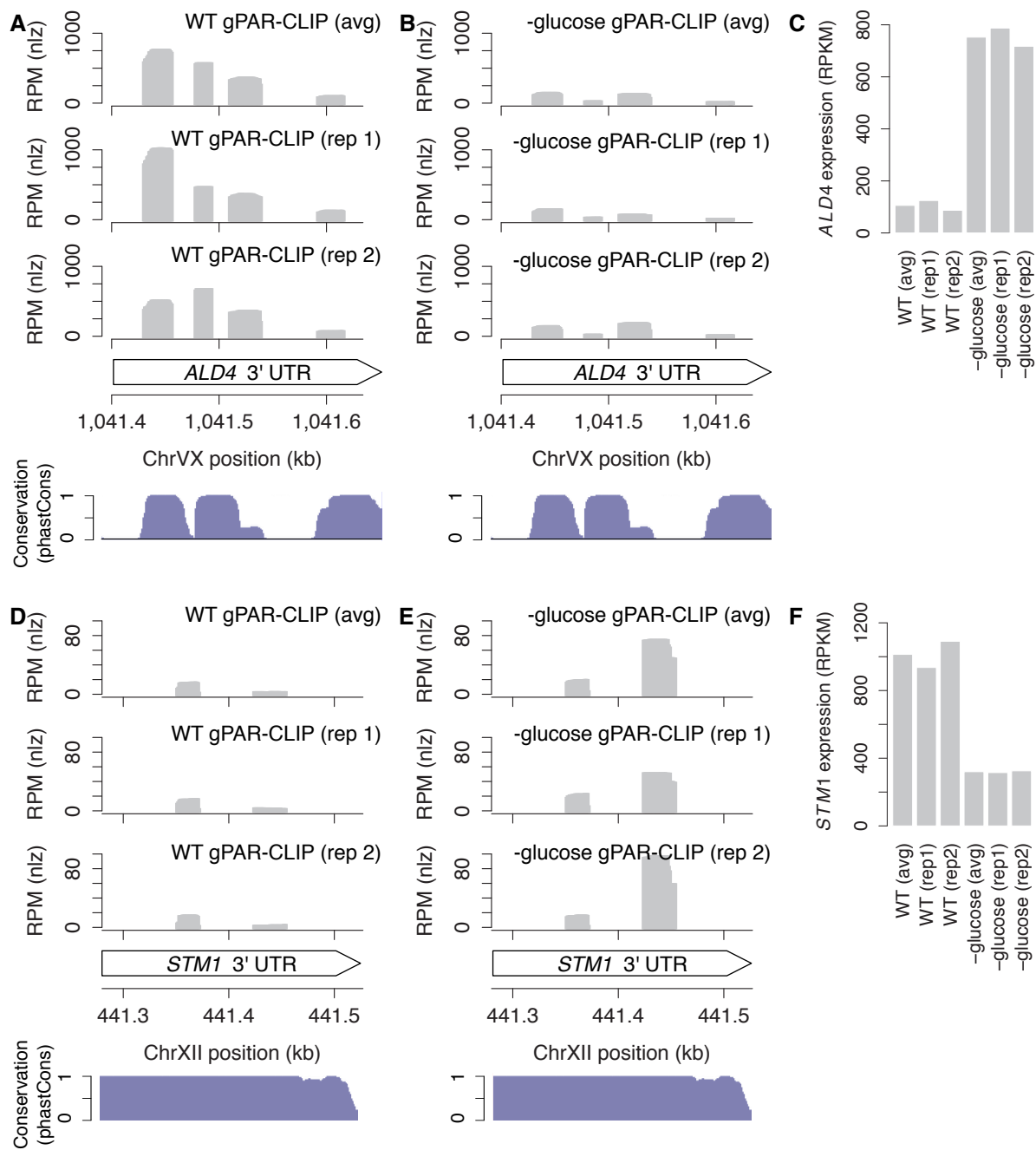


Figure S3.11: Same as Figure 3.6B-E but showing crosslinking site coverage and mRNA expression in individual replicate libraries.

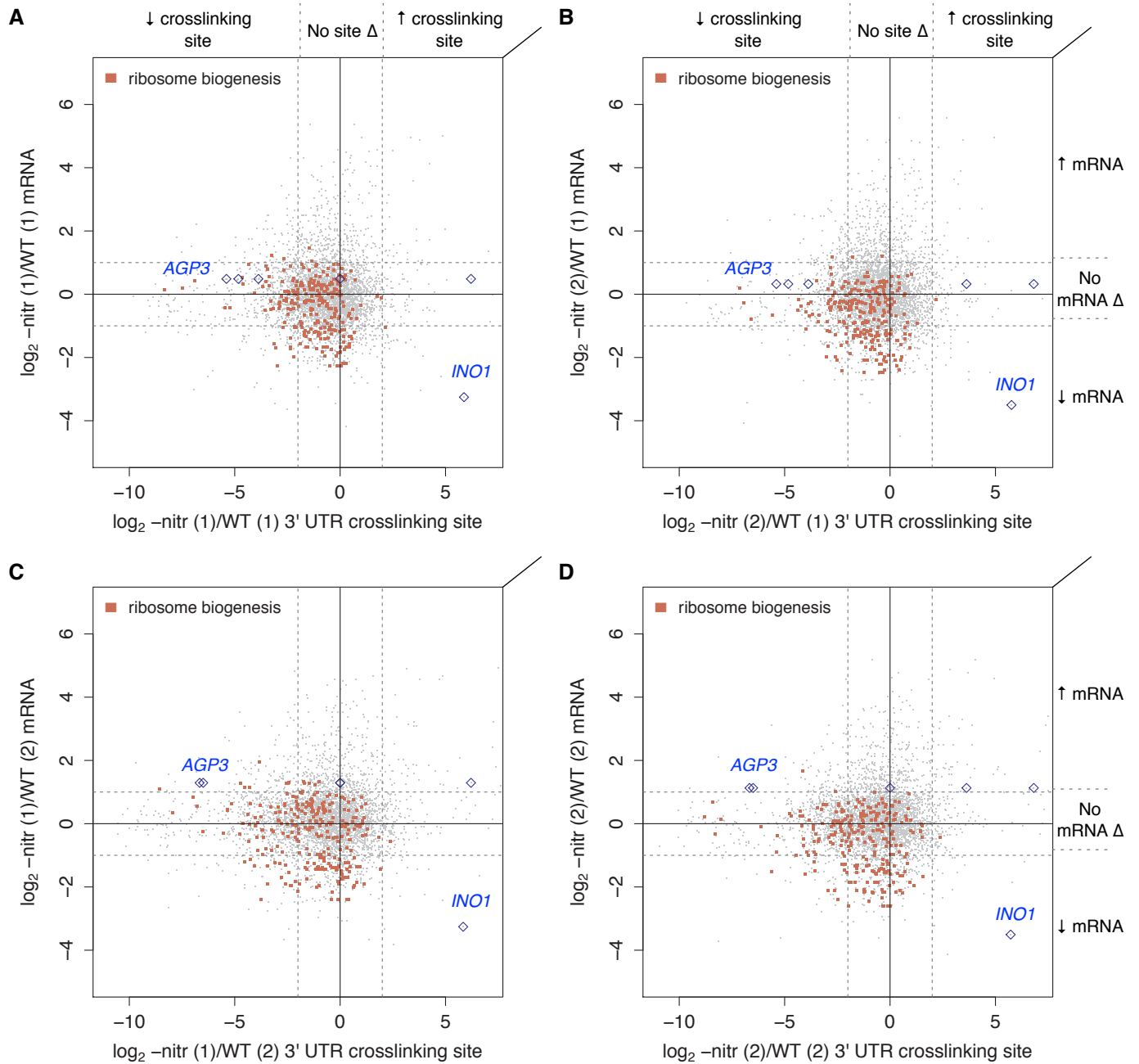


Figure S3.12: Changes in crosslinking site coverage from one replicate library each of WT and nitrogen starvation conditions are plotted versus changes in the corresponding mRNA from one replicate library each of WT and nitrogen starvation conditions. Dotted lines and colors are as in Figure 3.7A.

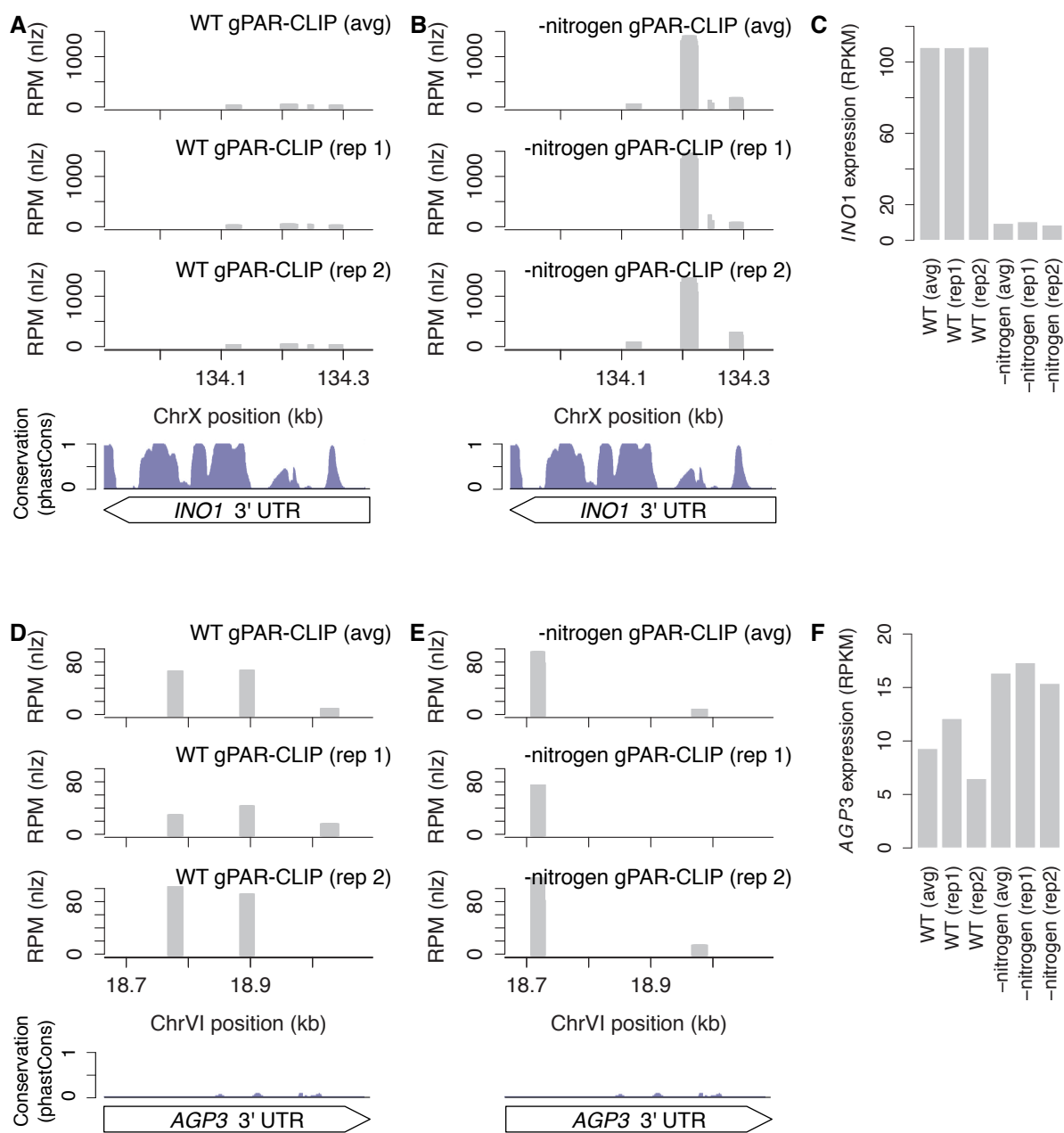


Figure S3.13: Same as Figure 3.7B-E but showing crosslinking site coverage and mRNA expression in individual replicate libraries.

CHAPTER IV

Mechanisms of PTGR by Pumilio family RBPs investigated by PAR-CLIP-seq

4.1 Introduction

RNA-binding proteins (RBPs) are crucial components of a diverse range of cellular processes. They play critical roles in post-transcriptional gene regulation (PTGR) by assembling RNAs and other protein co-factors into higher-order ribonucleoprotein (RNP) complexes to regulate RNA splicing, transport, storage, stability, and translation (Anderson and Kedersha, 2009; Glisovic et al., 2008; Shahbadian and Chartrand, 2012).

Puf (*Pumilio/FBF*) proteins are a family of highly conserved RBPs that serve a diverse set of roles including environmental stress response (Foat et al., 2005; Miller et al., 2014), germline stem cell maintenance (Ariz et al., 2009; Subramaniam and Seydoux, 2003; Suh et al., 2009), embryonic development (Datla et al., 2014; Kuo et al., 2009; Zhang et al., 1997), and control of neuronal plasticity, learning, and memory formation (Dubnau et al., 2003; Menon et al., 2004; Muraro et al., 2008). Mechanisms of Puf regulation typically include recognizing and binding short sequence elements at 3' untranslated regions (UTRs) to promote target deadenylation of poly(A) tails, degradation of the mRNAs, and translational

Official citations:

Freeberg MA, Han T, and Kim JK. Transcriptome-wide identification and target analysis of the Pumilio family of RNA binding proteins. (submitted)

repression (Goldstrohm et al., 2006, 2007; Gu et al., 2004; Hook et al., 2007; Lee et al., 2010; Miller and Olivas, 2011; Olivas and Parker, 2000; Quenault et al., 2011; Wickens et al., 2002). Puf proteins have also been shown to localize target mRNAs to specific subcellular regions, in particular to the tip of budding yeast cells (Gu et al., 2004; Saint-Georges et al., 2008; Vessey et al., 2010; Zipor et al., 2009). Interestingly, recent work has also identified an emerging role for Pufs in promoting target RNA expression (Kaye et al., 2009; Piqué et al., 2008; Suh et al., 2009), illustrating the diversity and importance of Puf-mediated PTGR.

The RNA-binding domain (RBD) of Puf proteins typically contains eight tandem repeats of ~ 36 amino acid residues that form three α -helices (Figure 4.1a). These repeats nestle together in a crescent shape with the inner surface forming the interface with RNA substrates (Edwards et al., 2001; Wang et al., 2002, 2009a). Each target RNA base is recognized through stacking and edge-on interactions with side chains of three specific amino acid residues (tripartite recognition motif, TRM) within each repeat with the most N-terminal TRM recognizing the most 3' nucleotide within the binding site (Campbell et al., 2014; Koh et al., 2011). Conservation of Puf repeats has led to an observed RNA recognition motif common to many Pufs starting with 5'-UGUA-3' followed by AU-rich sequences (Gerber et al., 2004, 2006; Goldstrohm et al., 2006, 2007; Gu et al., 2004; Hogan et al., 2008; Hook et al., 2007; Lee et al., 2010; Olivas and Parker, 2000; Zhang et al., 1997). Yeast Pufs also harbor multiple low-complexity (LC) domains annotated outside of their Puf RBDs, similar to LC domains observed globally for RBPs that promote aggregation of RNAs and RBPs into RNP granules (Castello et al., 2012; Chen and Moore, 2014; Han et al., 2012; Kato et al., 2012). Select Pufs from budding yeast and worm contain an additional pseudo-Puf repeat (R8'), located C-terminal to the last Puf repeat (R8), which forms a binding pocket for a cytosine 2nt upstream (+2) of the UGUA sequence (Qiu et al., 2012).

The *Saccharomyces cerevisiae* genome encodes six Puf proteins: Jsn1p (Puf1p), Puf2p,

Puf3p, Puf4p, Mpt5p (Puf5p), and Puf6p. Although all yeast Pufs contain tandem repeats of the Puf RBDs, each protein has its own set of unique targets that have been identified through either directed or global *in vitro* studies (Gerber et al., 2004, 2006; Hogan et al., 2008). Puf1p and Puf2p, which have an additional RNA-binding domain (RRM1) located N-terminal of the Puf repeats, are paralogs that arose from the whole yeast genome duplication event (Byrne and Wolfe, 2005; Kellis et al., 2004) and have been shown through a genome-wide *in vitro* affinity purification assay to bind mRNAs encoding membrane-associated proteins. Puf1p, in combination with other Pufs, mediates decay of two specific mRNAs, *HXK1* and *TIF1*, through binding 3'UTRs (Ulbricht and Olivas, 2008). Puf3p globally binds 220 mRNAs *in vitro* (Gerber et al., 2004). In a previous study, we identified 1,636 mRNAs by *in vivo* crosslinking, immunopurification, and deep-sequencing of bound RNAs (Freeberg et al., 2013). Puf3p directly binds target mRNAs encoding mitochondrial proteins, including *COX17* for which two well-characterized UGUA-containing Puf3p target sites have been validated (Houshmandi and Olivas, 2005; Jackson et al., 2004). Puf4p binds mRNAs encoding nuclear and nucleolar proteins in addition to ribosome biogenesis factors and ribosomal subunits (Foat et al., 2005; Gerber et al., 2004; Grigull et al., 2004). Many of these proteins are upregulated in *puf4* Δ mutants, suggesting that Puf4p functions as a repressor for these targets. Global *in vitro* analysis of Puf5p targets identify 224 mRNAs that encode proteins involved in nuclear functions, such as histone modification, chromatin remodeling, and transcriptional regulation, as well as components of the microtubule spindle (Gerber et al., 2004). Puf6p has one validated mRNA target, *ASH1*, which requires Puf6p binding for translational repression and localization to the yeast bud tip (Deng et al., 2008; Gu et al., 2004). Recently, a genome-wide ChIP-chip study identified 40 co-transcriptional targets of Puf6p, including several additional bud-localized mRNAs (Shahbadian et al., 2014), suggesting that Puf6p is first recruited to its targets during transcription before functioning in translational repres-

sion and mRNA localization. Despite these studies, the global set of Puf6p mRNA targets remains unknown.

In some cases, multiple Pufs bind cooperatively to regulate a single mRNA target. For example, regulation of *TIF1* decay depends on interactions at the 3'UTR with Puf1p and Puf5p, which can function individually but are both required for full regulation (Ulbricht and Olivas, 2008). Decay of *HXK1* mRNA depends on binding of Puf1p, Puf4p, and Puf5p (Ulbricht and Olivas, 2008). *HO* deadenylation is regulated by co-binding of Puf4p and Puf5p to adjacent sites on the 3'UTR and can both recruit the Ccr4p/Pop2p deadenylase complex (Hook et al., 2007). What remains to be discovered is the extent to which co-regulation by multiple Pufs occurs outside of these few isolated cases.

While informative, results from *in vitro* protein microarray studies have two major drawbacks for addressing the biological function of Puf proteins. First, the Puf:RNA interactions are identified in an artificial setting and thus might not reflect biologically relevant interactions or might reveal false positive interactions that do not occur *in vivo* due to distinct cellular localization of either the Puf protein or the RNA target. Second, nucleotide resolution of Puf:RNA binding site interactions is low - typically, entire transcripts are identified as Puf targets and presumed binding sites within a transcript are predicted *in silico* based on the existence of known Puf primary sequence motifs. For example, primary sequence motifs have been identified within target RNAs bound by Puf2-5p and used as a proxy for defining the precise sites of Puf:RNA interactions (Gerber et al., 2004; Hogan et al., 2008; Yosefzon et al., 2011). Differentiating between motif and non-motif sites and sites located within distinct genic regions of a transcript would aid in defining the precise, and perhaps multiple, roles of Puf proteins in regulating their target RNAs.

A recently developed method, termed photoactivatable ribonucleoside-enhanced crosslinking and immunopurification of an RBP of interest (PAR-CLIP), allows for the isolation and

identification by next-generation sequencing of precise RNA binding sites of RBPs *in vivo* (Hafner et al., 2010b). An advantage of using PAR-CLIP over other methodologies is the presence of T-to-C conversion events in the resulting deep sequencing reads that mark sites of covalent bond formation between metabolically incorporated photoactivatable ribonucleoside analog 4-thiouracil (4sU) and amino acid side chains of an RBP. PAR-CLIP followed by deep sequencing (PAR-CLIP-seq) has been used to identify comprehensively the *in vivo* RNA targets of RBPs in yeast (Baejen et al., 2014; Creamer et al., 2011; Jamonnak et al., 2011; Webb et al., 2014) and mammalian cells (Farazi et al., 2014; Gottwein et al., 2011; Hafner et al., 2010b; Wen et al., 2011). We recently implemented PAR-CLIP-seq to identify binding sites of Puf3p on its mRNA targets (Freeberg et al., 2013). In agreement with published observations, we identified the 5'-UGUAWAUA-3' conserved motif present in our high-confidence Puf3p sites, confirming the utility of this method to identify true targets of a Puf protein.

Despite the importance of Puf-mediated PTGR and the genome-wide identification of hundreds of mRNA targets *in vitro*, a comprehensive set of Puf target mRNAs in an *in vivo* cellular context remains to be determined. In this study, we present for the first time *in vivo* identification and characterization of the global set of direct target sites for a complete family of RBPs in an organism. We identify recognition motifs within direct binding sites of Puf1p and Puf2p that align with predicted tertiary structures of their Puf repeats. We also differentiate between multiple RNA recognition sequences for Puf3-5p that suggest distinct structural conformations of the RBP:RNA complexes. In addition, we show that Puf6p, which is structurally distinct from other canonical Pufs, preferentially binds rRNAs and snoRNAs at sites lacking a conserved primary motif. Combined with Puf6p interaction data, our observations indicate a role for Puf6p in pre-rRNA processing, perhaps through recognition of conserved secondary structures. Finally, we correlate features of Puf bind-

ing with various functional outcomes. We demonstrate that the majority of individual Puf binding sites are represented in a global annotation of RBP binding and show that subsets of direct Puf binding sites are gained or lost in a stress-specific manner. We also observe different functional classes of mRNA targets and effects on mRNA stability correlated with identified Puf recognition motifs. By mining PAR-CLIP-seq-derived sites directly bound by each Puf, we gain insight into how features of Puf:RNA binding affect PTGR.

4.2 Transcriptome-wide identification and target analysis of the Pumilio family of RNA binding proteins

Pufs bind over half of the yeast protein-coding transcriptome. To define the precise binding sites and interrogate RNA binding preferences for the Pumilio family of RNA binding proteins in yeast, we performed PAR-CLIP on strains expressing TAP-tagged versions of Puf1p-Puf6p. After RNase digestion of the bound RNAs to fragments protected by Puf binding, cDNA libraries were generated and subjected to Illumina deep sequencing. From mapped reads normalized to total library size, Puf binding sites were defined by overlapping reads with high-confidence sites containing at least one read with a T-to-C conversion event and passing an empirically determined minimum read count threshold (details in Methods, [Figure S4.1a-d](#)).

Using PAR-CLIP-seq, we identified at least one Puf binding site on over half of the protein-coding genes for yeast (3,006 of 5,911 annotated ORFs) including 32% of annotated 3'UTRs (1,793 of 5,684) where a majority of all RBP binding occurs (Freeberg et al., 2013; Matoulkova et al., 2012) ([Figure 4.1b](#)). Further, 47% of Puf mRNA targets (23.5% of all protein-coding genes) are bound by more than one Puf, suggesting that combinatorial control of PTGR by Puf proteins is more extensive than previously thought and highlighting the

pervasiveness of regulation by a single family of RBPs in an organism. We identified 331 (Puf1p), 602 (Puf2p), 1,441 (Puf3p), 1,005 (Puf4p), and 1,263 (Puf5p) novel Puf targets that contain Puf binding sites with equivalent RPM coverage to validated Puf binding sites (Figure S4.1c). Thus, PAR-CLIP-seq is more sensitive than *in vitro* methods for identifying Puf targets and has the additional advantage of pinpointing direct Puf:RNA interaction sites with nucleotide resolution.

Many examples of Puf-mediated PTGR involve direct interaction of Pufs with substrate 3'UTRs. To determine globally if Pufs interact predominantly with 3'UTRs, we calculated the proportion of normalized reads for each Puf mapping to 5' and 3' UTRs, CDS, and non-coding RNAs (ncRNA) (Figure 4.1c). The majority of Puf3/4/5p binding occurs at 3'UTRs (65-72%). Interestingly, while Puf1p and Puf2p RNA binding mostly occurs at 3'UTRs (58 and 49%, respectively), a large proportion (15 and 16%, respectively) of binding also occurs at 5'UTRs, suggesting that Puf1/2p might have distinct regulatory roles depending on where they bind their targets. Surprisingly, Puf6p binding is almost entirely absent from either 5' or 3' UTRs but enriched at snoRNAs and coding regions.

Binding of RBPs to different regions of target transcripts reflects different mechanisms of PTGR. For example, *C. elegans* GLD-1 binding at target CDSs promotes translational repression (Brümmer et al., 2013) while binding at target 3'UTRs is required for mRNA stabilization (Scheckel et al., 2012). We sought to distinguish between potential PTGR roles for each Puf depending on the set of targets bound at different genic regions. Puf targets with binding sites in their 5'UTRs, CDSs, or 3'UTRs were searched for enriched gene ontology (GO) terms (weighted Fisher's exact test $p < 1e-4$; details in Methods). Genes with 3'UTR Puf3p sites are enriched for many GO terms relating to the mitochondria, confirming previous observations based on whole-target analysis (Gerber et al., 2004; Hogan et al., 2008) (Figure 4.1d). Targets with 5'UTR or CDS Puf3p sites are enriched for different

GO terms, highlighting the subtle differences in the types of mRNAs targeted by Puf3p at different genic regions. mRNAs targeted by Puf4p to the 3'UTRs or CDS are enriched for GO terms relating to translation and ribosomes (Gerber et al., 2004), while Puf5p targets are enriched for cell wall-related GO terms agnostic of the genic region bound. Cell wall-related GO terms are also enriched among Puf2p, Puf4p, and Puf6p targets, regardless of genic region targeted. This could suggest that either Pufs, in general, might be important for regulating genes encoding cell wall components or that cell wall-related genes are bound non-specifically by Pufs.

We identify hundreds of novel *in vivo* mRNA targets for each Puf, including for Puf6p for which only one direct mRNA target was known. Our data indicate that features of Puf target binding should be investigated separately based on the genic regions with which each Puf interacts. By deriving binding sites *in vivo* by PAR-CLIP-seq, we can interrogate a much smaller subset of directly bound RNA regions for each Puf to understand how these proteins function in PTGR. In the following sections, we focus on and explore binding features of each Puf.

Puf1/2p binding motifs correlate with Puf protein domain structure. *In vitro* protein microarray studies revealed that Puf1p and Puf2p bind mRNAs encoding membrane-associated proteins (Gerber et al., 2004; Hogan et al., 2008). Within the 3'UTRs of these Puf2p targets, a motif was found consisting of two UAAU sequences separated by a 0-6 nt spacer (Yosefzon et al., 2011). Binding of Puf2p to this motif was validated *in vivo* for the Puf2p target 3'UTR of PMP2; however, the specificity of Puf2p for this motif in general has not been explored. Through *in vitro* yeast three-hybrid assays, Puf2p repeats R1/2/3 were shown to be important for interaction with the UAAU motif, and results for Puf1p suggest it has a similar binding motif. Using *in vivo* transcriptional shutoff assays to measure decay

rates, Ulbricht and Olivas (2008) show that Puf1p- and Puf5p-mediated regulation of *TIF1* decay can be disrupted by mutating a UGUA sequence in site #1 identified in the 3'UTR; however, whether Puf1p binds directly to this motif on *TIF1* and, if so, whether this motif is specific to *TIF1* or generalizable to other Puf1p targets remains unknown.

To determine the direct *in vivo* binding motifs for Puf1p and Puf2p, we hierarchically searched our PAR-CLIP-seq-derived binding sites for each Puf ranked by RPM coverage (details in Methods). Among Puf1p binding sites we identified a short motif consisting of a strong UAAU motif surrounded by AU-richness (Motif 1, present in 230 (26%) sites) and a UAA-repeat motif (Motif 2, present in 29 (3% sites) (Figure 4.2a). We did not find evidence of direct binding of Puf1p to site #1 in the *TIF1* 3'UTR identified by Ulbricht and Olivas (2008); however, we did find direct evidence of Puf5p binding to site #1, suggesting that Puf1p acts indirectly and Puf5p acts directly at this site to regulate degradation of *TIF1*. In our global Puf1p PAR-CLIP-seq dataset, sites containing Motif 1 are enriched at 3'UTRs and are significantly more efficiently bound than sites with no motif (Figure 4.2b,c). Further, we scanned all Puf1p binding sites and found 297 with at least one occurrence of the core UAAU sequence. On average, these sites had 2 copies of UAAUs most often overlapping to form a UAAUAAU sequence (Figure S4.2a,b). As shown for distinct genic regions, Puf1p target mRNAs are enriched for different sets of GO terms depending on which motif is present (Figure 4.2d). For example, only Puf1p targets with Motif 1-containing sites are enriched for the GO term “cellular response to starvation” suggesting a Motif 1-dependent link between Puf1p regulation and a starvation response.

We recovered the two known Puf2p binding sites in the *PMP2* 3'UTR that contain the sequences required for efficient binding by Puf2p (Yosefzon et al., 2011). Among all Puf2p binding sites, two significant motifs were identified through hierarchical motif searching: Motif 1 is UA-rich (present in 257 (13%) sites) and Motif 2 is UC-rich (present in 73 (4%) sites,

Figure 4.2e). Sites containing either motif are enriched at 3'UTRs and are more efficiently bound than sites with no motif (Figure 4.2f,g). While hierarchical searching identified no other significant motifs, 62 additional Puf2p sites contain short UA-repeats (data not shown). Like Puf1p, Puf2p binds different classes of mRNAs based on presence of a motif (Figure 4.2h).

We next investigated whether the motifs we identified within Puf1p and Puf2p binding sites could be explained by the arrangement of Puf repeats within their Puf domains. Puf1/2p each have eight Puf repeats: six are annotated in domain databases SMART and Pfam, but two more can be identified by protein sequence alignment of all yeast Puf proteins (Figure 4.1a). Puf repeats are characterized by having a small, nonpolar amino acid (most often glycine) immediately N-terminal of the TRM and a basic amino acid (most often lysine) immediately C-terminal of the TRM (Figure S4.2c). To better understand how Puf1p and Puf2p Puf repeats are arranged in space, we modeled their tertiary structures using the SWISS-MODEL protein structure homology-modeling server (Biasini et al., 2014) with the reported crystal structures of Puf3p (Zhu et al., 2009) and Puf4p (Miller et al., 2008) as templates (Figure 4.2i). Puf1/2p repeats R2-4 align with Puf3/4/5p repeats R2-4 and contain the canonical TRMs NQ-T (R2), TQ-W (R3), and NQ-Y (R4) that are predicted to recognize nucleotides U, A, and U, respectively (Figure S4.2d). Annotated Puf1/2p repeats R5-7 align to Puf3/4/5p repeats R5-7, they but do not contain canonical TRMs. The R5 TRMs are AR-R (Puf1p) and SR-R (Puf2p). AR-R is a rare TRM that might recognize A (Hall, 2014), although this has not been tested systematically, while the precise nucleotide bias of SR-R is unknown. Puf1/2p repeats R6 and R7 contain the non-canonical TRMs AT-L (Puf1p R6), TT-L (Puf2p R6), and SL-L (Puf1/2p R7). Leucine is present at the stacking residue positions in these repeats, the side chains of which are not electrostatic and do not protrude into the Puf protein cleft where the RNA is located. Thus, Puf1/2p repeats R6/7 most

likely do not form hydrogen bonds with interacting RNA and do not restrict the nucleotides located near this region of the protein. The two remaining Puf repeats, R1 and R8, are not annotated in protein domain databases but align to R1 and R8 of Puf3/4/5p. These repeats contain unrecognizable TRMs (LR-R/PH-T for Puf1p; LR-R/PY-Y for Puf2p), which might explain their absence in the SMART and Pfam databases. Based on their TRMs, Puf1p and Puf2p should recognize UA-rich motifs instead of the canonical UGUA motif observed for other Pufs, which is what we observe among our direct binding sites.

In a previous study, we identified RNA binding sites for all RBPs by global PAR-CLIP-seq (gPAR-CLIP-seq) in growing yeast and nutrient-starved yeast (Freeberg et al., 2013). Characterization of the global set of RBP-bound sites on the transcriptome revealed that 38% of sites located on 3'UTRs were >4-fold changed (increased or decreased binding by an RBP) upon starving yeast of either glucose or nitrogen. To determine the extent to which Puf1/2p target binding is dynamic, we interrogated how Puf binding is altered in response to nutrient starvation. We recovered 73% (142 of 194) of 3'UTR Puf1p binding sites and 76% (254 of 333) of 3'UTR Puf2p sites in the gPAR-CLIP-seq dataset (Figure S4.2e). Applying quantile normalization to binding site RPMs gave good correlation between datasets (Spearman's $\rho=0.49$ (Puf1p) and 0.60 (Puf2p); Figure S4.3a), indicating that gPAR-CLIP-seq data faithfully represents Puf1p and Puf2p binding sites and can be used as a proxy for assessing Puf binding sites under changing environmental conditions. Unexpectedly, while 3'UTR binding sites in the entire gPAR-CLIP dataset show 1.6-fold reduced overall RBP binding under glucose starvation conditions, Puf1p sites are bound 1.4-fold more and Puf2p sites are bound 1.5-fold more (t -test p -values= $3.85e-10$ and $1.81e-8$, respectively; Figure 4.2j). We did not observe differences in Puf1/2p binding site coverage in nitrogen starvation conditions (t -test $p>0.001$, Figure S4.3b) indicating that changes in Puf1/2p binding is glucose-dependent and not necessarily generalizable to other nutrient stress, such as loss of

nitrogen. Further, increased binding to the 3'UTRs of target transcripts in glucose starvation conditions depended on the presence of Motif 1 for both Puf1p and Puf2p, suggesting that these motifs function in a glucose-deprivation-dependent manner.

Taken together, the predicted Puf domain structures of Puf1p and Puf2p indicate the following RNA recognition motifs: 5'-?N_NUAU?-3' for Puf1p and 5'-?NN?UAU?-3' for Puf2p, where “N” is no constraint, “a” is low evidence for recognizing A, and “?” is no information regarding TRM specificity (Figure S4.2d). These predicted recognition elements are similar to the UAAU (Puf1p) and UA-rich (Puf2p) motifs we identified by PAR-CLIP-seq. The Puf1p UAAU motif contains an additional A between the Us not predicted by the Puf1p TRMs. To account for this extra base, Puf1p targets might exhibit “base flipping.” This phenomenon is observed among RNA substrates of Puf4p and *C. elegans* Pumilio protein FBF-2 whereby an extra RNA base is flipped away from and is unconstrained by the Puf repeat TRMs (Valley et al., 2012).

Puf3p target sites contain features of two distinct modes of binding. Puf3p is the yeast homolog of one the founding members of the Puf family of proteins, fly Pumilio, and human PUM1 and PUM2. The best-characterized role for Puf3p is promoting target mRNA degradation through direct interaction with the Ccr4p/Pop2p deadenylase complex (Lee et al., 2010; Tucker et al., 2002) or by affecting conformation of the Pab1p-mRNP structure and exposing the poly(A) tail to deadenylation by the Pan2p deadenylase (Lee et al., 2010). A conserved 5'-UGUAWAUA-3' RNA recognition motif was found enriched among direct Puf3p binding sites (Fazlollahi et al., 2014; Freeberg et al., 2013; McHugh et al., 2014) and has been shown to be required for Puf3p interaction with a limited set of known targets (Jackson et al., 2004; Olivas and Parker, 2000; Saint-Georges et al., 2008).

In our PAR-CLIP-seq dataset for Puf3p, we identify both Puf3p binding sites in the

COX17 3'UTR and one site in the *COX23* 3'UTR that are known to be required for deadenylation and degradation of these substrates (Jackson et al., 2004; Olivas and Parker, 2000; Saint-Georges et al., 2008). To determine globally the direct *in vivo* binding motifs for Puf3p, we hierarchically searched for enriched primary sequence motifs in the Puf3p binding sites. The top Puf3p motif (Motif 1) was found in 718 (18%) high-confidence binding sites including 471 (41%) 3'UTR sites (Figure 4.3a). Noting similar preference for either G or A at position 2 of this motif, we segregated sites containing Motif 1 into those with or without UGUA and observed two distinct sub-motifs: UGUAWAUA (Motif 1-UGUA) identified in 329 (8%) sites and UAUAWAUA (Motif 1-UAUA) identified in 389 (10%) sites (Figure 4.3a). Motif 1-UGUA highly resembles sequences identified as a consensus motif among mRNA targets of Puf3p by *in silico* searching of entire target 3'UTRs (Foat et al., 2005; Gerber et al., 2004; Hogan et al., 2008; Riordan et al., 2011) or direct Puf3p binding sites (Fazlollahi et al., 2014; Freeberg et al., 2013; McHugh et al., 2014).

Intriguingly, only Motif 1-UGUA, but not Motif 1-UAUA, is associated with a high prevalence of C at position +2 (Figure 4.3a and S4.4a) which corresponds to the C-binding pocket formed by Puf repeat R8' (Qiu et al., 2012). In contrast, positions upstream of Motif 1-UAUA are predominantly occupied by U or A, suggesting that Puf3p recognition of this motif is structurally distinct from recognition of Motif 1-UGUA such that Puf repeat R8' does not form hydrogen bonds with or constrain the nucleotide at position +2 of Motif 1-UAUA. Puf3p binding sites also harbor a second AU-rich motif (Motif 2), which is present in 188 (5%) sites including 144 (12%) 3'UTR sites (Figure 4.3a). Sites containing Motifs 1 and 2 are enriched at 3'UTRs, and Motifs 1-UGUA and 2 have significantly higher RPM coverage than sites without a motif (Figure 4.3b,c), which is suggestive of a higher affinity of Puf3p for these motifs. Like Puf1p and Puf2p, mRNAs targets of Puf3p with different motifs are enriched for exclusive sets of GO terms, with sites containing Motif 1-UGUA enriched

for mitochondria-related terms (Figure 4.3d).

To test our hypothesis that Puf3p recognition of Motif 1-UGUA-containing sites are structurally distinct from recognition of Motif 1-UAUA-containing sites, we used TEISER (Goodarzi et al., 2012) to search for enriched, short structural motifs (SSMs) in Puf3p target 3'UTRs (details in Methods). TEISER (Tool for Eliciting Informative Structural Elements in RNA) is a computational framework for identifying SSMs - both small hairpin structures and primary sequences - that contain the highest amount of mutual information to explain measurements from a genome- or transcriptome-wide assay. Within the set of 300 nucleotides located directly downstream of all yeast ORFs, which were ranked from highest to lowest cumulative 3'UTR Puf3p binding, we identified a single SSM (SSM1) with high mutual information to explain Puf3p binding (Figure 4.3e and S4.4b). The SSM1 primary sequence resembles Motif 1-UAUA; in fact, more than one quarter of Motif 1-UAUA-containing 3'UTR Puf3p sites contain at least 1 instance of SSM1 (Figure 4.3f). Further, Motif 1-UAUA sites contain, on average, ~5 instances of SSM1, indicating a strong preference for these sites to be highly structured (Figure S4.4c). To refine our search, we re-ran TEISER using a custom database of 3'UTRs supported by experimental RNA-seq data (details in Methods). We identified a second structural motif (SSM2) with a more divergent primary sequence (Figure 4.3g,h), but which is nonetheless enriched for being located in Motif 1-UAUA-containing sites at multiple copies and overlaps 55% of SSM1 sites (Figure S4.4d-f).

Taken together, our data suggest two distinct modes of Puf3p recognition of its targets at 3'UTRs. The first mode is characterized by a strong affinity for an unstructured primary sequence motif with a G at position 2 and a C at position +2 that are constrained by interactions with Puf3p Puf repeats. The second mode is characterized by a weaker affinity for a motif characterized by AU repeats in highly structured regions. It is possible that Puf3p recognition of targets via the second mode is just as robust as the first mode, but given

that PAR-CLIP-seq is biased against capturing structured double-stranded RNA (RNase T1 cleaves single-stranded RNA (Greiner-Stöffele et al., 2000)), we might recover fewer reads from, and thus under-represent, direct Puf3p target sites that form stable hairpin structures.

Two distinct Puf3p binding modes exhibit different target regulation. To determine whether differences in the two modes of Puf3p binding are biologically meaningful, we investigated different features of Puf3p targets. We downloaded steady-state mRNA half-life data derived from dynamic transcriptome analysis (Farazi et al., 2014) and observed that Puf3p targets with 3'UTR binding sites containing Motif 1-UGUA have significantly shorter half-lives (average 11.5 min) than non-targets (average 16.4 min, t -test $p=1.16e-7$; [Figure 4.4a](#)). Specifically, mitochondria-related genes with at least 1 Puf3p 3'UTR site containing Motif 1-UGUA have significantly shorter half-lives (average 9.5 min) than mitochondrial genes not containing Motif 1-UGUA (average 19.3 min; $p=2.17e-3$) ([Figure 4.4a](#)). No significant differences were detected for Puf3p mitochondrial gene targets with Motif 1-UAUA ($p=0.021$) or Motif 2 ($p=0.220$) versus non-targets (data not shown), indicating that repression of mitochondria-related mRNAs is specifically associated with Puf3p 3'UTR sites containing Motif 1-UGUA. Unexpectedly, Puf3p targets with 3'UTR binding sites containing Motif 2 have significantly longer half-lives (average 23.0 min, $p=4.32e-3$) compared to non-targets ([Figure 4.4a](#)), perhaps representing an activating role for Puf3p targeting mRNAs with this motif. No significant differences in half-lives were observed for targets of Puf3p with specific motifs at other genic regions ($p>0.005$, [Figure S4.5a](#)). These data suggest that Puf3p binding to target 3'UTRs results in two distinct outcomes: binding at Motif 1-UGUA-containing sites on target 3'UTRs is associated with target repression while binding of Puf3p to 3'UTRs at Motif 2-containing sites is associated with target stabilization.

To determine how Puf3p binding changes with environmental perturbations, we searched

within the gPAR-CLIP dataset and identified 72% (828 of 1,158) of our 3'UTR Puf3p sites (Figure S4.2e). Compared to 3'UTR binding sites in the entire gPAR-CLIP dataset (1.6-fold reduced binding), sites bound by Puf3p are overall significantly less affected (1.3-fold reduced; Figure 4.4b; t -test $p=3.54e-5$). However, Puf3p sites containing different motifs respond to glucose starvation differently. Sites containing Motif 1-UGUA show 2.3-fold reduced binding, which is significantly lower than all 3'UTR Puf3p sites ($p=9.18e-7$) and gPAR-CLIP sites ($p=1.05e-3$). Interestingly, genes with reduced binding at Motif 1-UGUA are up-regulated only 1.1-fold compared to all genes ($p=0.003$; Figure S4.6a), suggesting that Motif 1-UGUA might be a more specific regulator than a general regulator of target silencing during a glucose stress response. Indeed, when we narrow the list of Motif 1-UGUA sites to those only on genes annotated with the GO term “mitochondria”, we observe a 2.3-fold reduction of binding ($p=5.5e-10$, Figure 4.4c) as well as a significant 2-fold upregulation in mRNA levels compared to non-mitochondria-related Puf3p target genes ($p=1.24e-5$), which were unchanged (1.1-fold upregulated, Figure 4.4c). We see no change in Puf3p site coverage or gene expression in nitrogen starvation conditions ($p=0.5751$; Figure S4.5b), suggesting that Puf3p regulation of mitochondrial targets via Motif 1-UGUA is a specific response to changes in glucose levels but not nitrogen levels. Unexpectedly, sites with Motif 2 are significantly more bound than all gPAR-CLIP sites (1.3-fold up-regulated, $p=6.15e-8$), and the gene targets are significantly repressed (1.8-fold down-regulated, $p=4.45e-6$) compared to all genes, suggesting that these sites respond differently to glucose starvation than Puf3p sites with Motif 1-UGUA. We did not observe any significant difference in Puf3p binding site coverage in nitrogen starvation conditions (all $p>0.001$, Figure S4.3b), indicating that Puf3p binding does not globally respond to nitrogen starvation. Expression changes of 3'UTR Puf3p target genes in either glucose or nitrogen starvation conditions did not significantly differ ($p=0.421$ and $p=0.280$, respectively) based on presence of either structural motif SSM1 or

SSM2, suggesting that Puf3p binding at these structural motifs might not be involved in PTGR in response to these environmental stresses (Figure S4.5c-d).

Taken together, our data show that, in response to glucose starvation, Puf3p decreases binding to Motif 1-UGUA sites on genes involved in metabolic/mitochondrial processes, perhaps in an effort to promote ATP generation. This agrees with known metabolic remodeling that occurs during glucose starvation of yeast (Brauer et al., 2005; Gasch et al., 2000). We observe that primary sequence motifs direct distinct modes of Puf3p binding that target different sets of genes, have different effects on steady state target levels, and have different responses to environmental conditions. Moreover, only Puf3p sites containing Motif 1-UGUA, in particular sites on mitochondria-related genes correlate with shorter mRNA half-lives, while only Puf3p sites with Motif 2 correlate with longer half-lives, suggesting repressive and activating roles, respectively, for these Puf3p motifs. Our results provide *in vivo* evidence for the pervasiveness of a dynamic mRNA stability response to environmental stresses mediated by a single Puf.

Puf4p and Puf5p have distinct binding characteristics. Puf4p promotes degradation of endogenous target mRNA *HO* through direct binding of the Ccr4p/Pop2p complex to enhance deadenylation (Goldstrohm et al., 2006, 2007; Hook et al., 2007; Lee et al., 2010). Puf4p has also been shown to contribute to degradation of genes encoding ribosomal proteins and rRNA synthesis and ribosome assembly factors (Grigull et al., 2004). Based on *in vitro* selection studies (Campbell et al., 2014; Gerber et al., 2004, 2006) and determination of the crystal structure of Puf4p in complex with *HO* (Valley et al., 2012), a primary recognition motif of UGUAWAWUA has been suggested for Puf4p. However, besides the handful of identified Puf4p targets, the extent of Puf4p binding to the transcriptome and whether Puf4p specifically or generally functions in target repression remains unknown.

Using hierarchical motif searching, we identify a top motif (Motif 1) in 312 (10%) of all 3,002 Puf4p binding sites including 232 (43%) of 540 3'UTR sites. This motif contains two distinct sub-motifs: UGUAWAWUA (Motif 1-UGUA) identified in 198 sites and UAUAWAWUA (Motif 1-UAUA) identified in 114 sites (Figure 4.5a). Unlike Puf3p, Puf4p Motif 1-UGUA is not enriched for +2 C (Figure S4.4a), which agrees with a lack of the R8' Puf repeat in Puf4p (Qiu et al., 2012). The crystal structure of Puf4p in complex with target RNA revealed that RNA positions U5 and U7 are “flipped-out”; in fact, U5 forms no hydrogen bonds with Puf repeat R4 (Miller et al., 2008). Instead, the histidine (H688) side-chain in repeat R4 forms a stacking interaction with A6, suggesting that the U5 position could be occupied by any nucleotide. Indeed, our PAR-CLIP-seq-derived Puf4p binding sites are highly variable at positions 5 and 7 (Figure 4.5a); in addition, position 9 is also highly variable, suggesting that the nucleotide base at this position might also be “flipped-out” and unconstrained. We identify a second AU-rich motif (Motif 2) in 196 (7%) of all Puf4p binding sites including 132 (24%) of 3'UTR sites. Sites containing Motifs 1 or 2 are enriched at 3'UTRs, have significantly higher RPM coverage than sites without a motif, and occur within Puf4p 3'UTR sites on distinct sets of genes (Figure 4.5b-d).

Using TEISER and our custom 3'UTR database, we identified a short, structural motif (SSM3) enriched among 3'UTRs of Puf4p targets (Figure 4.5e). SSM3 is AU-rich and overlaps 25% of Puf4p 3'UTR binding sites containing Motif 2 and 17% of sites containing Motif 1-UAUA with an average of 5 and 3.5 instances of SSM3 per binding site, respectively (Figure 4.5f and S4.7a,c). As with Puf3p, these data suggest one mode of Puf4p target recognition driven by affinity for a primary sequence (Motif 1-UGUA) and a second mode of target recognition driven by affinity for an UA-rich, structural motif. When we examined mRNA half-life data of Puf4p targets, no significant differences were observed between genes targeted by Puf4p at different genic regions or with different motifs (Figure S4.7b). These

observations suggest that, while Puf4p is required for regulating target stability in a few validated cases, its global effect on mRNA stability might be too weak to detect through examination of global mRNA half-life measurements.

Similar to 3'UTR sites bound by Puf3p, sites bound by Puf4p are significantly less affected by glucose starvation (1.0-fold change) than all sites (1.6-fold reduced binding, *t*-test $p=1.20e-8$). Unlike Puf3p sites, however, both versions of Puf4p Motif 1 are associated with increased binding upon glucose starvation (1.1-fold increase for Motif 1-UGUA sites, 1.7-fold increase for Motif 1-UAUA sites). Combined with known decreased translation under stress, these observations support a model in which Puf4p binding to either Motif 1-UGUA or -UAUA on ribosome-related genes is increased under glucose deprivation in an effort to silence them and inhibit translation. Indeed, expression of genes bearing Puf4p sites with either Motif 1 is significantly reduced in glucose starvation conditions (Figure S4.6a).

Puf5p has a largely repressive role in PTGR of its targets (Quenault et al., 2011), many of which are bound at their 3'UTRs (Seay et al., 2006). Like Puf4p, Puf5p interacts with the Ccr4p/Pop2p deadenylase complex via direct binding to Pop2p (Goldstrohm et al., 2006) and also promotes recruitment of Dhh1p (decapping activator) and Dcp1p (decapping enzymes) to the 5' end of mRNAs targeted for degradation (Maillet and Collart, 2002; Riordan et al., 2011). In an *in vivo* yeast three-hybrid screen, a handful of direct targets of Puf5p were identified, and binding of Puf5p to these targets was confirmed to enhance their degradation (Seay et al., 2006). An *in silico* search through the full-length sequences of Puf5p target mRNAs identified an enriched motif that is similar to those for Puf3p and Puf4p, although the motif tends to be longer and more flexible in the number of nucleotides recognized after the core UGUA sequence (Gerber et al., 2004, 2006; Hogan et al., 2008).

We confirm the Puf5p binding site in the 3'UTR of *RAX2*, which is required for Puf5p direct binding to promote degradation of the *RAX2* transcript (Seay et al., 2006). Among all

Puf5p sites, the top motif (Motif 1), found in 497 (15%) of our PAR-CLIP-seq binding sites including 372 (28%) of 3'UTR sites, contains two distinct sub-motifs: UGUAW₄UW (Motif 1-UGUA) identified in 328 sites and UAUAW₄UW (Motif 1-UAUA) identified in 169 sites (Figure 4.6a). Very few Puf5p sites with Motif 1-UGUA contain a +2 C (Figure S4.4a), in agreement with the lack of a Puf R8' repeat (Qiu et al., 2012). The second most significant Puf5p motif (Motif 2) is a U-rich motif found in 312 sites with a strong prevalence of C at position 5. A third motif (Motif 3) was identified in 191 Puf5p binding sites and is similar to Motif 1 in that it includes UGUA at the 5' end of the motif. However, instead of four nucleotides separating the UGUA from the downstream UW part of the motif, Motif 3 only has three nucleotides separating the UGUA from the downstream UW. The variation of this region could reflect the flexibility of the structure of the Puf:RNA interaction at this location. Motifs 1-3-containing sites are enriched on 3'UTRs and are more strongly bound than Puf5p sites with no motif (Figure 4.6b,c). Strikingly, genes annotated with cell wall-related GO terms are promiscuously bound by Puf5p except when the Puf5p site contains Motif 1-UGUA (Figure 4.6d). This is in contrast to previous work that analyzed *in vitro* targets of Puf5p and saw enrichment for genes encoding chromatin modifiers and spindle pole body components (Gerber et al., 2004; Seay et al., 2006).

Despite roles for Puf5p in promoting degradation of a few known targets, Puf5p targets did not globally exhibit significantly shorter half-lives; in fact, genes with Motif 2-containing Puf5p sites in their 3'UTRs had significantly longer half-lives (Figure 4.6e and S4.8), similar to Puf3p targets with Motif 2 (Figure 4.4a). Combining gPAR-CLIP-seq data with Puf5p sites, we see very little difference between the global 1.6-fold reduced coverage of all 3'UTR sites and changes in Puf5p sites (although sites with Motif 1-UAUA are slightly, but significantly, less affected by glucose deprivation; Figure 4.6f).

Taken together, these results show subtle differences in sets of genes targeted by Puf4p

and Puf5p and target stability measurements based on the presence of one of many distinct primary sequence motifs or, in the case of Puf4p, by presence of a secondary structural motif. A common theme across Puf3/4/5p direct binding sites is the presence of “UGUA” versus “UAUA” as the first 4 nucleotides of the strongest motif appearing to be a major factor in the distinguishing between associations with different biological measurements. In addition, none of the Pufs so far examined appear to respond to nitrogen starvation in terms of increased or reduced binding to target 3'UTR sites (Figure S4.3b), suggesting that glycolysis and metabolic cues are upstream signalers of Puf-mediated PTGR, but not cues relating to translation (*i.e.* disruption in amino acid synthesis from lack of nitrogen).

Puf6p binds snoRNAs and rRNAs in snoRNP complexes. Puf6p is the least understood Puf protein in yeast. Puf6p is known to translationally repress and asymmetrically localize a single target, *ASH1*, to the bud tip of dividing yeast cell (Deng et al., 2008; Gu et al., 2004). Puf6p was recently shown to interact with Loc1p and She2p in the nucleus as part of a co-transcriptional mechanism to establish the post-transcriptional regulatory fate of bud-localized transcripts (Shahbadian et al., 2014). The major function of Puf6p, though, is in ribosome biogenesis and assembly in the nucleolus (Li et al., 2009c), where most of Puf6p is localized. A major component of ribosome biogenesis is the class of small nucleolar RNAs (snoRNAs), which are non-coding RNAs required for the maturation of pre-rRNAs through cleavage, pseudouridylation, and methylation mechanisms (Davis and Ares, 2006; Schattner et al., 2004; Torchet et al., 2005). From our PAR-CLIP-seq data, we found that Puf6p preferentially binds snoRNAs and coding regions of mRNAs (Figure 4.1c). We were unable to detect any statistically significant primary sequence motifs or short structural motifs despite searching within all sites, just CDS sites, or just snoRNA sites (data not shown). This agrees with a recent study showing that Puf6p promiscuously binds to double- and single-stranded

DNA and RNA (Qiu et al., 2014). We also observed no preferences for Puf6p binding different snoRNAs based on snoRNA type (box C/D or H/ACA), genomic organization, rRNA modification type, or rRNA species modified (Figures 4.7a and S4.9a).

In our PAR-CIP-seq dataset, Puf6p directly binds 66 of the 77 annotated snoRNAs with an average of 2.3 Puf6p sites per snoRNA and as many as 8 Puf6p sites on snR30. Puf6p also binds extensively to rRNAs and pre-rRNA regions (70% of total reads, Figure S4.9b), consistent with the nucleolar localization of Puf6p for 60S ribosome biogenesis (Gavin et al., 2002; Huh et al., 2003). Specifically, we identify a Puf6p-bound region on the C/D box U14 snoRNA that overlaps the Box C region and sits adjacent to a site that interacts with 18S rRNA (Figure 4.7b). We also identify two Puf6p sites on structured arms of the C/D box U3 snoRNA near the Box C regions (Figure 4.7c). Binding of snoRNAs and rRNAs by Puf6p, combined with several lines of evidence, points to a likely role for Puf6p in promoting snoRNA-mediated processing of pre-rRNAs or in ribosome assembly, rather than in pre-snoRNA processing. First, pre-snoRNA processing requires multiple factors including Pap1p, Trf4p, Trf5p, Nrd1p, Nab3p, Rrp6p, Sen1p, Pxr1p, and Rntp1. Searching BIOGRID (Stark et al., 2006), we found no published data supporting physical interactions between any of these factors and Puf6p. Second, Box C/D snoRNA-mediated maturation of pre-rRNAs to mature rRNAs requires Nop1p, Nop56p, Nop58p, Snu13p, and dozens of Utp proteins, while Box H/ACA snoRNA-mediated maturation of rRNAs requires Cbf5p, Gar1p, Nhp2p, and Nop10p (reviewed in (Woolford and Baserga, 2013)). Affinity capture-MS data in BIOGRID support a physical interaction between Puf6p and five of these factors (Nop1p, Nop56p, Nop58p, Cbf5p, and Gar1p), suggesting that Puf6p directly interacts with components responsible for pre-rRNA maturation. Taken together, these data indicate that Puf6p is most likely involved in the function of snoRNAs to modify pre-rRNAs or as a link between pre-rRNA processing and incorporation of rRNAs into mature ribosomal complexes,

not in the processing of pre-snoRNAs into their mature forms.

Upon nitrogen starvation, we observed a 1.6-fold decrease (t -test $p=2.1e-3$, **Figure 4.7e**) in snoRNA expression, which is not unexpected since a shortage of nitrogen inhibits amino acid biosynthesis and stalls translation, decreasing the need for ribosome biogenesis (Hinnebusch, 2005). Of the 150 Puf6p sites identified on snoRNAs by PAR-CLIP-seq, 107 (71%) are represented in the gPAR-CIP-seq dataset. Coverage of Puf6p-bound snoRNA sites increased 12.6-fold upon nitrogen starvation (**Figure 4.7d**). A milder effect was observed in glucose starvation conditions: snoRNA expression decreased 1.2-fold while Puf6p sites increased 3.6-fold in coverage (**Figure S4.9c,d**). These data suggest that Puf6p binding of snoRNAs is anti-correlated with expression snoRNAs. Taken together, our PAR-CLIP-seq data, combined with evidence of direct Puf6p and snoRNP component interactions, suggest a role for Puf6p as part of the snoRNP complex, perhaps to promote pre-rRNA maturation and assembly of ribosomal subunits. SnoRNA expression and snoRNA binding by Puf6p are greatly affected by availability of nitrogen and, to a lesser degree, glucose. By examining novel Puf6p mRNA targets, we can begin to infer additional roles for Puf6p in perhaps in trafficking of additional mRNAs, not just *ASH1*, to cell membranes and/or the tip of budding yeast cells.

Co-regulation of targets by multiple Pufs. Of all 3,006 protein-coding genes bound by at least one Puf, 47% (1,407) are bound by two or more Pufs while 37% (669 of 1,793) of Puf-bound 3'UTRs are bound by two or more Pufs, suggesting widespread co-regulation by this family of RBPs. Hundreds of targets are co-bound by any two Pufs (**Figure 4.8a**), with the greatest proportion of shared targets being co-bound by Puf3p, Puf4p, and Puf5p (**Figure 4.8a-c**). This agrees with the similar recognition motifs we identified within the direct binding sites of these Pufs (**Figures 4.3, 4.5, and 4.6**). Functional analysis of 3'UTRs targeted exclusively by two of these Pufs reveal striking features. For example, 3'UTR targets of Puf3p and Puf4p are enriched for genes involved in cellular responses to environmental

signals (*e.g.* oxygen radicals, copper ions) and vesicle localizations (Figure 4.8e). 3'UTR targets of Puf3p and Puf5p are enriched for mitochondrial-related genes, similar to functional annotations of all Puf3p targets. Finally, shared 3'UTR targets of Puf4p and Puf5p are enriched for ribosome biogenesis and localization factors, similar to functional annotations for Puf4p targets. Together, these observations highlight not only the pervasiveness of co-regulation by multiple Pufs but also that co-regulation occurs on functionally distinct sets of genes, hinting at diverse roles for single- and multi-Puf-dependent mechanisms of PTGR.

4.3 Materials and methods

Strains, media, and growth conditions

The following TAP-tagged strains for Puf1p-Puf6p were picked from a TAP-tagged yeast strain collection (Ghaemmaghami et al., 2003): (*MATa hisΔ1 leu2Δ0 lys2Δ0 ura3Δ0 [PUF]-TAP::HIS5*). Strains were grown as described in (Freeberg et al., 2013) in synthetic defined media.

PAR-CLIP and sequencing procedures

Puf1p-Puf6p PAR-CLIP was performed as described for Puf3p PAR-CLIP in Freeberg et al. (2013). Briefly, yeast were grown to mid-log-phase and irradiated with 365 nm UV. Crosslinked cells were lysed, treated with RNase T₁, and mixed with IgG magnetic beads to pull down each TAP-tagged Puf. Lysates were then subjected to RNase T₁ digestion, CIP treatment, 3' DNA linker ligation, 5' end phosphorylation, and SDS-PAGE. After nitrocellulose transfer, crosslinked RNAs were visualized by autoradiography. Bands corresponding to each Puf protein were excised and incubated with Proteinase K. RNAs were collected by centrifugation and loaded onto a 6% TBE UREA gel. Gel pieces corresponding to 70-90 nt RNA were

excised followed by amplification of the RNA fragments by RT-PCR. Amplicons were purified, run on a 10% TBE gel, and gel pieces corresponding to 96-116 bp DNA were excised. DNA fragments were amplified by PCR for two rounds and sequenced on an Illumina HiSeq 2000 sequencer.

Oligonucleotide sequences for constructing sequencing libraries

All primers used are as listed in Freeberg et al. (2013). Specifically, Index 1 through 6 bar-coded 3' DNA linker oligos and RT primers were used for Puf1p-Puf6p libraries, respectively.

Read processing

Read data were processed as described for Puf3p PAR-CLIP-seq in Freeberg et al. (2013). Briefly, reads were processed to remove linkers and sorted into libraries based on six-nucleotide barcodes. Next, reads were removed if they met any of the following criteria: <18 nucleotides, only homopolymer As, missing 3' adapter, 5'-3' adapter ligation products, 5'-5' adapter ligation products, low quality (more than 4 bases with quality scores below 10 or more than 6 bases with a quality score below 13). High-quality reads were mapped to the *S. cerevisiae* genome (S288C, sacCer3) with Bowtie (Langmead et al., 2009) using the following parameters: -v 3 (map with up to 3 mismatches), -k 275 (map at up to 275 loci), -best, and -strata.

Binding site generation

Reads were assembled into binding sites by aggregating overlapping reads harboring 0-2 T-to-C conversion events. Only binding sites containing at least 1 T-to-C conversion event were considered high-confidence binding sites. For each library, the counts of sequencing reads covering each position within a binding site was averaged and normalized to the total number of millions of mapped reads in that library. To filter off low-coverage binding sites, a reads

per million mapped reads (RPM) threshold for each library was empirically determined by simulating replicate data from each PAR-CLIP-seq dataset. Two sets of binding site RPM values were randomly sampled from all binding sites passing a minimum RPM threshold in a single dataset such that each sample contained 20% of the binding sites. A non-parametric two-sample Kolmogorov-Smirnov (K-S) test was performed on the two sets of RPM values and the resulting K-S test statistic was recorded. This test was repeated 10,000 times for each of 36 RPM threshold values ranging from 0 to 25. Mean K-S test statistic values were plotted for each RPM threshold value and a final binding site RPM threshold value for the library was chosen when the K-S test statistic stabilized ([Figure S4.1d](#)). The following empirical RPM thresholds were obtained: 1.5 RPM (Puf1p), 1.5 RPM (Puf2p), 3 RPM (Puf3p), 0.5 RPM (Puf4p), 5 RPM (Puf5p), and 3.5 RPM (Puf6p). After filtering, binding site RPM values were normalized to gene expression RPKM values from Freeberg et al. (2013).

Binding sites were annotated using custom scripts to known genomic elements in the *S. cerevisiae* genome (S288C, sacCer3). ORFs with unannotated untranslated regions (UTRs) were hierarchically assigned UTRs from the following: Nagalakshmi et al. (2008), Yassour et al. (2009), and a “predicted” UTRome defined by extending 80nt upstream from the start codon (for 5’UTRs) or 138nt downstream of the stop codon (for 3’UTRs) of ORFs still lacking a 5’UTR or 3’UTR. These lengths were chosen as the mean length of 5’UTRs and 3’UTRs, respectively, as determined by Nagalakshmi et al. (2008).

Computational analyses

Significance testing

All *t*-tests performed are Welch’s two-tailed unless otherwise noted. Tests are corrected for multiple hypothesis testing using the Bonferroni correction method. Differences were considered significant if p -value < 0.005, unless otherwise noted.

GO term enrichment

GO term enrichment was performed using the topGO (v2.18.0) R package (Alexa et al., 2006). GO terms were considered “enriched” if they passed a p -value threshold of $1.0e-4$ after Bonferroni correction. The following terms were omitted from figures due to their redundancy with other terms: cytosolic small/large ribosomal subunit (GO:0022626/0022625), cytoplasmic translation (GO:0002181), mitochondrial small/large ribosomal subunit (GO:0005763/0005762), structural constituent of ribosome (GO:0003735), glycolytic process (GO:0006096), cytosol (GO:0005829), translation (GO:0006412). Genes referred to in the text as “mitochondria-related” are those with annotated GO term “mitochondrion” (GO:0005739).

Hierarchical motif identification

For each Puf, sequences (plus 10 nt flanking each direction) of the top 200 binding sites were searched using MEME (Bailey and Elkan, 1994) for the single best motif, which was then used as input to MAST (Bailey and Gribskov, 1998) to find all occurrences of the motif in all binding sites. Sites containing the motif were then removed from the pool of all binding sites, and the process was repeated starting with the new set of top 200 binding sites until no significant motifs were found.

Half-life measurements

Measurements of steady-state half-lives for yeast transcripts were taken from the dynamic transcriptome analysis measurements in Miller et al. (2011).

TEISER

The following parameters for TEISER (Goodarzi et al., 2012) version 1.0 were used to ini-

tialize seeds: -min_stem_length 4 -max_stem_length 7 -min_loop_length 4 -max_loop_length 9
-min_inf_seq 4 -max_inf_seq 6 -min_inf 14 -max_inf 20.

The following parameters were used for initial structural motif searches: -expfile=<EXP>
-species=yeast -seedfile=<SEEDS> -exptype=continuous -ebins=30 -submit=1. <EXP>
is a file containing all yeast ORFs and the sum of Puf binding to each ORF's 3'UTR.
<SEEDS> is the initialized seed motif file generated in the first step.

The following parameters were used for the second round of structural motif searches:
-expfile=<EXP> -fastafile_dn=<FASTA> -species=yeast -seedfile=<SEEDS> -exptype=
continuous -doconservation=1 -dopagerun=0 -doonlypositive=0 -submit=1 -doremovedups=
0 -ebins=30. <EXP> is a file containing all yeast ORFs and the sum of Puf binding to
each ORF's 3'UTR. <SEEDS> is the initialized seed motif file generated in the first step.
<FASTA> is a custom fasta file of ORF 3'UTRs as described above.

4.4 Discussion and concluding remarks

The work presented here offers a comprehensive survey of the differences and similarities of substrate binding and rules for binding for an important family of conserved RBPs and represents a major step towards uncovering the pervasiveness of PTGR by a single family of RBPs. For the first time, the direct *in vivo* RNA binding sites of the Pumilio family of RBPs have been identified with nucleotide resolution, leading to key insights into their individual binding affinities, potential sequence determinants, dynamic changes in binding under nutrient-limiting conditions, and identities of novel RNA substrates. Largely through *in vitro* selection and single-gene analyses, the yeast Puf proteins, in aggregate, were thought to interact with 12% of the yeast protein-coding transcriptome (Gerber et al., 2004). By our PAR-CLIP-seq analyses, we observe that yeast Puf1-6p interact with over half of the

yeast coding transcriptome and one third of 3'UTRs, underscoring the importance of this family of RBPs in PTGR. In addition, the set of genes targeted by any one Puf protein are highly unique depending on whether Puf targets a gene at its 3'UTR, CDS, or 5'UTR (Figure S4.10). Further, we identified functionally distinct sets of targets for each Puf based on genic region, highlighting the necessity of identifying precise locations of Puf binding across a transcript. The use of PAR-CLIP-seq to identify direct Puf targets represents an improvement in specificity and relevance over *in vitro* experiments that identify whole transcript substrates out of biological context (Campbell et al., 2012, 2014; Gerber et al., 2004, 2006; Hogan et al., 2008).

Dual regulatory modes for Puf3p and Puf4p. Puf3p and Puf4p appear to bind target 3'UTRs via two distinct modes (Figures 4.3 and 4.5). In the first mode, each Puf recognizes a strong single-stranded motif characterized by the conserved UGUA sequence and, for Puf3p, a +2C. In the second mode, each Puf recognizes an AU-repeat motif characterized by overlap with a stable hairpin structure and, for Puf3p, lacking a +2C. These observations suggest that a single-stranded UGUA sequence is not the only feature necessary to promote a strong Puf3p- or Puf4p-RNA interaction. Further, this is the first time that two structurally distinct modes of binding have been characterized for a single RNA-binding domain in a Puf, hinting that while the structure of Puf repeat domains might be rigid, the interface between Puf and its substrate RNA might be more flexible and allow a range of interaction conformations to occur.

Evidence for Puf proteins as mRNA stabilizers. We show that presence of Motif 1-UGUA in Puf3p sites is correlated with significantly shorter mRNA half-lives than non-Puf3p targets (Figure 4.4a), in line with what has been experimentally shown for specific Puf3p targets (Jackson et al., 2004; Olivas and Parker, 2000; Saint-Georges et al., 2008). Interestingly, we observe that Puf3p target sites containing either Motif 1-UAUA or Motif

2 have longer half-lives than non-Puf3p targets, suggestive of a stabilizing role for Puf3p of these targets. The ability of Puf proteins to promote target expression is an emerging concept that has only been recently shown in a few isolated instances (Quenault et al., 2011). In *C. elegans*, Puf protein FBF acts as a usual repressor by binding the *gld-1* 3'UTR and, through physically interacting with the CCF-1/Pop2p deadenylase complex *in vitro*, represses *gld-1* expression (Suh et al., 2009). However, FBF can also enhance the activity of poly(A) polymerases GLD-2 and GLD-3 which in turn activates translation of the same *gld-1* target (Suh et al., 2009). Translational activation of another direct target of FBF, *egl-4*, in sensory neurons requires FBF binding to its 3'UTR in response to stimulation, although the mechanism of this regulation is still unclear (Kaye et al., 2009). In *Xenopus* oocytes, the Puf protein Pum is recruited to the 3'UTR of the cyclin B1 mRNA and stabilizes the interaction of the mRNA with the cytoplasmic polyadenylation element binding protein, thus promoting translation activation (Piqué et al., 2008). In *Trypanosoma brucei*, Puf9 stabilizes its mRNA substrates during S-phase, perhaps through outcompeting binding of a repressor to the same mRNA substrates (Archer et al., 2009). Our analysis opens up the possibility that stabilizing and/or activating Puf:RNA interactions are a much more pervasive phenomenon than previously considered and may co-exist with repressive interactions based on Puf interactions with other protein cofactors (as in the case with FBF) and/or substrate RNAs. It remains to be determined whether global activating Puf:RNA interactions are a result of mechanisms that antagonize other repressive RBP:RNA interactions, promote mRNA translation, or both.

Pumilio are diverse regulators of gene expression in response to nutrient stress. Mounting evidence supports a highly diverse regulatory role for Puf3p in adapting yeast to environmental perturbations. Recently, Puf3p was shown to bind its substrate mRNAs at maintained or increased levels upon oxidative stress to direct translational repression

or mRNA degradation, suggestive of a role in down-regulating expression of proteins targeted to mitochondria (Rowe et al., 2014). Another study revealed that Puf3p represses target mRNAs in glucose conditions when mitochondria are not needed but is inhibited by ethanol, galactose, and raffinose when mitochondria are required for growth, resulting in target stabilization (Miller et al., 2014). Through incorporation of global PAR-CLIP-seq data, we show that Puf1-6p respond differently to glucose and nitrogen starvation, and this difference appears to depend on the motif present in the Puf binding sites (Figures 4.2, 4.4-4.7). Specifically, Puf3p Motif 1-UGUA sites are enriched on mitochondria-related targets and appear less bound upon glucose starvation, resulting in higher levels of the associated mRNAs. We did not observe this trend for Puf3p Motif 2 sites which are not enriched on mitochondria-related genes, supporting the idea of distinct regulatory roles for Puf3p that depend on, at least in part, the nature of the interaction between Puf3p and its substrate mRNA.

Puf6p binds snoRNAs and rRNA in snoRNP complexes. One of the most surprising findings was the preference of Puf6p to bind snoRNAs instead of coding mRNAs (Figures 4.1 and 4.7). This feature sets Puf6p apart from its sister Pufs and represents an important insight in our understanding of the role of Puf6p in PTGR. Lack of identification of a primary sequence motif within Puf6p binding sites is in agreement with recent structural features of the crystal structure of the Puf6p homolog in humans, PUF-A (Qiu et al., 2014). The predicted structure of Puf6p based on PUF-A in combination with mutational studies of key Puf6p residues suggests that these Pufs are able to recognize and bind both single- and double-stranded RNA and DNA molecules, with no apparent sequence requirements. Instead, specific patches of basic residues on these proteins are required for efficient interactions with a nucleic acid molecule. Identification of Puf6p binding sites on almost all of the annotated snoRNAs in yeast, as well as extensive direct binding to rRNAs, strongly

supports Puf6p as a component of nucleolar snoRNPs that process pre-rRNAs into mature rRNAs. Identifying the distinct sites of Puf6p interaction on these non-coding RNAs provides a starting point for future work to determine the exact mechanism by which Puf6p is acting to regulate snoRNA-mediated rRNA maturation.

Taken together, our analysis of *in vivo* Puf substrates reveals that PTGR by Puf proteins is highly pervasive as evidenced by Pufs binding over half of protein-coding genes and a third of all 3'UTRs. PTGR by Pufs is also greatly coordinated with almost half of all Puf substrates bound by more than one Puf. Utilizing the direct Puf binding sites identified by PAR-CLIP-seq, we establish multiple primary recognition motifs for Puf1-5p, some of which coincide with predicted stable hairpin structures to establish distinct structural conformations of Puf:RNA complexes. The similarities and differences observed between features of Puf binding and responses to nutrient limitations reflect a highly dynamic role for Puf-mediated post-transcriptional gene regulation.

4.5 Data availability

Raw sequence data are available through the NCBI Gene Expression Omnibus using series entry GSE43747 (Puf3p) and GSE67064 (all other Pufs).

4.6 Acknowledgements

Ting Han performed the PAR-CLIP-seq experimental methods associated with this study. This work benefitted from thoughtful discussions with Aaron Goldstrohm and Danny Yang. The study and manuscript were conceived and written by MAF and JKK.

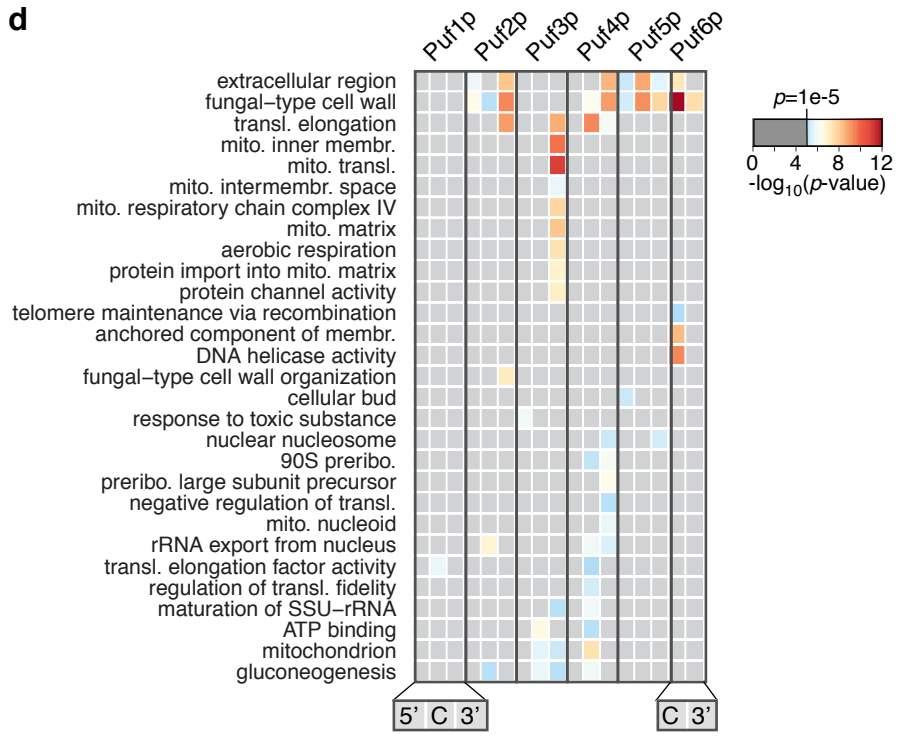
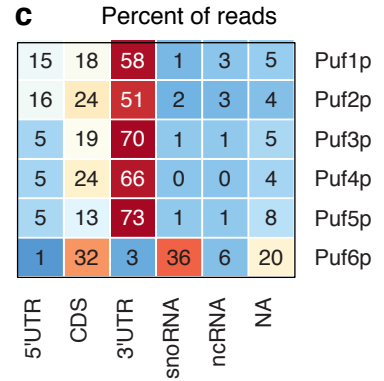
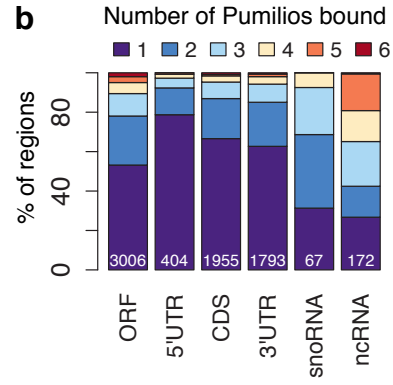
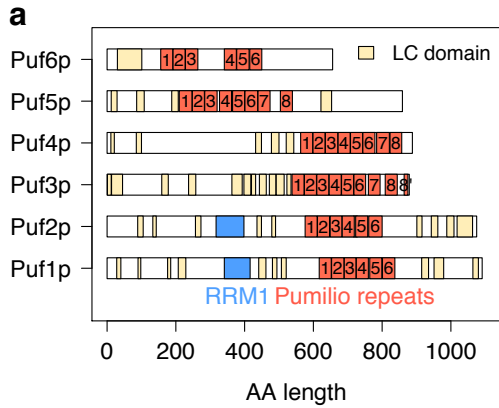


Figure 4.1: (a) Domain architecture for the six characterized yeast Puf proteins. (b) Direct Puf targets are bound by multiple Puf proteins. Numbers in white indicate the total number of elements bound by at least one Puf protein. Puf sites are identified on 7% (404 of 5,525) of annotated 5' UTRs, 33% (1,955 of 5,911) of coding regions, and 32% (1,793 of 5,684) of annotated 3'UTRs. (c) Puf1-5p bind primarily at 3'UTRs while Puf6p binds primarily at coding regions and snoRNAs. Heatmap colors represent row-normalized proportion of reads mapping to each genic region. Heatmap numbers indicate the percentage of all non-rRNA reads mapping to each genic region. NA: not annotated. (d) Puf targets are enriched for different GO terms depending on the genic region bound. Heatmap colors represent $-\log_{10}(p\text{-values})$ from enrichment testing.

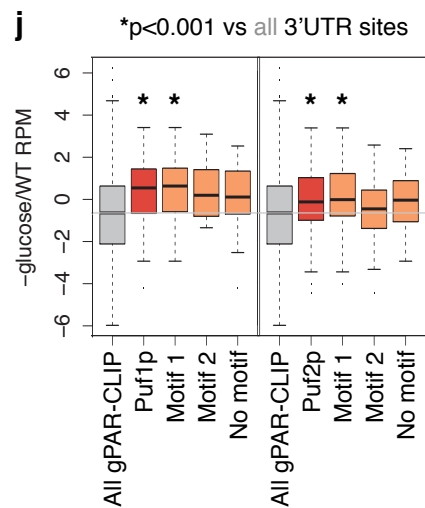
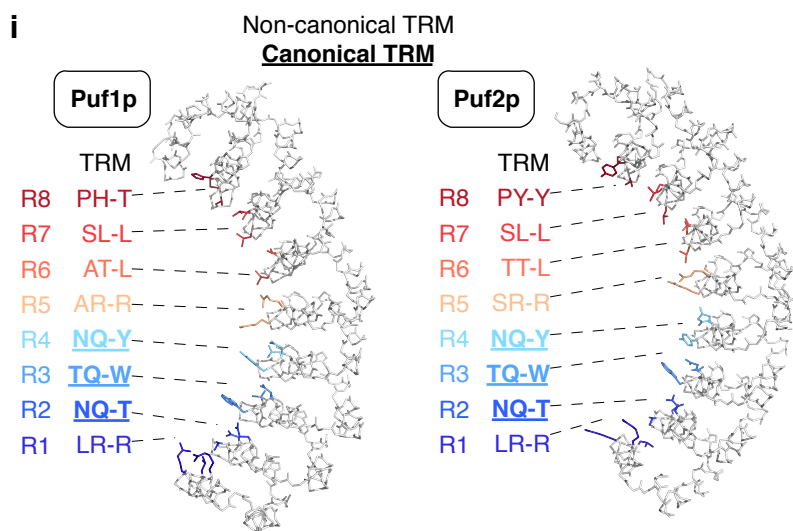
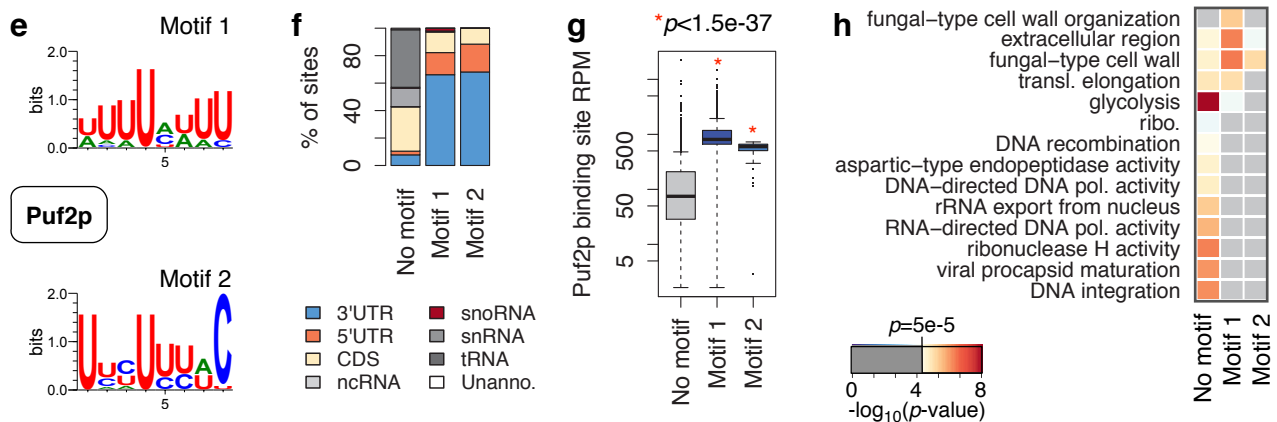
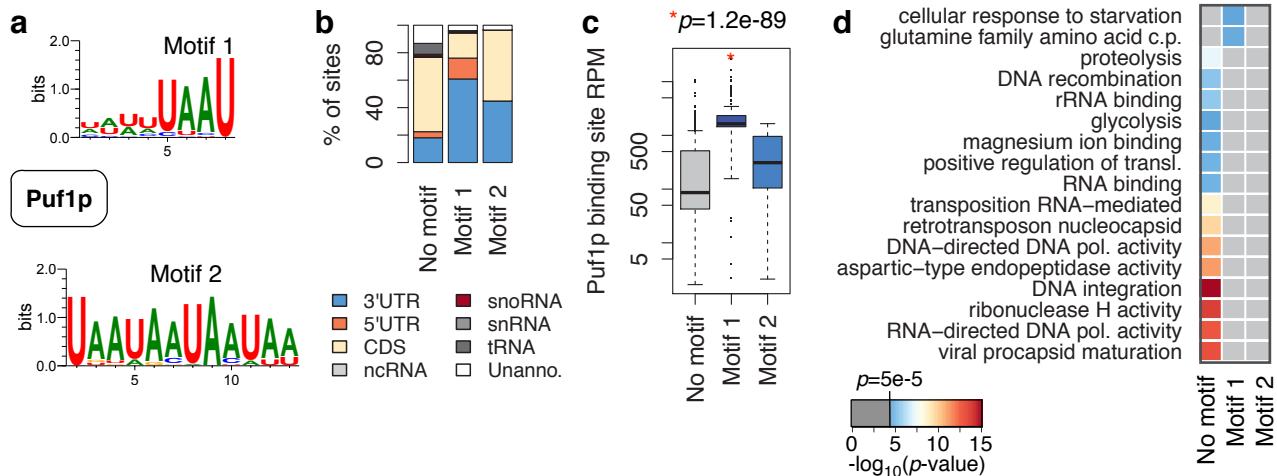


Figure 4.2: (a) Two UAAU-containing motifs are enriched among Puf1p binding sites. (b) Puf1p sites with Motif 1 are enriched among 3'UTRs. Puf1p sites with Motif 2 are enriched among 3'UTRs and CDS. (c) Puf1p sites with Motif 1 ($p=1.2e-89$), but not Motif 2 ($p=0.099$), are more strongly bound than non-motif-containing sites. (d) Genes with motif-containing Puf1p binding sites are enriched for distinct GO terms. (e) Two U-rich motifs are enriched among Puf2p binding sites. (f) Puf2p sites with Motif 1 or 2 are enriched among 3'UTRs. (g) Motif-containing Puf2p binding sites are more strongly bound than non-motif-containing sites ($p<1.5e-37$). (h) Genes with motif-containing Puf2p binding sites are enriched for distinct GO terms. (i) Predicted Puf1p and Puf2p Puf repeat domain structures. Colored are the side chains of amino acids in indicated TRMs. (j) Puf1p and Puf2p sites on 3'UTRs containing Motif 1 are significantly more bound upon glucose starvation than all sites ($p<0.001$).

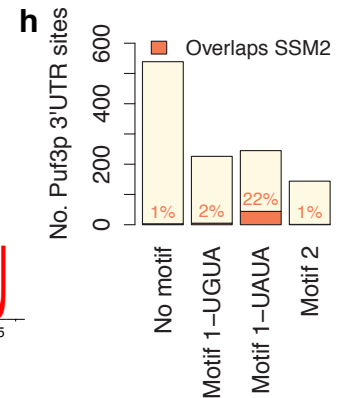
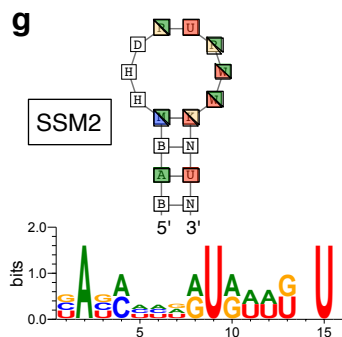
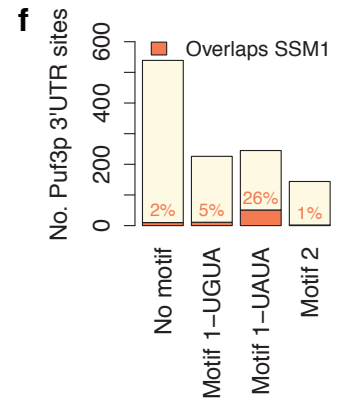
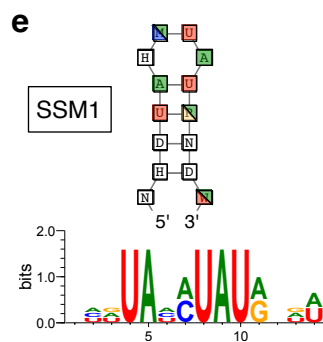
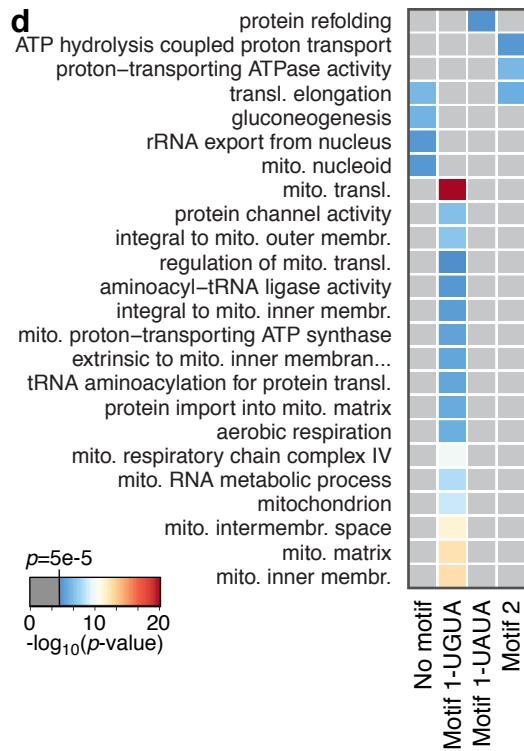
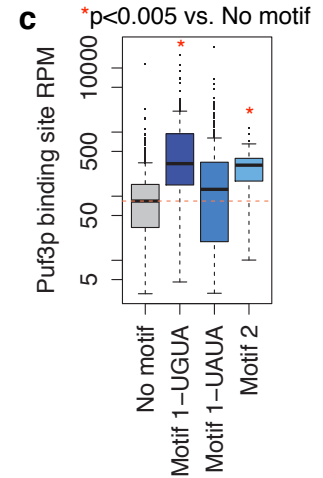
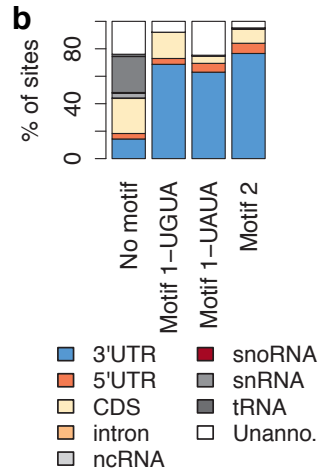
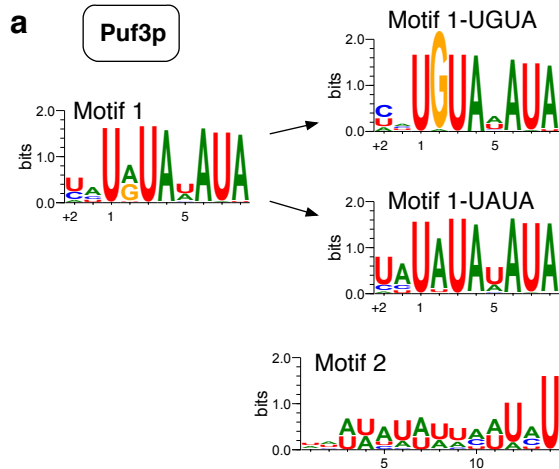


Figure 4.3: (a) Three motifs were found enriched among Puf3p target binding sites. (b) Motif-containing Puf3p sites are preferentially located within 3'UTRs. (c) Motif 1-UGUA- and Motif 2-containing sites display strongest Puf3p binding. (d) Puf3p binds targets annotated with distinct GO terms depending on the motif present. (e) SSM1 is identified by TEISER within annotated Puf3p 3'UTR targets. (f) SSM1 overlaps Motif 1-UAUA-containing Puf3p 3'UTR sites. (g) SSM2 is identified by TEISER within custom annotated Puf3p 3'UTR targets. (h) SSM2 also overlaps Motif 1-UAUA-containing Puf3p 3'UTR sites.

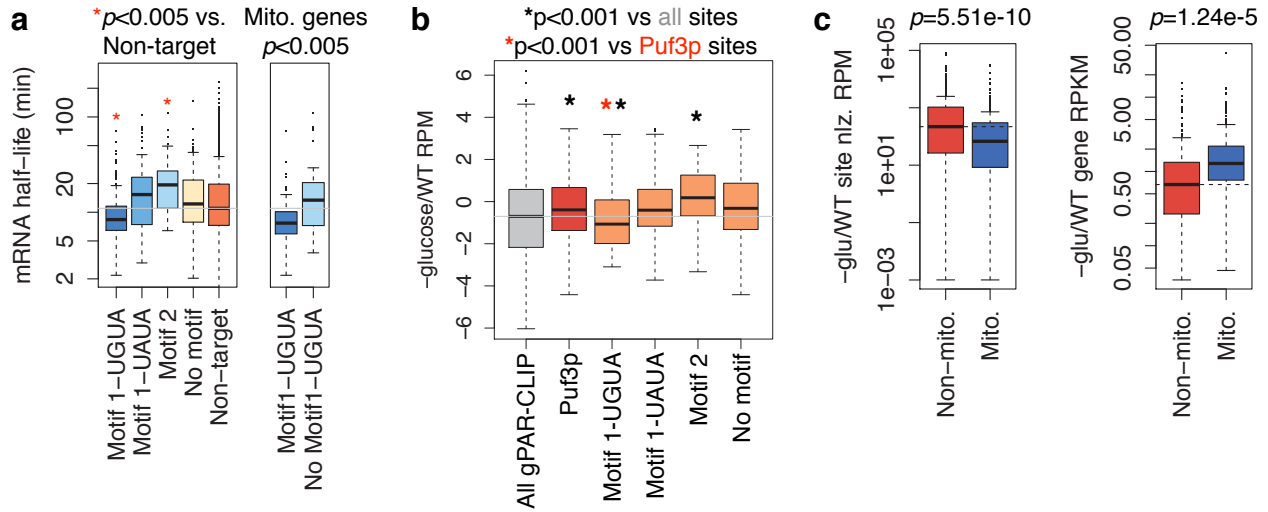


Figure 4.4: (a) Motif 1-UGUA in 3'UTRs is correlated with shorter half-lives for all Puf3p targets (left) and mitochondria-related targets (right). Motif 2 in 3'UTRs is correlated with longer half-lives for all Puf3p targets. (b). Puf3p sites on 3'UTRs containing Motif 1-UGUA are significantly less bound upon glucose starvation than all sites and all other Puf3p sites ($p < 0.001$). (c) Puf3p binding sites on mitochondria-related targets have significantly reduced binding under glucose starvation conditions compared to non-mitochondria-related Puf3p binding sites ($p = 5.5e-10$). Correspondingly, expression of the genes with mitochondria-related Puf3p 3'UTR sites is significantly up-regulated upon glucose starvation ($p = 1.2e-5$).

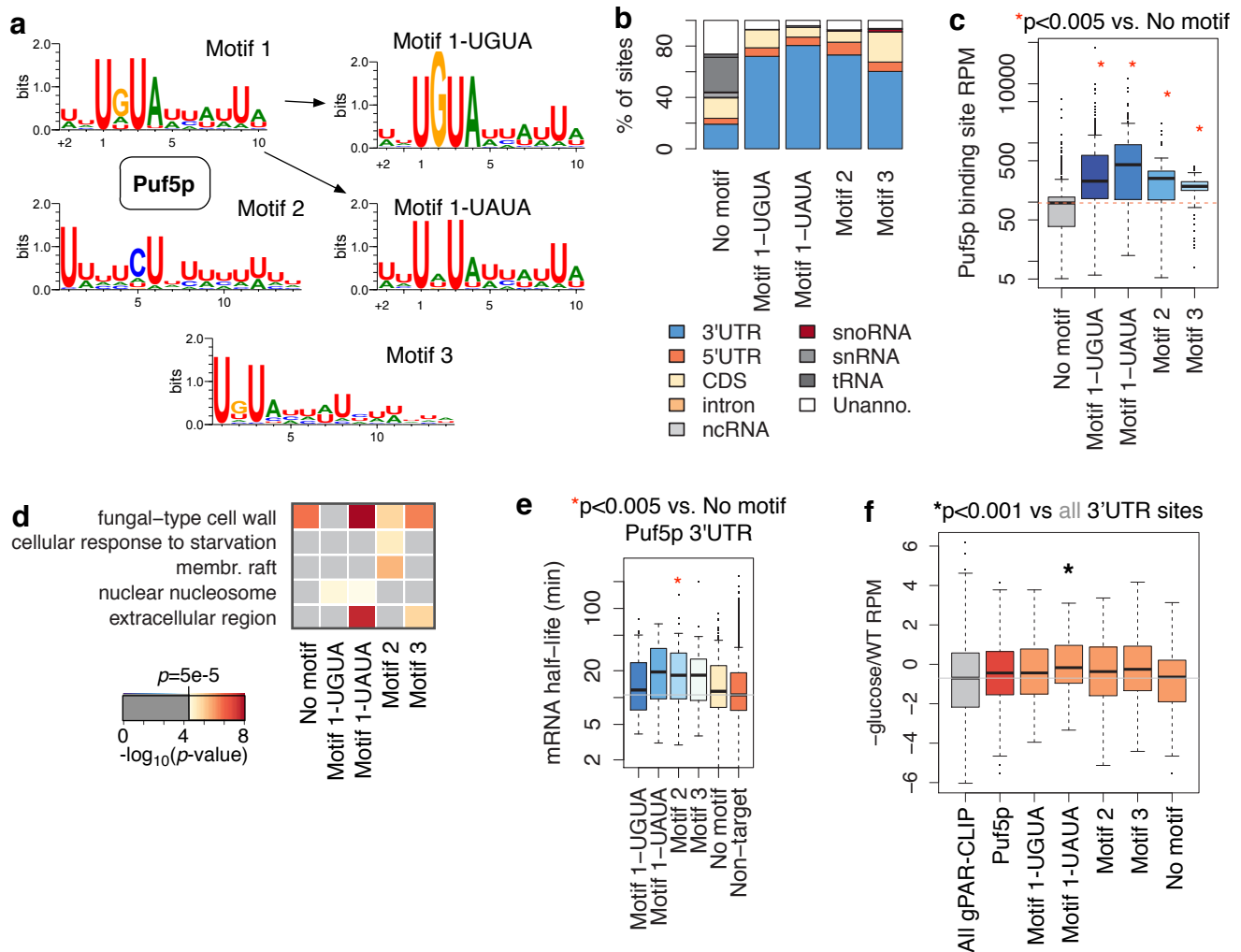


Figure 4.6: (a) Four motifs were found enriched among Puf5p target binding sites. (b) Motif-containing Puf5p sites are preferentially located within 3'UTRs. (c) Motif 1- and Motif 2-containing sites display strongest Puf5p binding. (d) Puf5p binds targets annotated with distinct GO terms depending on the motif present. (e) Genes with Puf5p sites on 3'UTRs containing Motif 2 have significantly longer half-lives than non-Puf5p target genes ($p < 0.005$). (f) Only Puf5p sites on 3'UTRs containing Motif 1-UAUA are significantly more bound upon glucose starvation than all sites ($p < 0.001$).

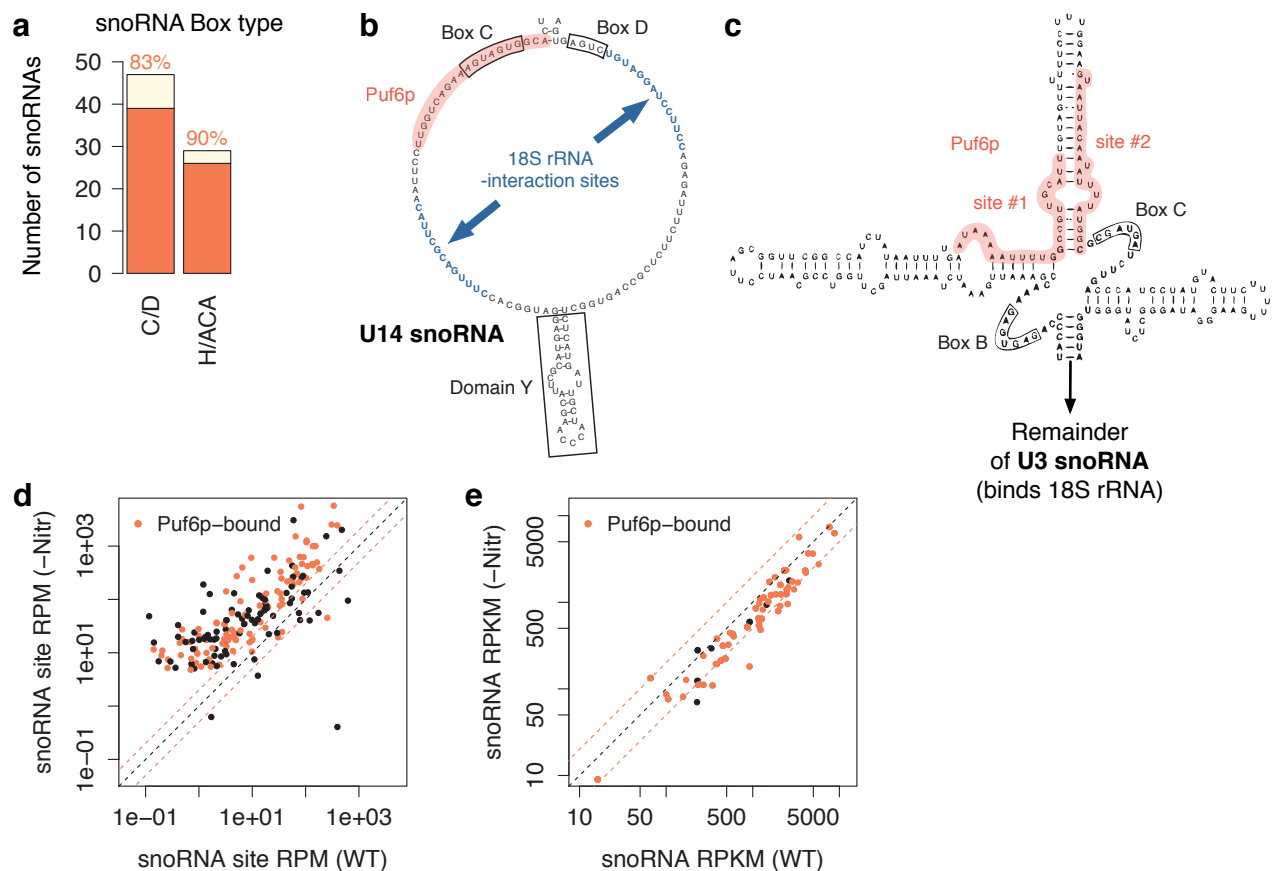


Figure 4.7: (a) Puf6p binds extensively to both C/D and H/ACA box snoRNAs. (b) Puf6p binds U14 snoRNA at a ssRNA region outside rRNA-interactions sites and overlapping Box C element. (c) Puf6p binds U3 snoRNA at dsRNA regions outside rRNA-interactions sites and Box elements. (d-e) In response to nitrogen starvation, snoRNAs expression is down-regulated (d) while Puf6p binding to snoRNAs is increased (e).

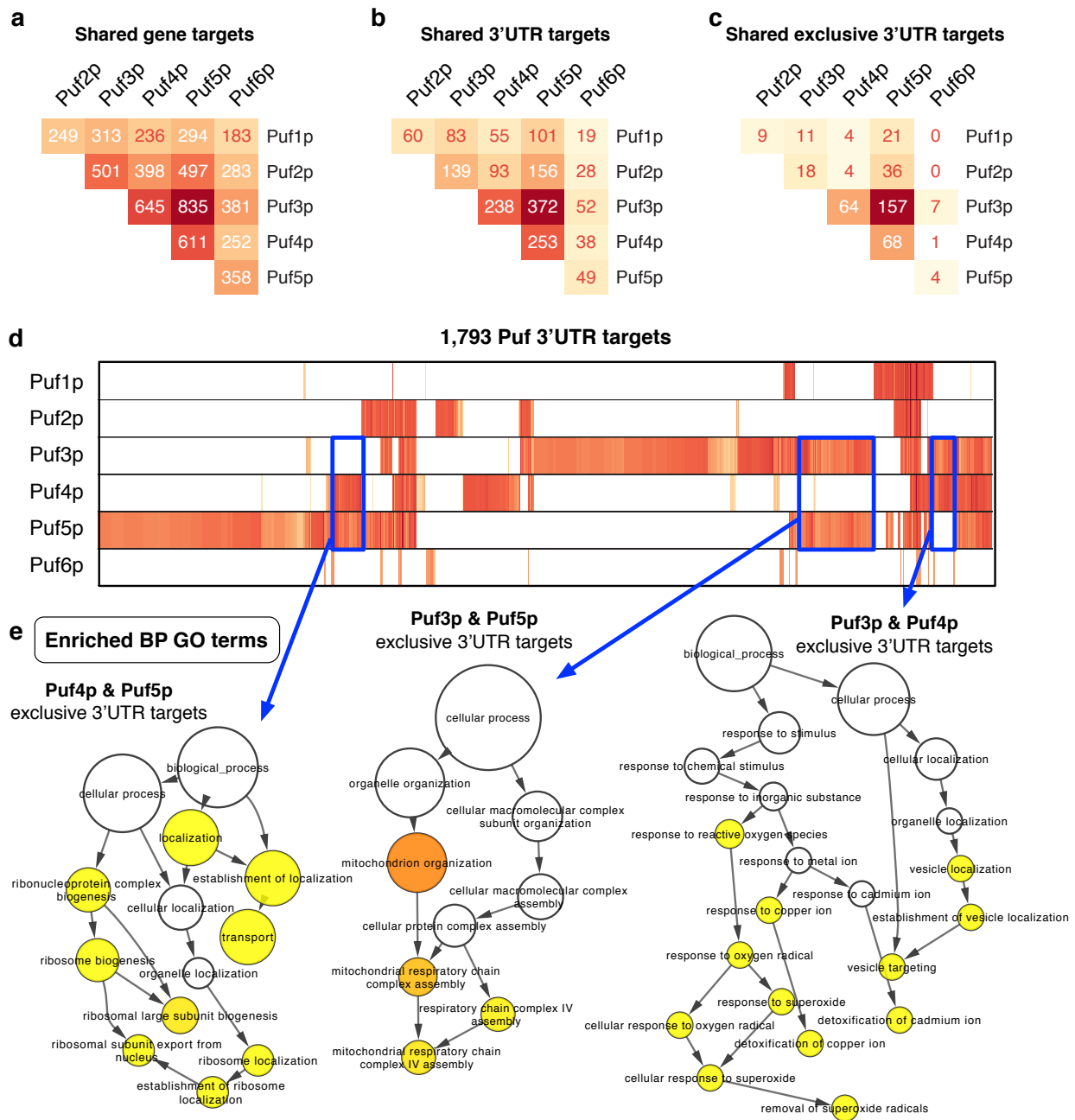


Figure 4.8: (a-b) Hundreds of gene (a) and 3'UTR (b) targets are shared between pairs of Pufs. (c) Puf3p and Puf5p co-bind 157 target 3'UTRs (not bound by any other Puf, exclusive) while Puf3p/Puf4p and Puf4p/Puf5p have over 60 exclusively shared 3'UTR targets. (d) Heatmap showing which Pufs bind the 1,793 co-targeted 3'UTRs. (e) Networks of GO: Biological Process terms enriched among exclusive 3'UTR shared targets of the indicated Puf pairs. Highlighted terms have p -values ≤ 0.05 .

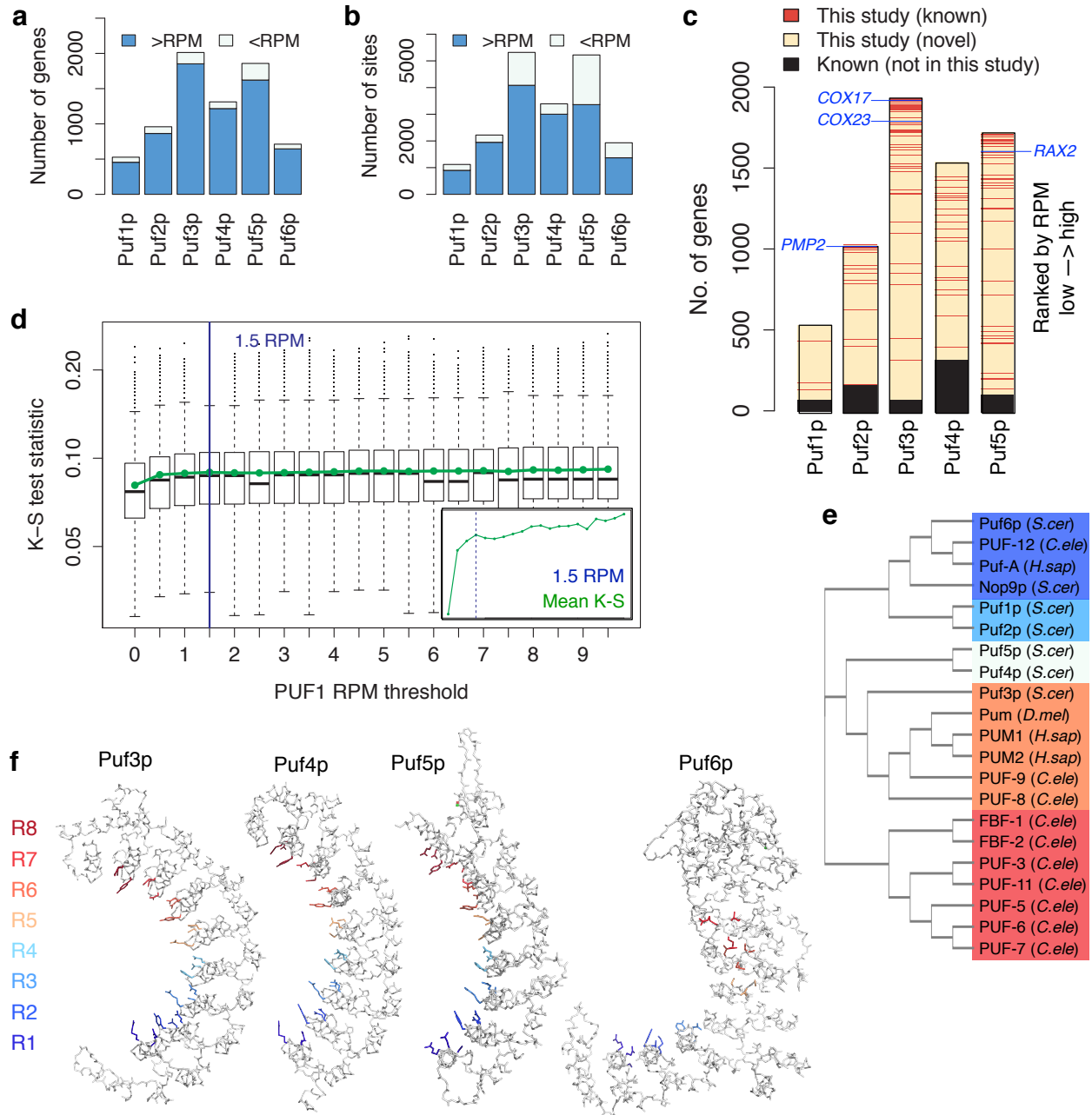
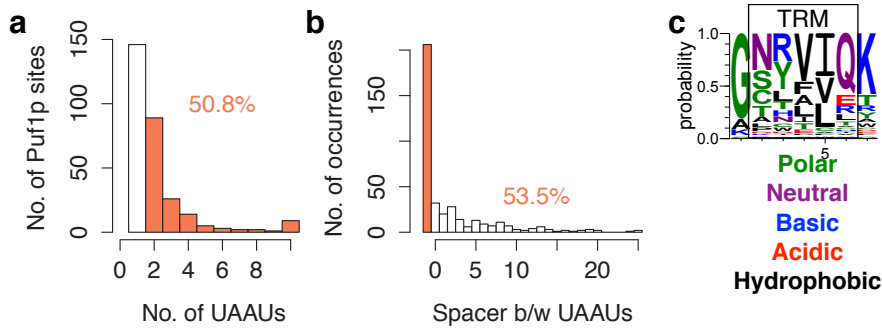


Figure S4.1: (a) The number of genes identified by PAR-CLIP-seq as targets of each Puf protein. The numbers of genes containing at least one site passing an empirical RPM threshold are indicated in blue. (b) The number of direct binding sites identified by PAR-CLIP-seq for each Puf protein. The numbers of sites passing an empirical RPM threshold are indicated in blue. (c) Overlap of PAR-CLIP-seq-identified gene targets of Puf1-5p and targets identified by previous studies. Genes in each column are ranked from high to low (top to bottom) total normalized read counts in our PAR-CLIP-seq. Dark grey regions represent genes not identified in our study. Blue genes indicated known Puf targets. (d) Determination of empirical RPM threshold for Puf1p binding sites. Plotted are distributions of K-S test statistics from 10,000 tests for each threshold value (details in Methods). Inset: zoomed means of K-S test statistics with dotted line indicating the threshold chosen. (e) Phylogenetic tree representing yeast, worm, fly, and human Pumilio proteins assembled using ClustalOmega and Pumilio protein sequences. (f) Crystal (Puf3p, Puf4p) and predicted (Puf5p, Puf6p) structures of Pumilio proteins with TRM amino acid side chains colored from R1 (blue) to R8 (red).



d

		Canonical TRMs								
Puf repeats:	C-	R8	R7	R6	R5	R4	R3	R2	R1	-N
Puf1p TRM:		PH-T	SL-L	AT-L	AR-R	NQ-Y	TQ-W	NQ-T	LR-R	
predicted nt:	5'-	?	N	N	a	U	A	U	?	-3'
Puf2p TRM:		PY-Y	SL-L	TT-L	SR-R	NQ-Y	TQ-W	NQ-T	LR-R	
predicted nt:	5'-	?	N	N	?	U	A	U	?	-3'
Puf3p TRM:		NQ-Y	SE-N	NQ-Y	CQ-R	NQ-H	CQ-R	NQ-Y	SQ-R	
predicted nt:	5'-	U	G	U	A	U	A	U	A	-3'
Puf4p TRM:		NQ-Y	SE-N	NQ-Y	CQ-C	NQ-H	TQ-R	NQ-Y	CQ-R	
predicted nt:	5'-	U	G	U	A	U	U	U	A	-3'
Puf5p TRM:		NQ-Y	SE-N	NQ-Y	CQ-C	NQ-H	TQ-R	NQ-Y	CQ-R	
predicted nt:	5'-	U	G	U	A	U	U	U	A	-3'

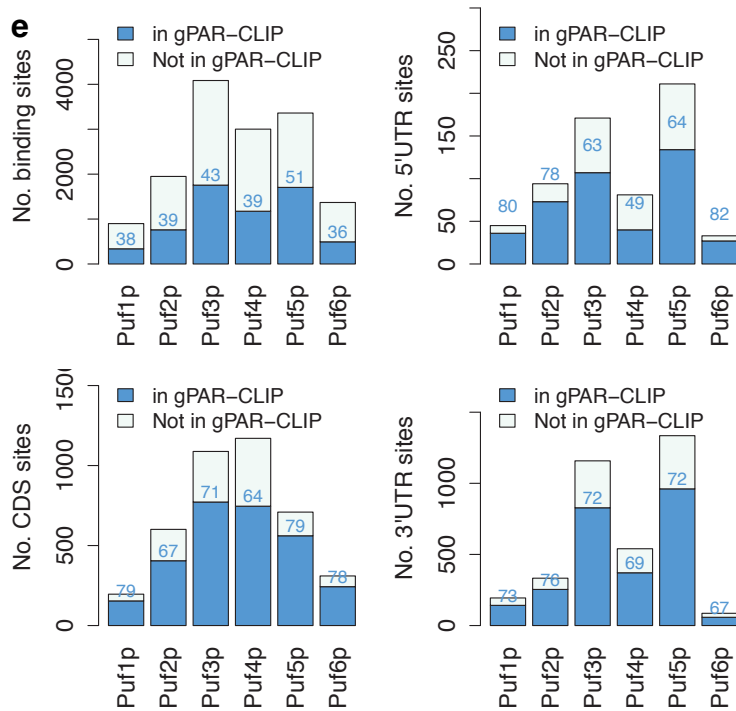


Figure S4.2: (a) Over half of Puf1p sites contain at least 2 repeats of UAAU. (b) Over half of Puf1p sites with 2 or more instances of UAAU contain tandem repeats: UAAUAAU. (c) Puf TRMs are generally bounded by small non-polar residues (N-terminal) and basic residues (C-terminal). Logos plot derived from alignment of Puf repeats from Puf1-6p. (d) Puf1p-Puf5p TRMs and predicted RNAs recognized by the TRMs. (e) Proportion of PAR-CLIP-seq sites identified in global PAR-CLIP-seq datasets.

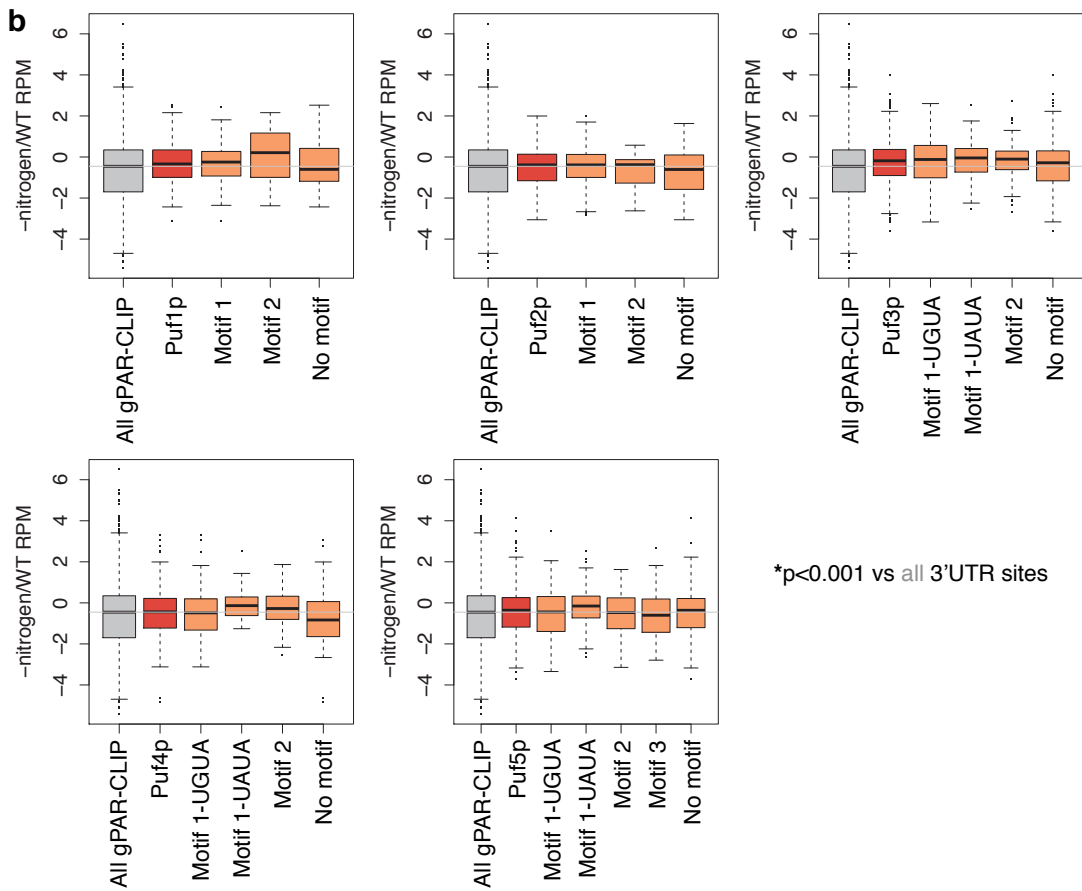
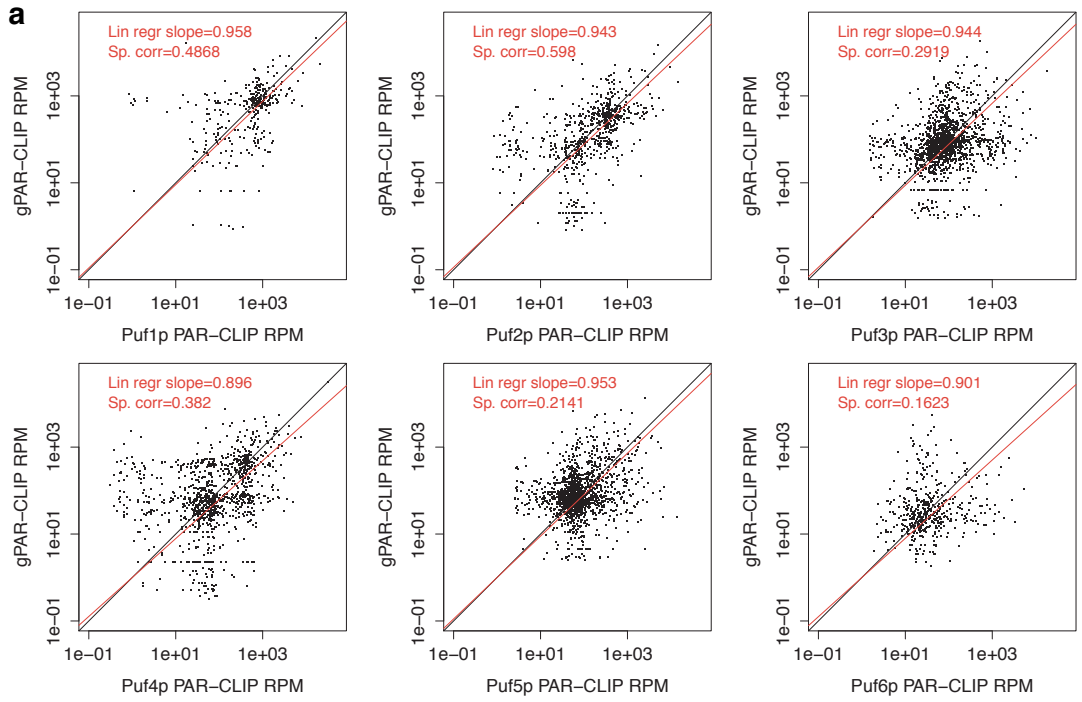


Figure S4.3: (a) Rank-normalized RPM values for PAR-CLIP-seq and gPAR-CLIP-seq binding sites that overlap show strong positive correlation. (b) No significant binding site changes were observed for any Puf with any motif upon nitrogen starvation ($p > 0.001$).

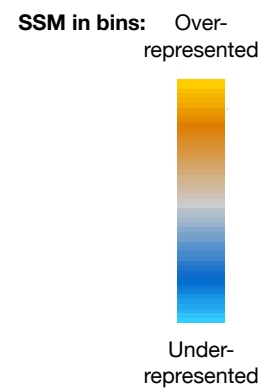
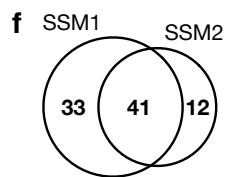
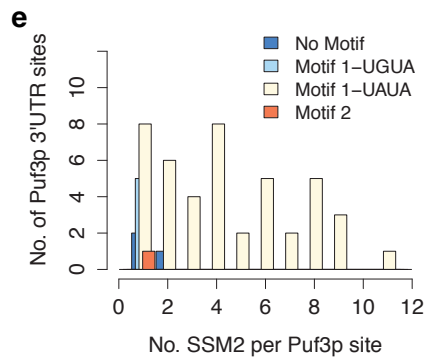
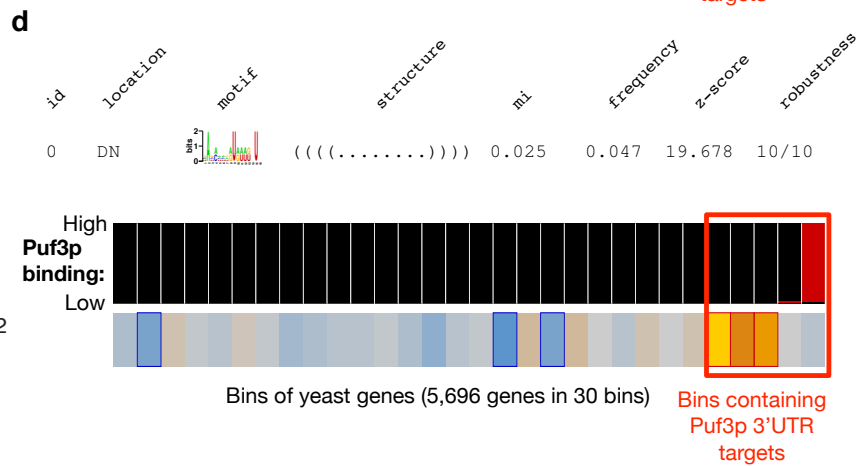
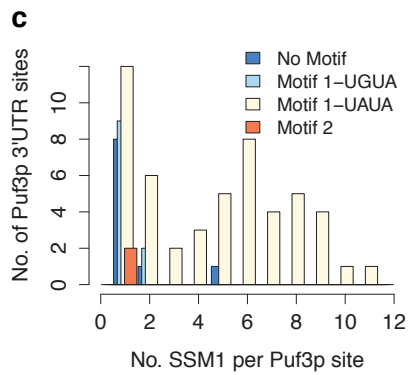
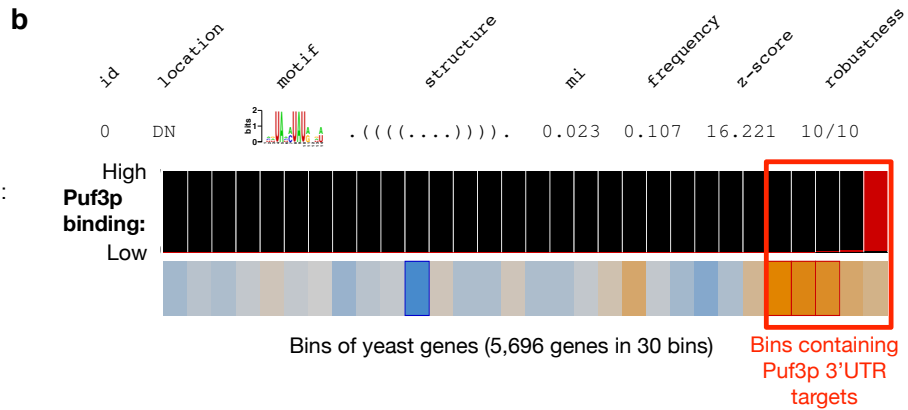
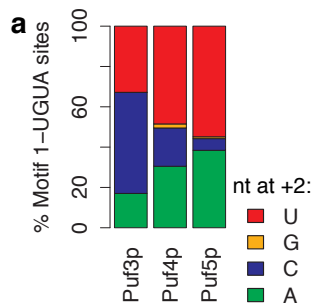


Figure S4.4: (a) Only Puf3p Motif 1-UGUA-containing sites are enriched for C at position +2 from the UGUA. (b) TEISER output showing primary sequence and secondary structure motifs for SSM1. All yeast genes are binned by total Puf3p binding at 3'UTRs. The top 5 bins contain all Puf3p 3'UTR target genes. Remaining genes are randomly assigned to 25 remaining bins. Yellow colored blocks represent bins with over-represented occurrences of SSM1 (red outline = statistical significance), while blue colored blocks represent bins with under-represented occurrences of SSM1 (blue outline = statistical significance). SSM1 was identified by searching set of 300 nucleotides directly downstream of all yeast ORFs. (c) Motif-1-UAUA-containing 3'UTR Puf3p sites harbor multiple instances of SSM1. (d) Same as (b) but SSM2 was identified by searching custom 3'UTR database described in Methods. (e) Motif-1-UAUA-containing 3'UTR Puf3p sites harbor multiple instances of SSM2. (f) Number of Puf3p 3'UTR binding sites containing SSM1, SSM2, or both.

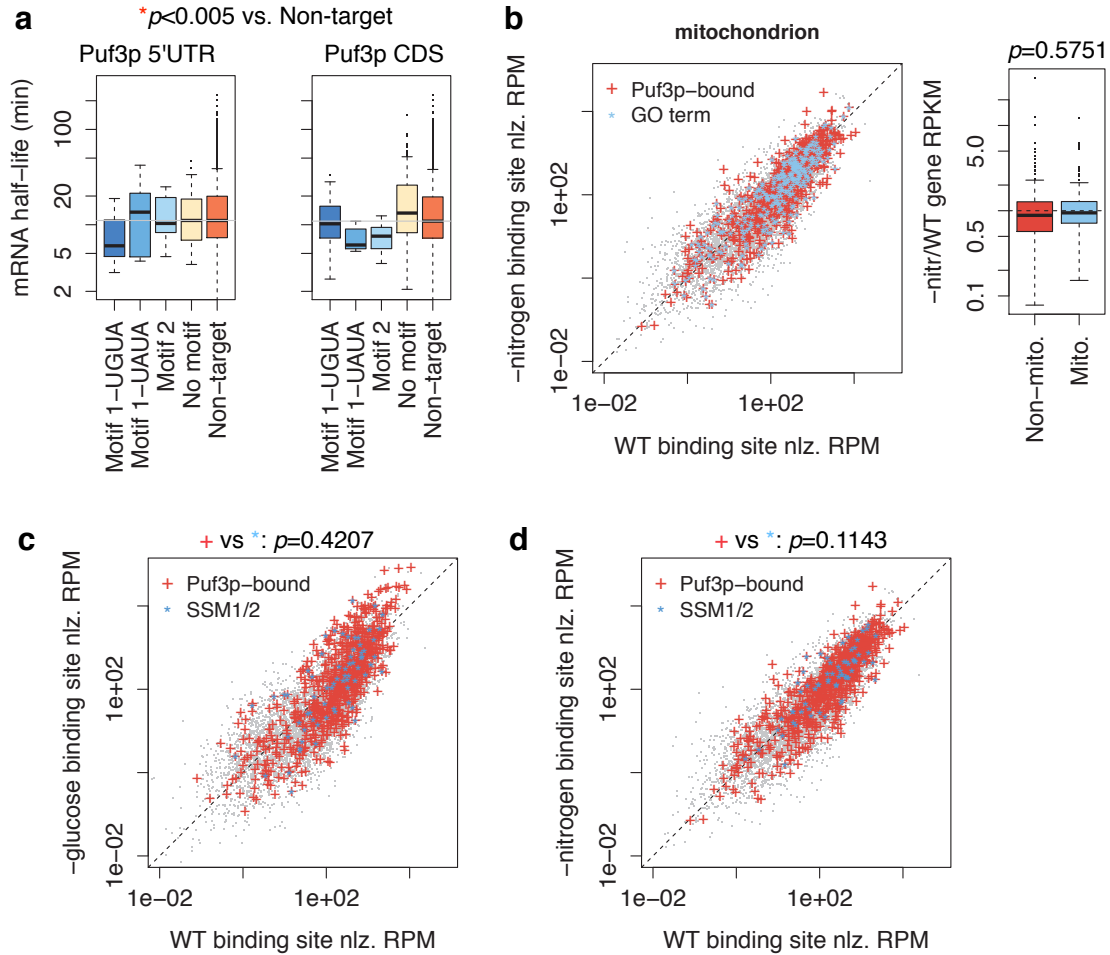


Figure S4.5: (a) Presence of different Puf3p motifs in 5'UTR or CDS binding sites does not significantly correlate with half-life measurements (all $p > 0.005$). (b) No significant difference is observed between binding site coverage or gene expression for Puf3p-bound sites on mitochondria-related genes upon nitrogen starvation ($p > 0.005$). (c-d) Motif 1-UAUA-containing Puf3p sites show no difference in response to glucose (c) or nitrogen (d) starvation whether they overlap SSM1/2 or not ($p = 0.607$).

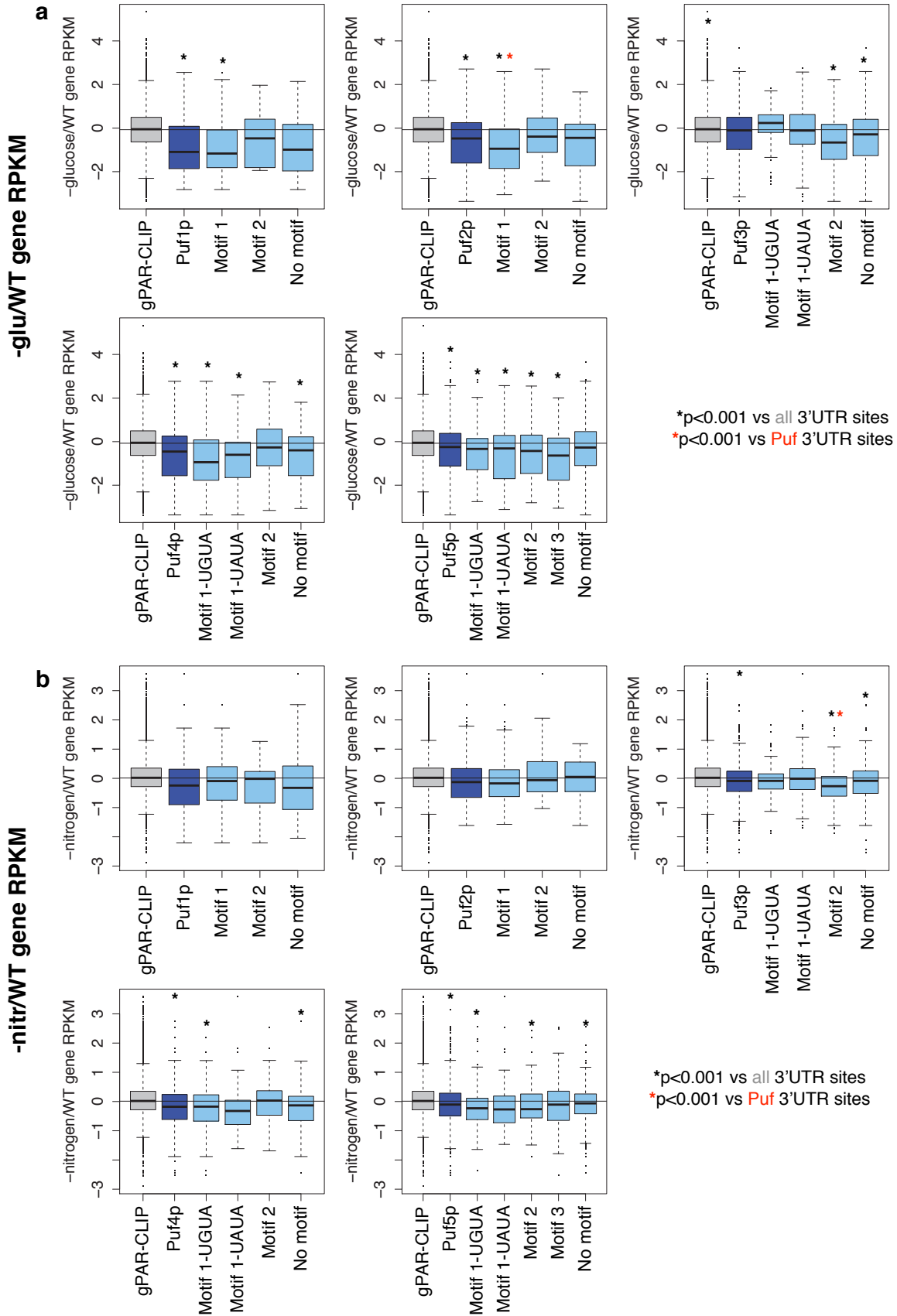


Figure S4.6: (a-b) Gene expression changes for all 3'UTR Puf targets based on motif upon glucose (a) or nitrogen (b) starvation.

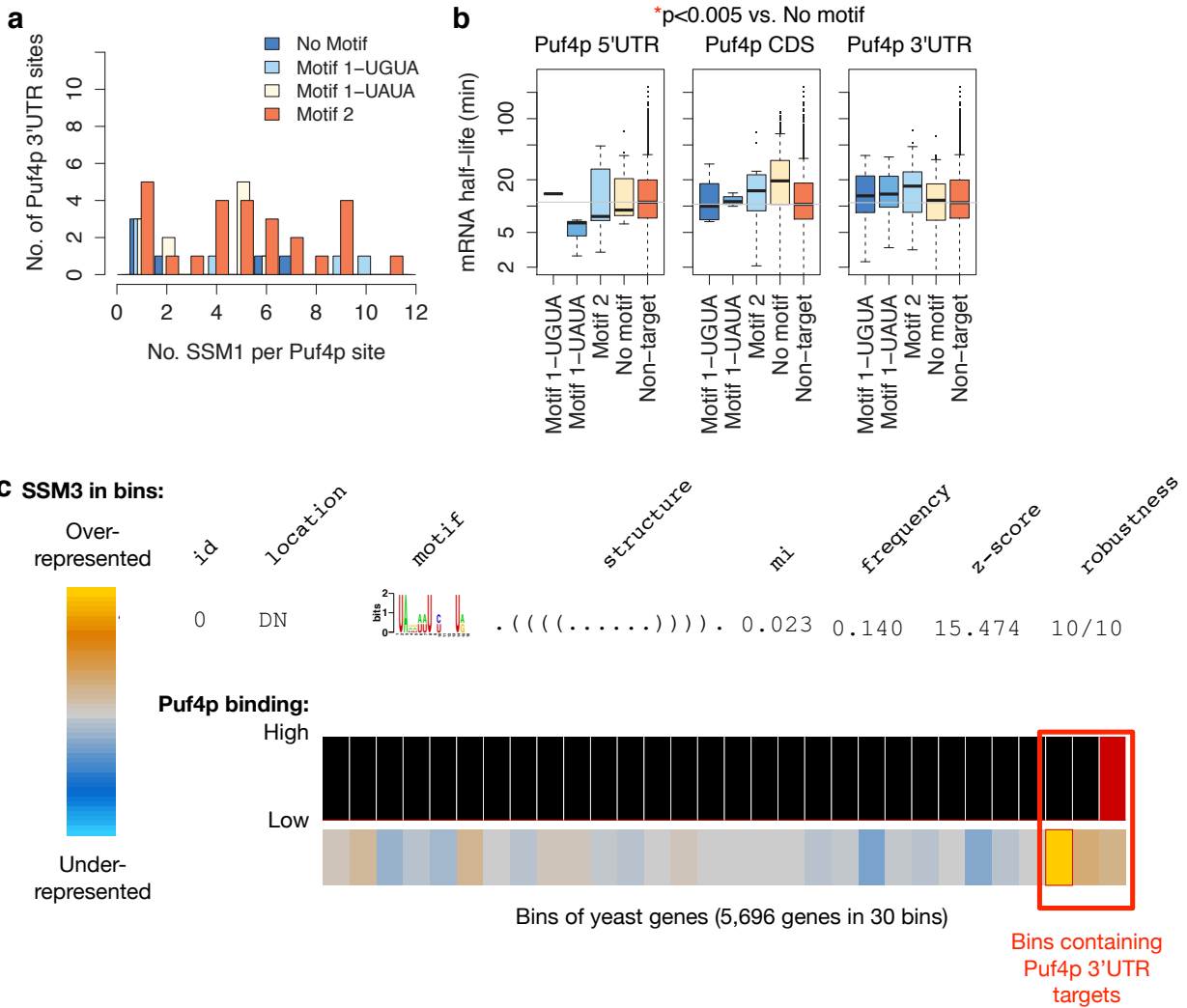


Figure S4.7: (a) Motif-2-containing 3'UTR Puf4p sites harbor multiple instances of SSM3. (b) Presence of different Puf4p motifs in 5'UTR, CDS, or 3'UTR binding sites does not significantly correlate with half-life measurements (all $p > 0.005$). (c) SSM3 was identified from Puf4p target 3'UTRs as in Figure S4d with custom 3'UTR dataset.

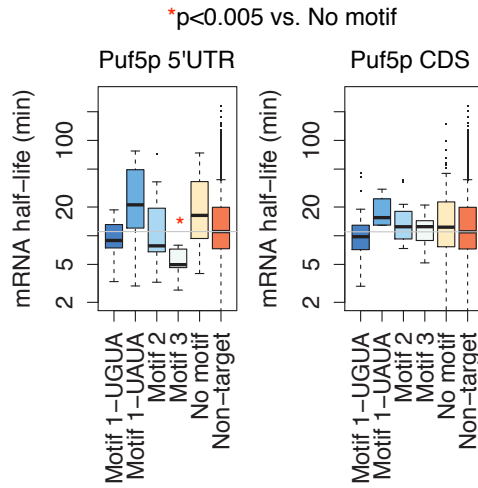


Figure S4.8: Presence of different Puf5p motifs in 5'UTR or CDS binding sites does not significantly correlate with half-life measurements (all $p > 0.005$) except for Motif 3 in 5'UTR sites.

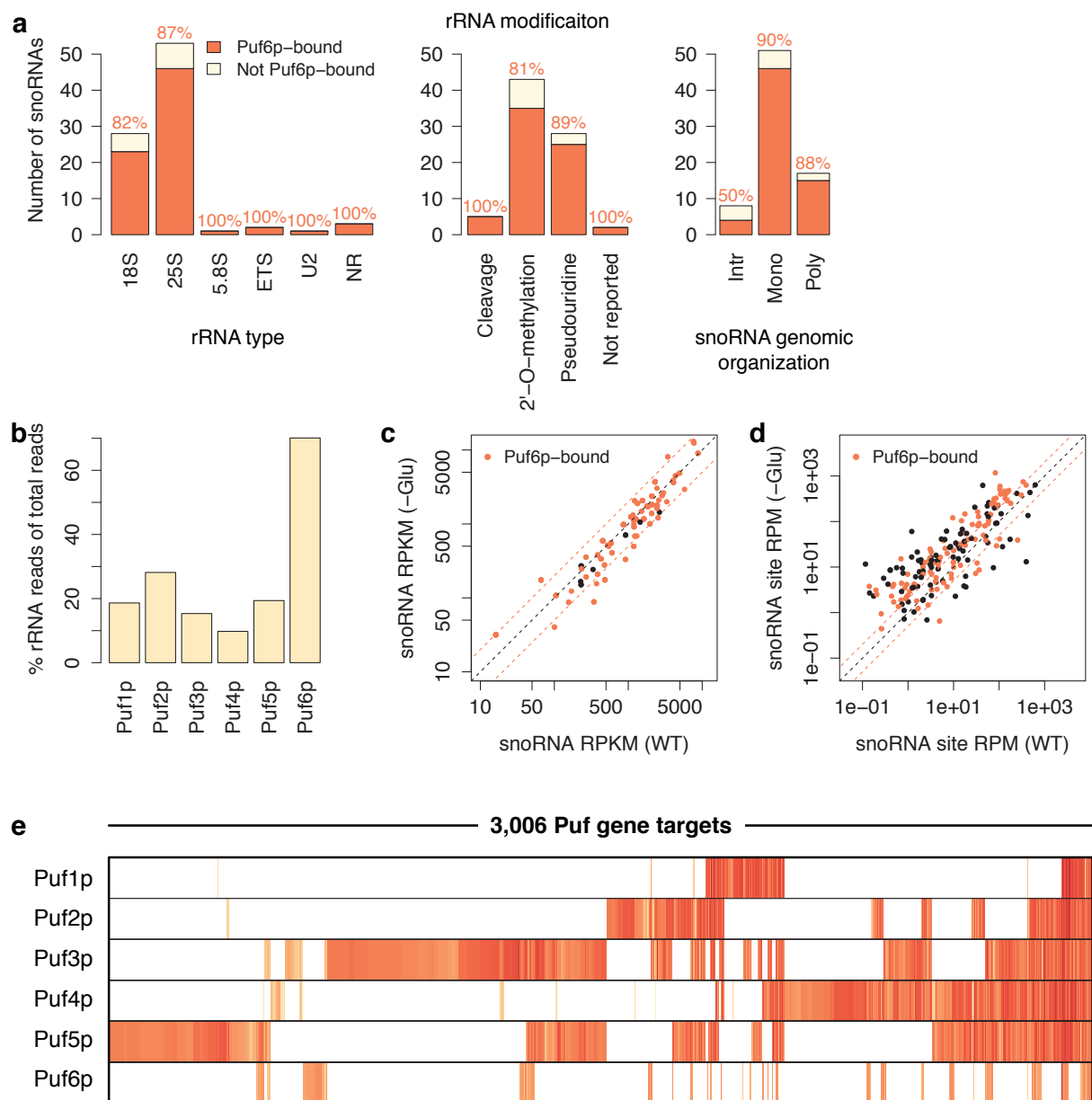


Figure S4.9: (a) Puf6p binding of snoRNAs does not correlate with any snoRNA features. (b) Puf6p binds proportionately more rRNAs than other Pufs. (c-d) Puf6p binding of snoRNAs (d) and snoRNA expression (c) upon glucose starvation responds similarly to nitrogen starvation albeit to a lesser degree. (e) Heatmap shows which Pufs bind the 3,006 co-targeted genes.

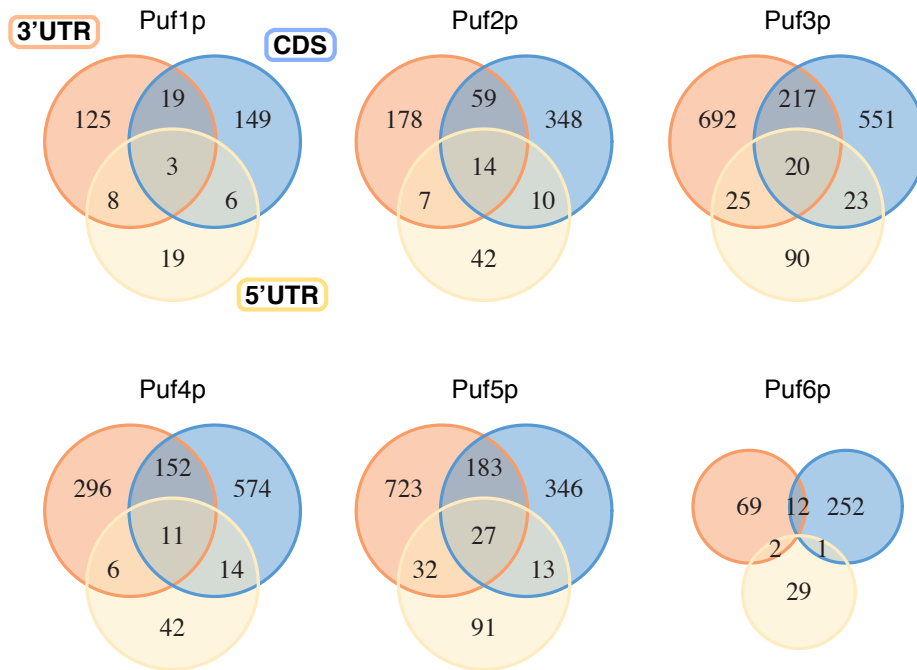


Figure S4.10: Very little Puf binding occurs on different genic regions of the same genes.

CHAPTER V

Specific RBP-directed PTGR investigated by PAR-/CLIP-seq

In this Chapter, I present two studies of specific RBP-mediated PTGR that rely on deep sequencing to refine the mechanism by which the RBPs act on their targets. In the first study, I collaborated with Danny Yang to define the global set of direct mRNA target sites for *C. elegans* Pumilio PUF-9 and the miRNA Argonaute ALG-1 by performing HITS-CLIP. Through analysis of their binding sites, I show that these RBPs co-target similar sets of genes in a manner dependent on not only a primary sequence motif for PUF-9, but also through stable RNA secondary structures. In the second study, I collaborated with Ting Han and Sung Ki Hong in Ken Inoki's laboratory to discover the direct binding sites of the mTOR-interacting RBP LARP1. I show that LARP1 binds pyrimidine-rich regions at 5'UTRs of ribosomal protein-encoding and other translation-related mRNAs to regulate their translation under conditions in which mTOR is inactive.

5.1 PUF-9 regulates developmental timing with miRNAs in *C. elegans*

5.1.1 Introduction

Co-regulation by Pumilio and miRNAs is a conserved feature of PTGR. Mounting evidence suggest that regulation by Pumilio and miRNAs/RISC (miRISC) is pervasive throughout metazoans and is vital for regulation of cellular differentiation and developmental processes. Human PUM1, for example, is required for repression of tumor suppressor *p27* by miR-221/222 (Kedde et al., 2010). In response to growth factor stimulation, PUM1 increases affinity for and binds a Pumilio response element (PRE) in the *p27* 3'UTR which induces a local RNA secondary structural conformation favoring interaction with and suppression by miR-221/222. Human PUM1 and PUM2 also have been shown to bind to the 3'UTR of the E2F3 oncogene, enhancing the activity of multiple miRNAs to silence E2F3 (Miles et al., 2012). Dysregulation of either *p27* or E2F3 leads to proliferation and cell cycle progression defects that are hallmarks of cancer pathogenesis; thus, PTGR of these genes through Pumilio- and miRNA-mediated mechanisms is important.

To computationally explore the physical characteristics of Pumilio and miRNA co-binding of substrate mRNAs, Incarnato et al. (2013) developed the MREdictor algorithm that predicts miRNA target sites on regions of structurally inaccessible 3'UTRs in the context of proximal PREs. While this work was inspired by the *p27* and E2F3 examples of Puf/miRNA co-regulation, it revealed a pervasive trend of miRNA target sites residing in poorly accessible regions on target 3'UTRs proximal to potential Pumilio, and perhaps other RBP, interaction sites.

Interactions between the Pumilio and miRISC post-transcriptional regulatory pathways is vital for regulating the larval-to-adult transition in the nematode *Caenorhabditis elegans*. Nolde et al. (2007) observed that loss of *puf-9* enhanced *let-7(-)* miRNA hypomorphic phe-

notypes, including vulval bursting. They also showed that PUF-9 binds *hbl-1* mRNA, a known target of *let-7*, at sites in the 3'UTR containing PREs as well as *let-7* target sites. Given these results, we wondered whether any of the other 12 Pumilio proteins encoded by the *C. elegans* genome interacted either physically or genetically with the miRNA pathway. We also wondered if Pumilio/*let-7* co-regulation was specific to these two targets (*lin-14* and *hbl-1*), to all *let-7* targets, other miRNA families, or to all miRNAs in general.

***puf-9* genetically interacts with *let-7*.** We screened the 13 conserved Pumilio of *C. elegans* for enhancement of *let-7(-)* vulval bursting and identified only *puf-9* as a genetic enhancer of this phenotype. While *C. elegans* exhibited no vulval bursting in wild-type animals (0%) and only weak bursting in *let-7(-)* (2%) and *puf-9(-)* (6%) animals, *puf-9(-);let-7(-)* double mutant animals show enhanced bursting (22%). Bursting in both *puf-9(-)* and *puf-9(-);let-7(-)* animals is rescued to wild-type levels (0% and 1%, respectively) by introduction of a PUF-9::GFP transgene. Similarly, bursting is suppressed in *puf-9(-);lin-41(-)* animals (1%) in which the *let-7* target *lin-41* is disrupted. Taken together, our data suggest a direct link between regulation of developmental timing by *let-7* and *puf-9*. We hypothesize that PUF-9 physically associates with additional *let-7* target mRNAs, including *lin-41*, to regulate developmental timing in *C. elegans*.

5.1.2 PUF-9 and miRISC directly interact at sequence-specific and highly structured regions of co-targeted genes

PUF-9 and miRISC co-target genes at directly overlapping sites. Given the specific interactions observed between *puf-9* and two *let-7* targets, we wondered whether globally PUF-9 directly interacted with other *let-7* targets and perhaps, more broadly, with targets of other miRNAs. To determine the specific sites of PUF-9 interactions, we isolated animals at embryo, L1, and gravid adult stages and performed CLIP. Isolated RNAs were

barcoded and subjected to deep sequencing using the Illumina HiSeq platform. In parallel we isolated and sequenced RNAs bound by the miRISC Argonaute ALG-1 in the same stages. Reads were mapped using Bowtie2 (Langmead and Salzberg, 2012) to the *C. elegans* genome version WS220. PUF-9 and ALG-1 binding sites were identified as clusters of overlapping reads; binding site coverage was quantified as the number of reads in the clusters normalized to the number of millions of mapped reads (RPM) in each sequencing library and to the RPKM value of the gene to which the cluster maps. We set an empirical minimum RPM threshold to 0.5 RPM to filter off low-confidence sites. In our datasets, PUF-9 binds 2.5k-5.5k unique genes, while ALG-1 binds 200-1k unique genes.

Similar to previous ALG-1 binding studies, we observed a majority of ALG-1 bound to 3'UTRs and exonic regions as well as mature miRNAs (Fig. 5.1A). PUF-9 also bound 3'UTRs in agreement with previous characterizations of Puf proteins as 3'UTR-interacting regulators of mRNAs stability (Nolde et al., 2007; Quenault et al., 2011). PUF-9 targets are enriched for DAVID functional categories related to growth and development, aging, the cytoskeleton, reproduction, and germline development (Fig. S5.1A). We searched 3'UTR PUF-9 binding sites (PBSs) for over-represented sequence motifs using MEME (Bailey and Elkan, 1994) and identified a 5'-UGUA-3' sequence followed by an A/U-rich region with consensus 5'-HWWW-3' (Fig. 5.1B). Together, these sequences form a full *Pumilio* recognition element (PRE) similar to those identified for other *Pumilio*s. Of the top 20% PBSs (based on RPM coverage) in each stage, 50-60% contain a PRE and another 20-30% contain UGUA (Fig. 5.1C) suggesting that presence of a PRE/UGUA promotes strong PUF-9:target interactions. Indeed, read coverage at 3'UTR PBSs with a full PRE or just UGUA was significantly higher than at PBSs with no motif (Welch's two-tailed *t*-tests: $p < 10^{-17}$ (PRE) and $p < 10^{-3}$ (UGUA) in any stage; Fig. 5.1D). Using phastCons conservation scores derived from aligning the *C. elegans* genome to six other nematode genomes, we observed that 3'UTR PBSs with a full

PRE were significantly more conserved than sites with no motif (Fig. S5.1B), suggesting a functionally conserved role for PUF-9 binding targets at sites with a full PRE motif.

To determine whether PUF-9 and ALG-1 preferentially bind common target mRNAs, we identified genes at each developmental stage with at least one PBS or at least one ALG-1 binding site (ABS) in the 3'UTR and determined the number of genes at each stage with both PUF-9 and ALG-1 binding. 70-80% of mRNAs bound by ALG-1 were also bound by PUF-9 (Fig. 5.1E) suggesting extensive co-regulation by PUF-9 and miRISC. To determine if the number of common targets of PUF-9 and ALG-1 is more than expected by chance, we randomly selected the same number of genes in each gene set from the list of all genes expressed at each stage (Hillier et al., 2009) with the probability of selecting each gene corresponding to its expression level in the dataset. From the set of random PUF-9 and ALG-1 “targets” we calculated the number of shared genes. This process was repeated 100,000 times, and an empirical p -value was calculated as (No. of times observed more shared genes in random control)/100,000. Indeed, PUF-9 and ALG-1 co-target more genes at 3'UTRs than expected by chance ($p < 1e-5$ for all stages). Surprisingly, 60% of PUF-9/ALG-1 co-targeted genes in gravid adults contained a PBS with a full PRE compared to only 40% of genes targeted only by PUF-9 (Fig. S5.1D), suggesting that PREs are especially important for recruiting PUF-9 to ALG-1 target genes in gravid adult animals.

Based on the close proximity of known PUF-9 and *let-7* target sites on *hbl-1* and *lin-41* 3'UTRs, we wondered whether globally PUF-9 and ALG-1 bind in close proximity on common target 3'UTRs. We defined the proximity of PBSs and ABSs as the distance in nucleotides between the midpoints of two binding sites. Surprisingly, we observed 73%, 51%, and 78% of 3'UTR ALG-1 binding located within 25nt of a PBS in embryo, L1, and gravid adult animals, respectively (Fig. S5.1C), suggesting widespread co-binding between PUF-9 and miRISC. Furthermore, when we centered all PBSs on their midpoints and summed

ALG-1 binding RPM surrounding each PBS, we observed strong ALG-1 binding directly overlapping PBSs only at 3'UTRs (Fig. 5.1F). This trend was observed regardless of whether we examined all ABSs on the same 3'UTR as the PBS or limited the analysis to just the closest ABS to each PBS (data not shown). Finally, we observed an enrichment of PREs in PBSs located within 25nt of an ABS in gravid adult animals (40% of PBSs \leq 25nt versus 21% of PBSs $>$ 25nt), but not in embryos or L1 animals, supporting the idea that the PRE may be especially important for recruiting PUF-9 to sites near miRISC binding specifically in gravid adult animals. Taken together, our data support a PUF-9/miRISC co-regulatory network in which PUF-9 and ALG-1 bind a common set of target mRNAs for post-transcriptional regulation.

PUF-9 target sites are highly structured. Previous work in BJ primary fibroblast cells showed that the human Pumilio PUM1 binds an energetically stable stem-loop structure in the 3'UTR of *p27* to induce a structural change allowing miR-221/222-loaded miRISC access to its recognition site (Kedde et al., 2010). Even though we identified a primary sequence motif enriched among PBSs, we wondered whether PUF-9 in *C. elegans* could also bind highly structured RNA. Using the secondary structure prediction software RNAfold (Lorenz et al., 2011), we predicted the minimum free energy (MFE, $-\Delta G$ in kcal/mol) of all 3'UTR PBSs extended by 15nt in either direction. As a control, we also predicted the MFE of RNA regions located directly upstream and downstream of each PBS (control region length matched the length of its associated PBS). Strikingly, PBS MFEs were significantly lower than their matched upstream and downstream control regions (Fig. 5.2A, Fig. S5.2A), suggesting that PUF-9 preferentially binds structured regions in 3'UTRs. Furthermore, PBSs with a PRE are more structurally stable than PBSs without a PRE, suggesting that both primary sequence and secondary RNA structure contribute to PUF-9 binding at target 3'UTRs. In both of our predictions of PUF-9:*lin-41* target sites at PBS1, half of the

PUF-9 sequence motif nucleotides are paired while the remaining nucleotides are unpaired, leaving them available for base-specific recognition by PUF-9 Puf repeat domains. This observation suggests that PUF-9 might initially recognize unpaired nucleotides in the PBS base-specifically followed by melting of the paired RNA region, much like the PUM1:*p27* interaction, to allow full base-specific recognition of the PBS.

Based on our observation that PBSs overlapping ABSs in gravid adult animals contain a higher proportion of PREs than PBSs away from ABSs, we next compared the MFE of PBSs located ≤ 25 nt from an ABS with PBSs located > 25 nt from an ABS. PBSs that are closer to an ABS are significantly more structured compared to PBSs that are away from an ABS (Welch's two-tailed *t*-test $p=6.6e-4$; Fig. 5.2B). No difference in RNA stability was observed in embryos or L1 animals ($p=0.14$ and $p=0.17$, respectively; Fig. S5.2B), suggesting that this phenomenon is specific to regulation of targets in gravid adult animals. Predicted structures with the lowest MFE show how PREs within PBS1 and PBS2 on the *lin-41* 3'UTR might be part of hairpin structures (Fig. 5.2C) and how the *let-7* target site (LCS1) is capable of forming a hairpin with the PRE in PBS1 (Fig. 5.2D).

We immunopurified PUF-9 and successfully validated a subset of HITS-CLIP-identified PUF-9 targets by qPCR. Additional molecular genetic and biochemical experiments are required to fully characterize the importance of proximal PUF-9- and ALG-1-mediated PTGR of developmental timing in gravid adult *C. elegans*.

5.2 LARP1 displays dual mode of RNA binding to regulate translation of ribosomal protein-encoding genes

5.2.1 Introduction

Mechanistic target of rapamycin complex 1 (mTORC1) is a strong promoter of translation and translation initiation and plays a major role in driving cellular proliferation (Shimobayashi and Hall, 2014). When cells are subjected to metabolic stress, mTOR, an essential component of mTORC1, is sequestered in stress granules that house translational components including mRNAs, initiation factors, and poly(A)-binding proteins (PABPs) (Kedersha et al., 2013; Wippich et al., 2013; Thedieck et al., 2013). Allosteric mTORC1 inhibitors (*e.g.* rapamycin) can partially inhibit global protein synthesis but almost completely block translation of mRNAs containing pyrimidine-enriched sequences (PES) such as 5' terminal oligopyrimidine (5'TOP), 5'TOP-like, and pyrimidine-rich translation elements (PRTEs) (Huo et al., 2011). Inhibiting mTORC1 activity blocks its phosphorylation and inhibition of eIF4E-binding proteins (4EBPs), which can then inhibit the eIF4F translation initiation complex. The eIF4F complex is required to stimulate translation of PES-containing mRNAs, which might explain why mTOR inhibitors preferentially block translation of these transcripts. Despite work illuminating the components of this post-transcriptional regulatory pathway, the precise mechanism by which the eIF4F complex preferentially stimulates translation of PES-containing mRNAs remains unknown.

La-related proteins (LARPs) are a conserved family of RBPs that contain a La RBD. Two members of the LARP family, LARP1 and LARP4B, promote mRNA stability and translation through interactions with PABP1 (Aoki et al., 2013; Bousquet-Antonelli and

Official citation:

Hong SK, **Freeberg MA**, Kamath A, Han T, Yao Y, Fukuda T, Suzuki T, Kim JK, and Inoki K. LARP1 functions as a molecular switch for mTORC1-mediated translation of an essential class of mRNAs. (in revision).

Deragon, 2009; Burrows et al., 2010; Schäffler et al., 2010; Tcherkezian et al., 2014). Another LARP protein, LARP1, associates with components of the active eIF4F complex and 5'TOP mRNAs to regulate stability and translation of these mRNAs (Aoki et al., 2013; Burrows et al., 2010; Tcherkezian et al., 2014). Recent studies have proposed that LARP1 associates with PABP1 and translation initiation factors to promote translation of 5'TOP mRNAs that are sensitive to cellular mTORC1 activity (Tcherkezian et al., 2014); however, the molecular mechanisms by which LARP1 regulates the translation of such specific mRNAs remain unknown.

In this study, we show that in a HEK293 cell line, LARP1 physically associates with mTORC1, and that mTORC1 activity is required for interaction with LARP1. We also show that LARP1 indirectly associates with PABP1 and initiation factors through binding of common mRNAs. In fact, LARP1 anchors mTORC1 to LARP1-interacting mRNAs in a manner dependent on mTOR activity, which explains mTORC1-dependent translation initiation of LARP1-interacting mRNAs. We confirm that LARP1 scaffolds mTORC1 to LARP1-interacting mRNAs (*e.g.* 5'TOP-containing mRNAs) to allow mTORC1 to locally phosphorylate 4EBPs and S6K1 to promote the translation of LARP1-bound mRNAs.

5.2.2 LARP1 binds 5'UTR pyrimidine-rich regions of translation-related transcripts in a mTOR activity-dependent manner

While several recent studies have demonstrated that LARP1 associates with 5'TOP mRNAs (Tcherkezian et al., 2014; Burrows et al., 2010; Aoki et al., 2013), the comprehensive identity and sequence characteristics of mRNAs that preferentially interact with LARP1 have not been defined. To address this gap, we performed PAR-CLIP of endogenous LARP1 in HEK293T cells in the presence or absence of an mTOR inhibitor (PP242), followed by deep sequencing of the LARP1-bound RNA substrates. Sequenced reads were mapped using

Bowtie (Langmead et al., 2009) to the human transcriptome, clustered to derive LARP1 binding sites, and filtered to retain clusters containing 0-2 T-to-C conversion events and passing an empirically-derived reads per million mapped reads (RPM) threshold (details in Methods; Fig. S5.3A). In parallel, replicate mRNA-seq experiments were performed in the presence or absence of PP242 to quantify gene expression and normalize LARP1 binding sites to mRNA abundance levels. We identified 1,200 and 1,900 LARP1 binding sites on 1,000 and 1,500 mRNAs in the presence or absence of PP242, respectively (Fig. S5.3A,B).

Analysis of genes bound by LARP1 in growing conditions revealed enrichment for Gene Ontology terms related to translation. LARP1 bound 137 translation-related genes including 42 genes encoding ribosomal proteins (RP), as well as genes involved in cellular differentiation and development (Fig. 5.3A). Under mTOR-inactive conditions, more translation-related genes were bound by LARP1 (230) including 94 RP-encoding genes (Fig. 5.3A). These results indicate that LARP1 substrates are enriched for mRNAs encoding factors involved in translation, and that this interaction is enhanced under conditions of mTOR inactivation.

To identify where LARP1 binds across a transcript, we summed LARP1 binding coverage across the 5'UTR, CDS, and 3'UTR regions of its targets, which were separated into non-translation-related (non-TR) genes, translation-related (TR) genes, and the subset of TR genes encoding RPs (Fig. 5.3B). Strikingly, LARP1 binding at 5'UTRs of TR genes more than doubled upon mTOR-inactivation, and binding on RP genes increased from 0% to 17% (Fig. 5.3B). To further explore this observation, we plotted the accumulation of LARP1 binding under growing and mTOR-inactive conditions along normalized gene lengths (Fig. 5.3C). Across non-TR mRNAs, LARP1 preferentially associated with CDSs and 3'UTRs, but was almost completely absent from 5'UTRs. In contrast, LARP1 bound most strongly to 3'UTRs of TR and RP mRNAs under growing conditions. Importantly, under conditions of mTOR inactivation, LARP1 accumulated at 5'UTRs, with the majority of 5'UTR binding

occurring on RP transcripts (Fig. 5.3C).

Since LARP1 regulates PES-containing mRNAs, including 5'TOP sequences (Tcherkezian et al., 2014), we searched the 58 and 92 5'UTR LARP1 binding sites under growing and mTOR-inactive conditions, respectively, for a consensus motif using MEME (Bailey and Elkan, 1994). We identified 6 consecutive pyrimidines in all 5'UTR LARP1 binding sites, suggesting that LARP1 binds directly to PESs (Fig. 5.3D). Importantly, 5'UTR LARP1 binding sites rarely overlapped with 5'TOP sequences, which are located at the 5'-most end of 5'UTRs; instead, LARP1 binds predominantly at the 3'-most end of 5'UTRs (Fig. 5.3C). In fact, 5'TOP-containing 5'UTRs bound by LARP1 are more pyrimidine-rich at their 3' ends than those not bound by LARP1 (Fig. 5.3E). To confirm that LARP1 binds PESs within target 5'UTRs, we compared pyrimidine-richness of LARP1-bound regions to non-LARP1-bound regions of these 5'UTRs and observed a significantly higher proportion of pyrimidines in LARP1-bound regions under both growing (Welch's two-tailed t -test: $p=1.4e-15$) and mTOR-inactive ($p=1.2e-18$) conditions (Fig. 5.3F). Taken together, our data suggest that LARP1 specifically recognizes and binds PESs at the 3'-end of 5'UTRs for a subset of TR and RP transcripts *in vivo*.

LARP1 also binds CDSs of non-TR genes and 3'UTRs of TR and RP mRNAs. We identified GA-rich motifs in 9-15% of these sites under both conditions (Fig. S5.3C). LARP1-bound 3'UTR regions are slightly, but significantly, enriched for higher G-content than non-LARP1-bound regions on the same 3'UTRs (Fig. S5.3D). These motifs are similar to ones identified for RRM domain-containing RBPs in a recent systematic *in vitro* study characterizing the sequence-specific recognition sites for RBPs across 24 eukaryotes (Ray et al., 2013), suggesting a dual RNA-binding mode, which has now been confirmed by 5'UTR and 3'UTR reporter binding assays (data not shown). The LARP1 La domain may be responsible for recognizing and binding pyrimidine-rich regions at the 3'end of target 5'UTRs while its putative RRM

domain (Bayfield et al., 2010) recognizes and interacts with 3'UTRs, perhaps at GA-rich regions.

The relationship between LARP1 binding and decreases in translational efficiency (TE) upon mTOR inactivation are paralleled in mouse embryonic fibroblasts. We obtained measurements of changes in mouse transcript TE upon treatment of cells with Torin1 (Thoren et al., 2012). Thirty-three percent of human homologs of mouse genes exhibiting decreased TE were bound by LARP1 in mTOR-inactive conditions compared to only 12% and 14% of genes showing no change in or increased TE, respectively (Fig. S5.3E). We next asked if increased pyrimidine richness observed at 3' ends of LARP1-bound 5'UTRs is functionally linked to mTOR-dependent changes in TE rates. We compared 5'UTR pyrimidine content of mouse RP-encoding mRNAs exhibiting the greatest changes in TE to those exhibiting the least changes in TE and saw no difference at the 5'-most region of the 5'UTRs (Fig. S5.3F). Strikingly, however, the 3'-most 5'UTR region of the most affected genes contained a significantly higher proportion of pyrimidines compared to the least affected genes (Welch's two-tailed t -test: $p=0.036$), indicating that pyrimidine richness at LARP1-interacting regions of 5'UTRs is correlated with strong decreases in TE upon mTOR inactivation. Together, these results suggest that the relationship between LARP1 binding and decreased TE is conserved from mouse to human.

Of the 88 annotated human ribosomal proteins, mRNAs encoding 84 were expressed in our mRNA-seq libraries under both conditions. Summing LARP1 binding site coverage of these genes confirms increased LARP1 binding at 5'UTRs and slightly decreased LARP1 binding at CDS and 3'UTR upon mTOR inactivation (Fig. 5.3G). We verified the specific interaction between endogenous LARP1 and mRNAs encoding RpS6, S3A, S18, and L26 by RIP assays followed by quantitative PCR (Fig. S5.3G). Taken together, these observations raise the intriguing possibility that the function of LARP1 in regulating RP mRNA trans-

lation may be context-dependent: the interaction of LARP1 with PESs in the 5'UTRs of RP mRNAs may have an inhibitory role, whereas its interaction with 3'UTRs may exert a positive role in RP mRNA translation.

LARP1 functions as a molecular switch for RP mRNA translation. Through biochemical experiments, we confirmed that LARP1 interacts with the RpL32 5'UTR under starvation conditions, but that in response to growth factors the interaction is reduced in an mTOR activity-dependent manner. In contrast, the binding of LARP1 to the RpL32 3'UTR was increased in response to mTOR activation. The consequences of LARP1 interaction with the 5'UTR or 3'UTR of RpL32 to regulate translation were examined using luciferase reporters. Expression from a reporter containing both UTRs was higher compared to a reporter with just the 5'UTR, suggesting that the 3'UTR of RpL32 is important for promoting translation. Under mild serum starvation conditions, a 5'UTR-mutated reporter produced more luciferase protein compared to a wild-type reporter. These data support a model in which LARP1 stimulates translation of RpL32 through its interaction with the 3'UTR while its association with the 5'UTR negatively regulates RpL32 translation.

Loss of LARP1 function causes inefficient RP translation elongation. So far we have observed that LARP1 can function as both suppressor and stimulator for RP mRNA translation in response to mTOR activity. Examination of polysome and monosome fractions for RP mRNAs in LARP1 knockdown cells suggested that efficiency of translation elongation is compromised by loss of LARP1. Through *in vivo* labeling experiments to measure the rate of *de novo* RP protein synthesis, we showed that knockdown of LARP1 significantly reduced levels of newly synthesized RPs, and that prolonged LARP1 knockdown decreased RP expression and significantly reduced global protein synthesis. Together, these observations suggest that LARP1 has multiple roles in translation. Under low mTORC1 activity conditions, LARP1 prevents aberrant assembly of initiation factors onto RP mRNAs; how-

ever, under growth conditions, LARP1 assists efficient formation of active 80S to promote translation elongation of RP mRNAs.

LARP1 plays an essential role in cancer cell proliferation. Consistent with our observations in HEK293T cells, prolonged knockdown of LARP1 decreased the expression of multiple RPs in colon, breast, and prostate cancer cell lines and dramatically suppressed cancer cell proliferation. Interestingly, expression of LARP1 and RP proteins was well correlated and often enhanced in cancer cell lines compared to non-transformed cells. The levels of LARP1 mRNA were also generally increased in metastatic prostate cancer cell lines. These observations support the idea that LARP1 is necessary for RP mRNA translation and suggest that enhanced LARP1 expression might fulfill the elevated potential of dysregulated PI3K-mTORC1 for demanding ribosome biogenesis and proliferation in a subset of cancer cell lines.

5.3 Materials and methods

5.3.1 PUF-9 HITS-CLIP and data analysis

Isolation and sequencing of PUF-9-bound RNA

PUF-9-GFP and GFP-ALG-1 HITS-CLIP were performed similarly to Puf3p PAR-CLIP (Freeberg et al., 2013) with modifications for application in *C. elegans*. Worms expressing PUF-9-GFP (high-copy Is[*puf-9-promoter::puf-9-gfp::puf-9-3'UTR*]) or GFP-ALG-1 (high-copy Is[*alg-1-promoter::gfp-alg-1::alg-1-3'UTR*]) were grown at 25°C on 15cm NGM plates seeded with HB101 bacteria. Synchronized L1 worms were prepared by hypochlorite preparation of embryos and nutation for 22-26 hours in M9. Gravid adults were grown for 50 hours, washed 3 times in M9 solution, and nutated in M9 for 15 minutes. Embryo samples were prepared by hypochlorite preparation of Day 1 gravid adults (50-60 hours). Embryos were nutated for 22-26 hours in M9 and L1 worm samples were filtered through 20 μm filters.

Live samples (worms or embryos) were washed once with water, resuspended in 10ml water on 15cm glass cell culture dishes and placed on ice. Samples were irradiated with 254nm UV at 150 mJ/cm² four times in a Stratalinker 2400 (Stratagene). Samples were centrifuged for 1 minute at 3000xg and water was removed. Samples were flash frozen in liquid nitrogen and ground to a fine powder in a ball-mill homogenizer (Retsch MM400) chilled with liquid nitrogen.

100mg of cryopulverized sample powder was resuspended in 1ml Lysis Buffer (1xPBS, 0.5%NP-40, 1xComplete Mini Protease Inhibitor, EDTA-free). Lysates were cleared by sequential spins at 1,300Xg for 5 minutes and 20,000Xg for 10 minutes at 4°C. 500μl of clarified lysate was passed through a Costar Spin-X filter (Corning), mixed with RNase T₁ (Fermentas) to 1 U/μl, and incubated at 24°C for 15 minutes (with 15 s shaking at 1,000 rpm followed by a 2 minute rest interval on a Thermomixer) followed by a 5 minute incubation on ice.

Lysates were mixed with 25μl anti-GFP magnetic bead slurry (prepared by DMP-mediated crosslinking of 5ug monoclonal anti-GFP antibody (Invitrogen clone 3E6) to 100μl Protein A Dynabeads (Invitrogen) and resuspended in 200μl Lysis Buffer) to pull down PUF-9-GFP or GFP-ALG-1 at 4°C for 1 hr. Beads were washed twice with 1x PBS + 0.1% NP-40, then incubated with 20 μl of 50 U/μl RNase T₁ (Fermentas, 1:20 dilution in 1? PBS) at 24°C for 15 minutes on a Thermomixer (15 s shaking at 1,000 rpm followed by a 2 minute rest interval), followed by a 5 minute incubation on ice. Beads were washed twice with wash buffer (1x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40), twice with high-salt wash buffer (5x PBS, 0.1% SDS, 0.5% deoxycholate, 0.5% NP-40) and twice with 1x PNK buffer (50 mM Tris pH 7.4, 10 mM MgCl₂, 0.5% NP-40).

Beads were incubated with 20 μl of CIP mix (50 mM Tris pH 7.9, 100 mM NaCl, 10 mM MgCl₂, 0.5 U/μl calf intestinal alkaline phosphatase (CIP); NEB) at 37°C for 15 minutes,

with 15 s shaking at 1,000 rpm followed by a 2 minute rest interval on a Thermomixer. After CIP treatment, beads were washed twice with 1x PNK+EGTA buffer (50 mM Tris pH 7.4, 20 mM EGTA, 0.5% NP-40) and twice with 1x PNK buffer.

Beads were incubated with 20 μ l of ligation mix (50 mM Tris pH 7.4, 10 mM MgCl₂, 0.5 mM DTT, 2 μ M pre-adenylated 3' DNA linker, 25% PEG-8000, 10 U/ μ l T4 RNA ligase 2, truncated K227Q; NEB M0351S) at 16°C overnight (\geq 16 h), with 15 s shaking at 1,000 rpm followed by a 2 minute interval on a Thermomixer. After linker ligation, beads were washed three times with 1x PNK+EGTA buffer.

Beads were mixed with 15 μ l of 1x PNK+EGTA buffer and 15 μ l of 4x NuPAGE LDS sample buffer (Invitrogen NP0007), and incubated at 70°C for 10 minutes. Beads were removed, and the supernatant was loaded onto NuPAGE 4-12% Bis-Tris gel (Invitrogen) and run at 150 V for 1 hr. The gel was transferred to Protran BA 85 nitrocellulose membrane (pore size 0.45 μ m, Whatman) using Novex wet transfer at 30 V for 1 hr. A membrane band corresponding to each RNP was excised and transferred to a microfuge tube.

A diagnostic CLIP experiment was performed in parallel, for the purpose of identifying RNP bands on nitrocellulose membranes. Extract preparation, RNP immunoprecipitation, and RNase T₁ digestion were performed as described above. 5' end phosphorylation was performed on-bead in 15 μ l of PNK mix (70 mM Tris pH 7.6, 10 mM MgCl₂, 5 mM DT, 0.75 μ l P³² γ ATP (6000 Ci/mmol 10 mCi/ml Perkin Elmer BLU502Z500UC), 1 U/ μ l T4 polynucleotide kinase (NEB)) and incubated at 37°C for 15 minutes. After SDS-PAGE and transfer, crosslinked RNAs were visualized by autoradiography.

Non-radioactive membrane bands were excised for RNA isolation and library construction. Excised membranes were incubated with 200 μ l of 4 mg/ml Proteinase K prepared in 1x PK buffer (100 mM Tris pH 7.5, 50 mM NaCl, 10 mM EDTA) for 20 minutes at 37°C on a Thermomixer. We added 200 μ l of 7 M urea prepared in 1x PK buffer to the tube followed

by another 20 minute incubation at 37°C. The Proteinase K digestion reaction was mixed with 1 ml of phenol:chloroform:isoamyl alcohol 25:24:1 (Sigma-Aldrich P2069) by vortexing and spun for 5 minutes at 20,000Xg. The liquid phase was transferred into a new tube, mixed with 50 μ l of 3 M NaOAc, 1 ml of 100% ethanol and 1 μ l of 15 mg/ml Glycoblue (Invitrogen), and precipitated for 2 h at -80°C. RNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature followed by two washes with cold 75% ethanol.

RNA pellets were air-dried briefly, resuspended in 10 μ l of PNK mix (70 mM Tris pH 7.6, 10 mM MgCl₂, 5 mM DTT, 1 mM ATP, 1 U/ μ l T4 polynucleotide kinase (NEB)) and incubated at 37°C for 30 minutes. The reaction was combined with 90 μ l of H₂O and 100 μ l of phenol:chloroform:isoamyl alcohol 25:24:1, mixed well and spun for 5 minutes at 20,000Xg. The liquid phase was mixed with 12.5 μ l of 3 M NaOAc, 250 μ l of 100% ethanol, 1 μ l of 15 mg/ml glycoblue and precipitated for 2 h at -80°C. RNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol.

RNA pellets were resuspended in 10 μ l of ligation mix (50 mM Tris pH 7.5, 10 mM MgCl₂, 10 mM DTT, 1 mM ATP, 0.1 mg/ml bovine serum albumin, 2 μ M 5' RNA linker, 1 U/ μ l T4 RNA ligase (Fermentas EL0021), 10% DMSO) and incubated at 15°C for 2 h.

Ligation reaction was terminated by adding 10 μ l of 2x formamide gel loading buffer (Invitrogen AM8546G), heated for 2 minutes at 70°C and then quickly chilled on ice. Samples were loaded onto a 6% TBE UREA gel (Invitrogen EC6865BOX) and run at 150 V for 45 minutes. After staining with 1x Sybr Gold Stain (Invitrogen S-11494), a gel piece corresponding to a 70 to 90 nucleotide RNA (80 to 100 nucleotide single-stranded DNA) was excised, crushed, and soaked in 400 μ l of 0.3 M NaOAc overnight at room temperature. After removing gel pieces, the solution was combined with 1 ml of 100% ethanol and 1 μ l of 15 mg/ml glycoblue and precipitated for 2 h at -80°C. RNAs were collected by centrifugation

for 20 minutes at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol. After brief drying, RNAs were resuspended in 15 μ l of H₂O.

The ligated RNA (10 μ l) was combined with 2 μ l of 5 μ M RT primer, heated at 65°C for 5 minutes, and then quickly chilled on ice, and followed by the addition of 1 μ l of water, 1 μ l of 10 mM dNTP, 1 μ l of 0.1 M DTT, 4 μ l of 5x First strand buffer, and 0.5 μ l of SuperScript III Reverse transcriptase (Invitrogen, 200 U/ μ l). The RT reaction was kept at 50°C for 45 minutes, 55°C for 15 minutes and 90°C for 5 minutes. A test PCR was performed with 2.5 μ l of RT product in 20 μ l PCR mix: 1x AccuPrime PCR buffer I, 0.5 μ M P5 long primer, 0.5 μ M P7 primer, 0.2 μ l AccuPrime Taq High Fidelity (Invitrogen, 5 U/ μ l). PCR was carried out with an initial 3 minute denaturation at 98°C, followed by 14 to 22 cycles of 80 s denaturation at 98°C, 90 s annealing and extension at 65°C, and termination with a final 5 minute extension at 65°C. PCR product (15 μ l) was collected after 14, 18, and 22 cycles and analyzed on a 10% TBE gel (Invitrogen) at 150 V for 1 h to determine the optimal amplification cycles (the lowest cycle number required to generate 96 to 116 bp amplicons detected by Sybr Gold staining).

Preparation of sequencing libraries

A 50 μ l PCR reaction was carried out with the determined cycle number. Amplicons were purified using DNA clean and concentrator-5 (Zymo D4013; Irvine, CA, USA), run on 10% TBE gels at 150 V for 1 h and stained with Sybr Gold. A gel piece corresponding to 96 to 116 bp DNA was excised, crushed, and soaked overnight in 400 μ l 0.3 M NaOAc at room temperature. After removing gel pieces, the solution was combined with 1 ml of 100% ethanol and 1 μ l of 15 mg/ml glycoblue and precipitated for 2 h at -80°C. DNAs were collected by centrifugation for 20 minutes at 20,000Xg at room temperature, followed by two washes with cold 75% ethanol. After brief drying, amplicons were resuspended in 20 μ l of H₂O. Purified amplicons (5 μ l) were used to seed a second round of PCR in 50 μ l:

1x AccuPrime PCR buffer I, 0.5 μ M Illumina Primer A, 0.5 μ M Illumina Primer B, 0.2 μ l AccuPrime Taq High Fidelity for 6 to 12 cycles. Second PCR amplicons were purified with DNA clean and concentrator-5 (Zymo) and sequenced on an Illumina HiSeq 2000 sequencer. Barcodes (5' RNA linker, barcoded 3' DNA linkers) and primers were exactly the same as in Freeberg et al. (2013).

Sequence read processing

HITS-CLIP reads were processed to remove linkers and low-quality reads. Reads from ALG-1 HITS-CLIP libraries were mapped to miRBase v19 using BWA 0.6.2-r126 (Li and Durbin, 2009), and reads mapping to miRNA sequences were set aside. Remaining ALG-1 HITS-CLIP reads and all PUF-9 HITS-CLIP reads >15nt were aligned to the masked *C. elegans* genome version WS220 using BWA 0.6.2-r126. Optimal alignments were kept based on mismatches per read length. Read clusters were generated from single nucleotide overlapping reads and Gaussian smoothing was applied to resolve multi-peak clusters as described in Freeberg et al. (2013). Read clusters were annotated to genic regions from WS220, including predicted 3'UTRs which we defined as 1.5kb downstream from any ORF with no annotated 3'UTR. Clusters mapping anti-sense to annotated genes were removed from downstream analyses. Cluster read counts were normalized to the total number of million mapped reads in each library (RPM) and additionally to gene RPKM values (Hillier et al., 2009) for the gene to which each cluster was aligned.

Functional category enrichment analysis

DAVID (Huang et al., 2009a,b) was used to derived enriched functional categories for genes targeted by PUF-9 in each stage. Background gene lists were restricted to genes expressed in each specific stage as determined by the modENCODE group (Hillier et al., 2009).

Identification of sequence motifs

MEME (Bailey and Elkan, 1994) was used to search 3'UTR PUF-9 binding sites in each stage for over-represented sequence motifs. The following parameters were used: -zoops -minw 6 -maxw 9.

RNA structure prediction

RNAfold from the ViennaRNA Package 2.0 (Lorenz et al., 2011) was used to determine optimal minimum free energy (MFE) scores for all PUF-9 and ALG-1 binding sites. Each binding site was extended 15nt in either direction for prediction. As matched controls for each binding site, RNA regions immediately upstream and downstream of each site were chosen, with the same length, and used as input to RNAfold to derive MFE scores. The RNAfold web server (<http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi>) was used to generate example RNA structures for select PUF-9 binding sites.

5.3.2 LARP1 PAR-CLIP-seq and data analysis

Isolation and sequencing of LARP1-bound RNA

To label cells with 4-thiouridine (4sU), cells were seeded in two 15 cm plates to grow overnight to reach 70% confluency and on the next day, cells were incubated with 100 μ M 4sU for 14 hours. After washing with cold PBS, cells were irradiated in the CL-1000 ultraviolet Crosslinker (UVP) on ice with 150 mJ/cm². Cross-linked cells pellets were collected by scrapping and lysed with 600 μ l PAR-CLIP lysis buffer (0.1% SDS, 0.5% deoxycholate, 0.5% NP-40 in PBS without Mg²⁺, Ca²⁺) on ice for 10 minutes. To remove DNA, 10 μ l of RQ1 DNase was added into each tube and tubes were incubated at 37°C for 10 minutes with gentle rocking. Afterwards, 1U/ μ l of RNase T₁ (Fermentas) was supplemented and the lysates were incubated at room temperature for 15 min with gentle rocking. Lysates were spun at 4°C for 10 minutes at 14,000 rpm. The soluble fractions were incubated

with LARP1 antibody-Dynabeads Protein A for 1 hour at 4°C. Immunoprecipitates were collected on the magnetic stand and were washed three times with IP-wash buffer (50 mM HEPES-KOH [pH 7.5], 300 mM KCl, 0.05% NP-40, 0.5 mM DTT, and EDTA-free protease inhibitor). Immunoprecipitates were resuspended in 40 μ l of IP-wash buffer containing 50 U/ μ l RNase T₁ (Fermentas) and incubated at room temperature with gentle rocking for 15 minutes followed by incubation on ice for 5 min. Immunoprecipitates were washed with high-salt wash buffer (50 mM HEPES-KOH [pH 7.5], 500 mM KCl, 0.05% NP-40, 0.5 mM DTT, and EDTA-free protease inhibitor) and three times with the PAR-CLIP lysis buffer, and washed twice with high-salt wash buffer (0.1% SDS, 0.5% deoxycholate, 0.5% NP-40 in 5X PBS without Mg²⁺, Ca²⁺) followed by twice washing with PNK (polynucleotide kinase) buffer (50 mM Tris-Cl pH 7.4, 10 mM MgCl₂, 0.5% NP-40). For visualization of crosslinked RNAs, immunoprecipitates were incubated in 40 μ l of the PNK mixture (1 μ l of P³² γ ATP, 4 μ l of 10X PNK buffer [NEB], 2 μ l of T4 PNK enzyme [NEB], 33 μ l of water) for 30 minutes at 37°C. Labeled immunoprecipitates were washed three times with PNK buffer and resuspended with 30 μ l of 2X NUPAGE LDS sample buffer (15 μ l of 1 X PNK with 15 μ l of Novex LDS sample buffer). Denatured samples were resolved in 4-12% NUPAGE Bis-Tris gel and transferred to nitrocellulose membrane at 30 V for 1 hr using NuPAGE transfer buffer. The membrane was expose to X-ray film at -80°C for 4 hours to visualize crosslinked RNAs. Molecular biology procedures for cloning LARP1-bound RNA fragments were described previously (Freeberg et al., 2013).

Sequence read processing

PAR-CLIP-seq and mRNA-seq reads were processed to remove linkers and low-quality reads. Reads were aligned to the human transcriptome version GRCh37 using Bowtie (Langmead et al., 2009) allowing for up to 3 mismatches with the following parameters: -v 3 -k 100 -best -strata -phred33-quals. mRNA-seq reads aligning perfectly to the transcriptome were

kept; reads aligning perfectly to multiple loci were distributed evenly among the mapped positions. Transcript RPKM values were calculated as the number of reads per million mapped reads aligning to a transcript normalized to transcript length in kilobases. Replicate mRNA-seq libraries had a high Pearson correlation coefficient ($R^2=0.9998$), so transcript RPKM values were averaged from the two libraries. PAR-CLIP-seq reads with 0-2 T-to-C mismatches were clustered into peaks with at least 1 overlapping nucleotide. Clusters were smoothed with a Gaussian smoothing technique as described in Freeberg et al. (2013). Clusters with at least 1 read containing a T-to-C conversion event were kept as LARP1 binding sites, and all reads containing 0-2 T-to-C conversion events were summed per binding site and normalized to the number of million mapped reads per library (RPM). Binding site RPM values were additionally normalized to gene RPKM and multiplied by 1,000 to account for the kilobase normalization of gene RPKM values.

Gene ontology term enrichment analysis

GO term enrichment analysis was performed using the topGO Bioconductor package for R (Alexa et al., 2006). The background gene list was restricted to genes with reads in at least one of our replicate mRNA-seq libraries. The Fisher's exact test was used to measure the significance of enriched GO terms, and p -values were corrected for multiple testing using the Bonferroni correction method. GO terms with adjusted p -values < 0.001 were manually grouped into the following categories: translation, cell differentiation and development, protein localization to the ER, regulation of signaling response, response to stimulus, antibody production, metabolism, and other.

Identification of sequence motifs

Sequence motifs were searched for within 5'UTR LARP1 binding sites found under growing and mTOR inactive conditions, respectively, using MEME with default parameters (Bailey and Elkan, 1994). Binding site sequences were extended by 15nt up- and down-stream for

this search. Additionally, sequence motifs were searched for within CDS and 3'UTR LARP1 binding sites under growing and mTOR inactive conditions using the same parameters.

5.4 Discussion and concluding remarks

5.4.1 PTGR by PUF-9/miRISC is physically linked on 3'UTRs

Examination of PUF-9 and ALG-1 (as a proxy for miRISC) binding sites across developmental stages in *C. elegans* revealed extensive co-regulation of targets, more specifically, at sites that are directly bound by both RBPs (Fig. 5.1F). Our results highlight the pervasiveness of a mechanism that, until now, had only been observed in a few cases (*e.g.* human PUM1 and miR-221/222 regulation of *p27*). Since both Pumilios and miRNAs function by recognition of conserved primary sequence motifs in target 3'UTRs, it is intriguing to consider that the positive selection constraints on each set of motifs might be strengthened if the two motifs could base-pair to form a stable hairpin structure, such as in the *p27* model. Indeed, we showed that in the PUF-9 binding site on the *lin-41* 3'UTR that overlaps with the *let-7* target site, the PUF-9 motif and the *let-7* recognition sequence can pair in the stem of a stable hairpin structure (Fig. 5.2D). A cursory comparison of the PUF-9 recognition motif with sequences recognized by each miRNA in *C. elegans* revealed that only a few out of the hundreds of miRNAs could form hairpins similar to *let-7*. This does not rule out the possibility, though, that the proximity of PREs and miRNA target sites can influence RNA secondary structure.

The few validated examples of co-regulation by Pufs and miRNAs show that binding is cooperative: PUM1 and miR-221/222 on *p27* (Kedde et al., 2010), PUM1/2 and miRNAs on E2F3 (Miles et al., 2012), PUF-9 and *let-7* on *lin-41* and *hbl-1* (Nolde et al., 2007). This makes sense given that, in general, both Pufs and miRNAs promote degrada-

tion/deadenylation of their target mRNAs. However, the proximity of PREs and miRNA target sites could also be evidence for competitive binding. We chose a threshold of 25 nucleotides to identify “close” PUF-9 and ALG-1 binding sites which allows for the possibility of both RBPs to contact a single RNA region. Identification of sites in which the PRE and miRNA target site are too close to sterically allow both RBPs to bind could indicate putative competitive mechanisms of PTGR. Competitive co-regulation, which is probably not as pervasive as cooperative co-regulation, could be linked to less common roles observed for Pumilio in mRNA localization and stability (Quenault et al., 2011).

5.4.2 LARP1 is a model for multiple RBDs promoting multiple PTGR mechanisms

Examination of LARP1 substrates by PAR-CLIP-seq revealed an intriguing PTGR mechanism for the essential class of ribosomal protein-encoding mRNAs. LARP1, in response to intracellular cues mediated by mTORC1, directly binds either 5'UTRs or 3'UTRs to suppress or enhance translation, respectively. Since LARP1 contains both a La RBD and an RRM RBD, each RBD could be responsible for recognizing different RNA substrates under different conditions. Indeed, our PAR-CLIP-seq data revealed a pyrimidine-rich sequence motif among direct LARP1 binding sites located in substrate 5'UTRs (Fig. 5.3D), while a GA-rich motif was discovered among 3'UTR sites (Fig. S5.3C). One hypothesis to explain RP mRNAs having different LARP1 motifs is that each motif could be responsible for promoting different PTGR mechanisms. Indeed, subsequent biochemical assays confirmed that disruption of LARP1 binding at either 3'UTRs or 5'UTRs produces different functional consequences relating to translation of the target. Given that RBD modularity is a feature of many RBPs (Lunde et al., 2007), it is possible that a dual mode of PTGR for a single RBP is much more common than has been previously thought. To disentangle the roles of

each LARP1 RBD in PTGR of its substrates, functional assessments of translation rates on LARP1 substrates could be performed by introducing transgenes expressing versions of LARP1 with mutations in either the La RBD or the RRM RBD. In addition, PAR-CLIP-seq performed on the RNA substrates of these LARP1 RBD mutants could help define which LARP1 direct target sites interact with which RBD.

5.5 Data availability

PUF-9 HITS-CLIP sequence data are currently unavailable to the public but will be submitted to the NCBI Gene Expression Omnibus upon publication of the PUF-9 study. LARP1 PAR-CLIP-seq and mRNA-seq sequence data are available through the NCBI Gene Expression Omnibus using series entry GSE59599.

5.6 Acknowledgements

Danny Yang and Jungsook Park performed a majority of the genetic and molecular biology experiments associated with the PUF-9 study. Ting Han generated the PUF-9 and ALG-1 HITS-CLIP sequencing libraries. This work was supported by the National Science Foundation Open Data IGERT grant 0903629 (MAF), the NIH National Research Service Award T32-GM07544 (DY), and the National Institutes of Health grant GM088565 (JKK).

Sung Ki Hong performed a majority of the genetic and molecular biology experiments associated with the LARP1 study. Ting Han generated the LARP1 PAR-CLIP sequencing libraries. We thank Arul Chinnaiyan and Shaomeng Wang for providing breast and prostate cancer cell lines, respectively, Eric Fearon for helpful advice, Ivan Topisirovic for instructions on polysome fractionation, and Philip Gafken (Fred Hutchinson Cancer Research Center)

for the proteomics study. This work was supported by grants from the NIH (DK083491) and the Ruddon Research Fund in Cancer Biology.

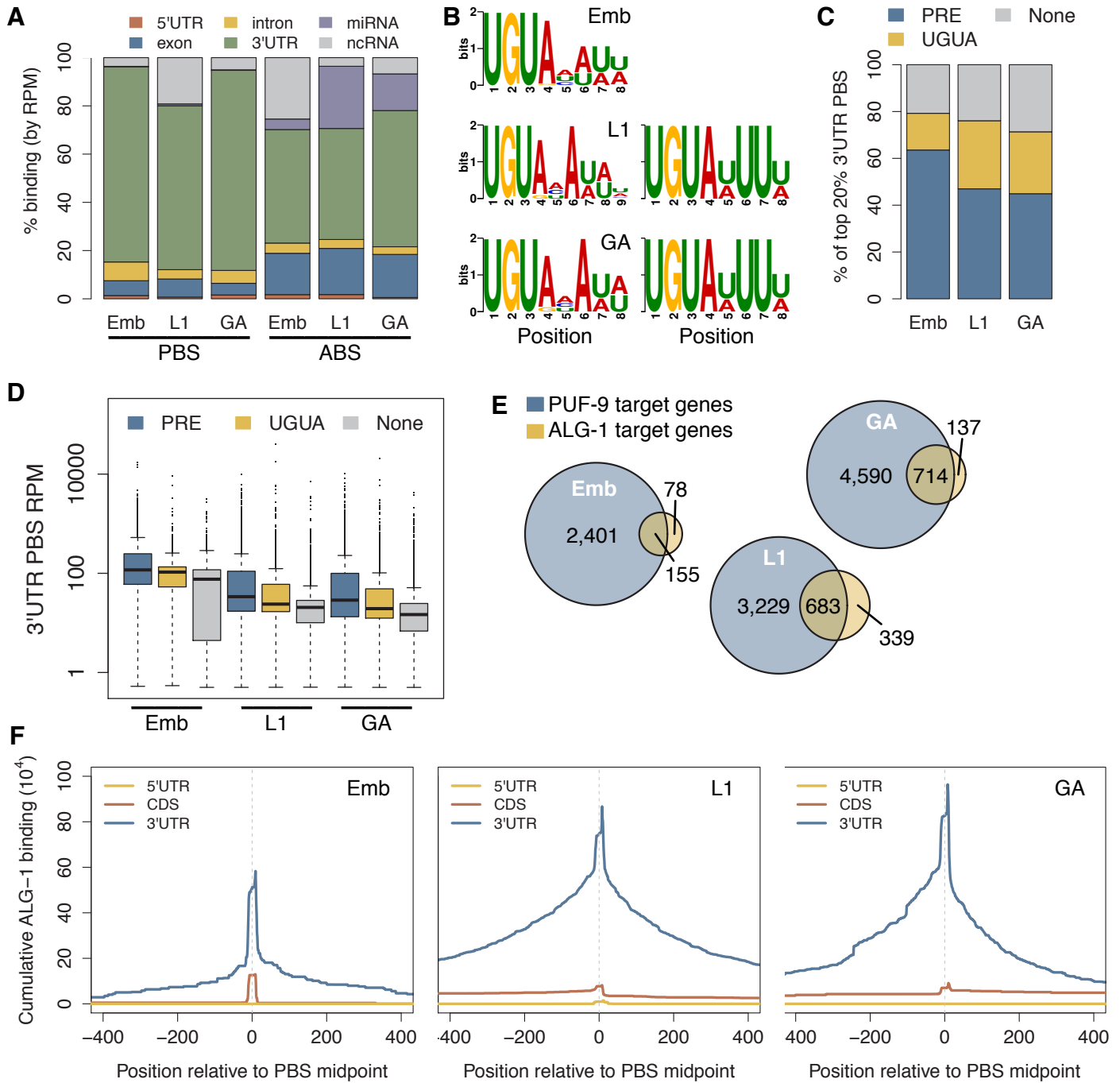


Figure 5.1: (A) PUF-9 directly binds target 3'UTRs; ALG-1 directly binds miRNAs and target 3'UTRs. (B) Full Pumilio recognition elements (PRE) are identified within 3'UTR PUF-9 binding sites (PBSs) in all stages. (C) 70-80% of the most strongly bound PBSs contain either a full PRE or the first few nucleotides of the PRE (UGUA). (D) 3'UTR PBSs with a full PRE or UGUA motif are more efficiently bound by PUF-9 as measured by binding site RPM coverage. (E) A majority of ALG-1 target genes are also bound by PUF-9. (F) ALG-1 binding accumulates at PBSs on co-targeted genes.

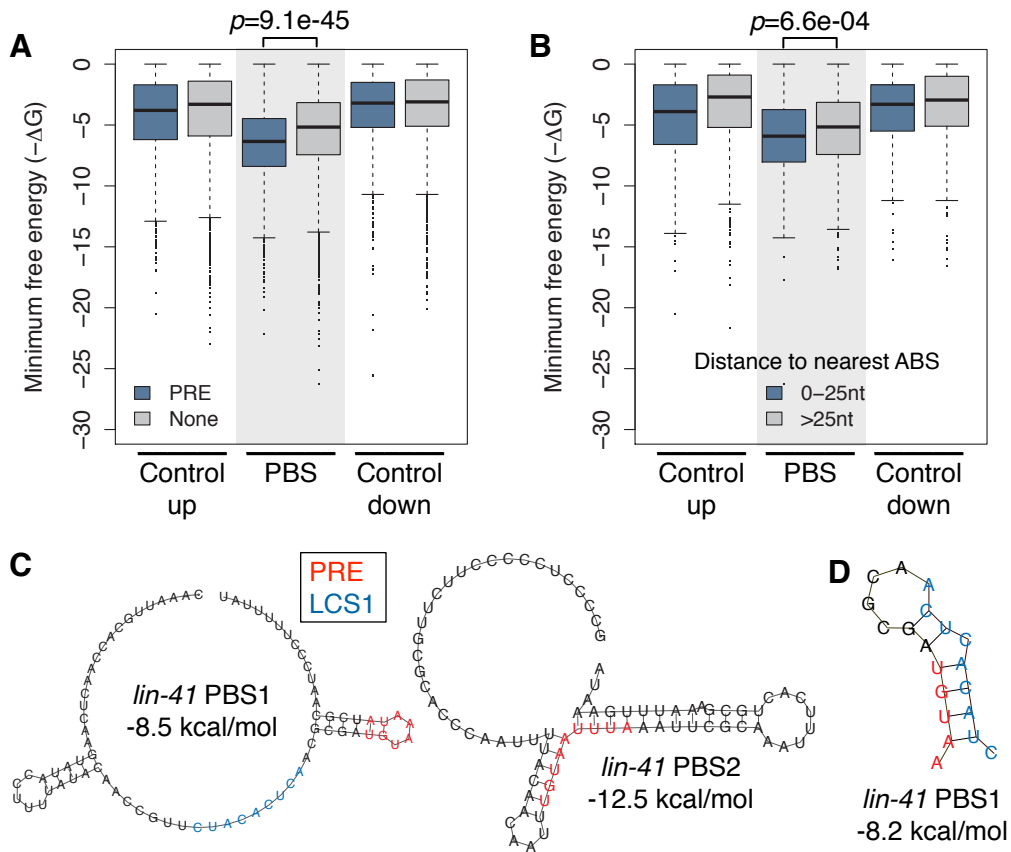


Figure 5.2: (A) PBSs with or without a full PRE are more structured than RNA regions immediately upstream or downstream. RNA structure stability is predicted by RNAfold and plotted are the minimum free energy scores for the best predicted structure of each PBS and control RNA. PBSs with a full PRE are significantly more structured than PBSs with no motif. (B) Same as A but showing that PBSs within 25nt of an ALG-1 binding site are more structured than PBSs located farther away from an ABS. (C) Best predicted RNA structure for the two PBSs on the *lin-41* 3'UTR. Red letters denote the full PRE sequence, and blue letters denote the conserved *let-7* miRNA seed site. (D) Predicted highly structured hairpin formed by direct base pairing of the PRE and *let-7* seed site in PBS1.

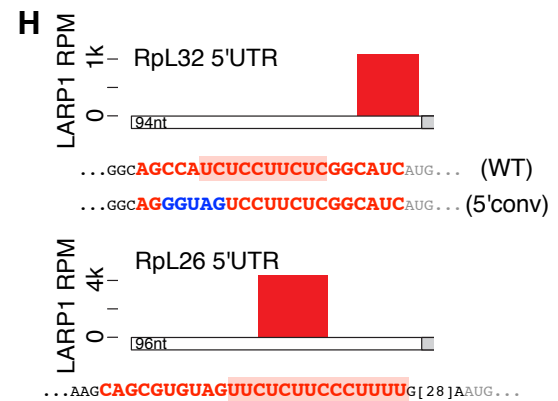
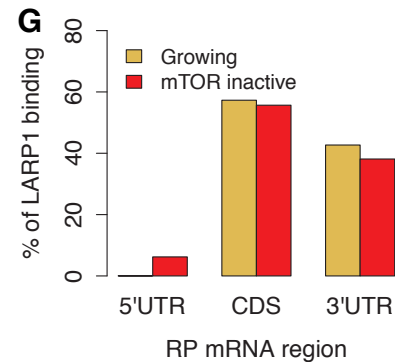
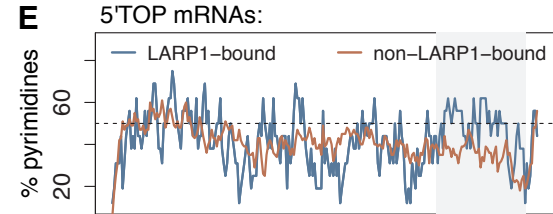
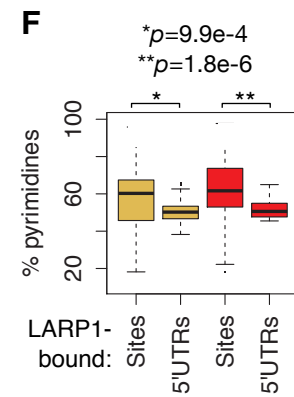
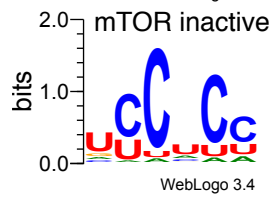
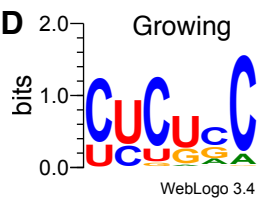
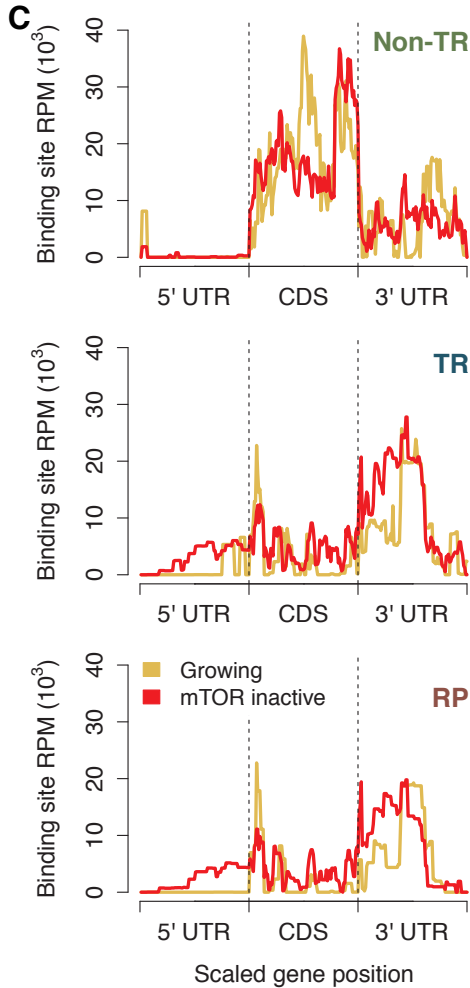
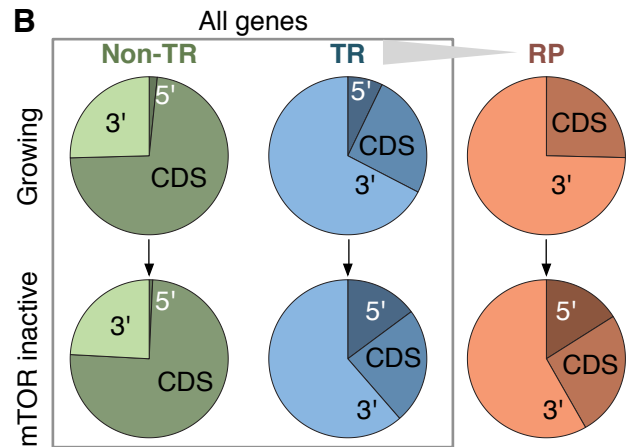
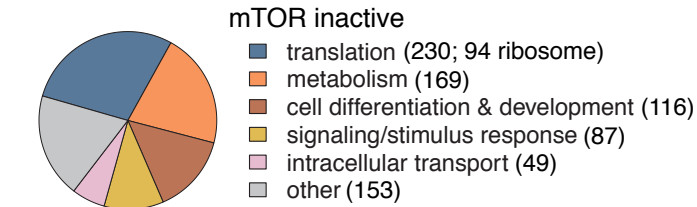
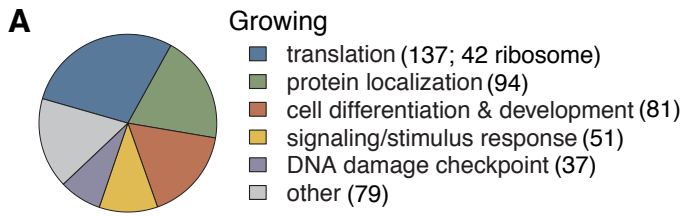


Figure 5.3: (A) LARP1-bound genes are most enriched for GO terms related to translation including RP genes. (B) Upon mTOR inactivation, LARP1 binding at 5'UTRs increases on TR and RP genes. (C) LARP1 binding at TR and RP 5'UTRs under mTOR-inactive conditions tends to occur at the 3'end. (D) LARP1 binds directly to pyrimidine-enriched sequences in 5'UTRs. (E) The LARP1 binding sites at the 3' end of 5'TOP-containing 5'UTRs are enriched for pyrimidines. (F) LARP1-bound sites on 5'UTRs are enriched for pyrimidines compared to the rest of the 5'UTR sequence. Welch's two-tailed *t*-test: * $p=1.4e-15$ and ** $p=1.2e-18$. (G) LARP1 binding on RP-encoding mRNAs is gained at 5'UTRs upon mTOR inactivation and slightly decreased at CDS and 3'UTRs. (H) mRNAs encoding RpL26 and RpS24 have a single 3'UTR LARP1 site at their 3'UTRs under mTOR-inactive conditions. Red letters highlight the LARP1 binding site and red boxes indicated the pyrimidine-enriched sequence. Start codons are in grey.

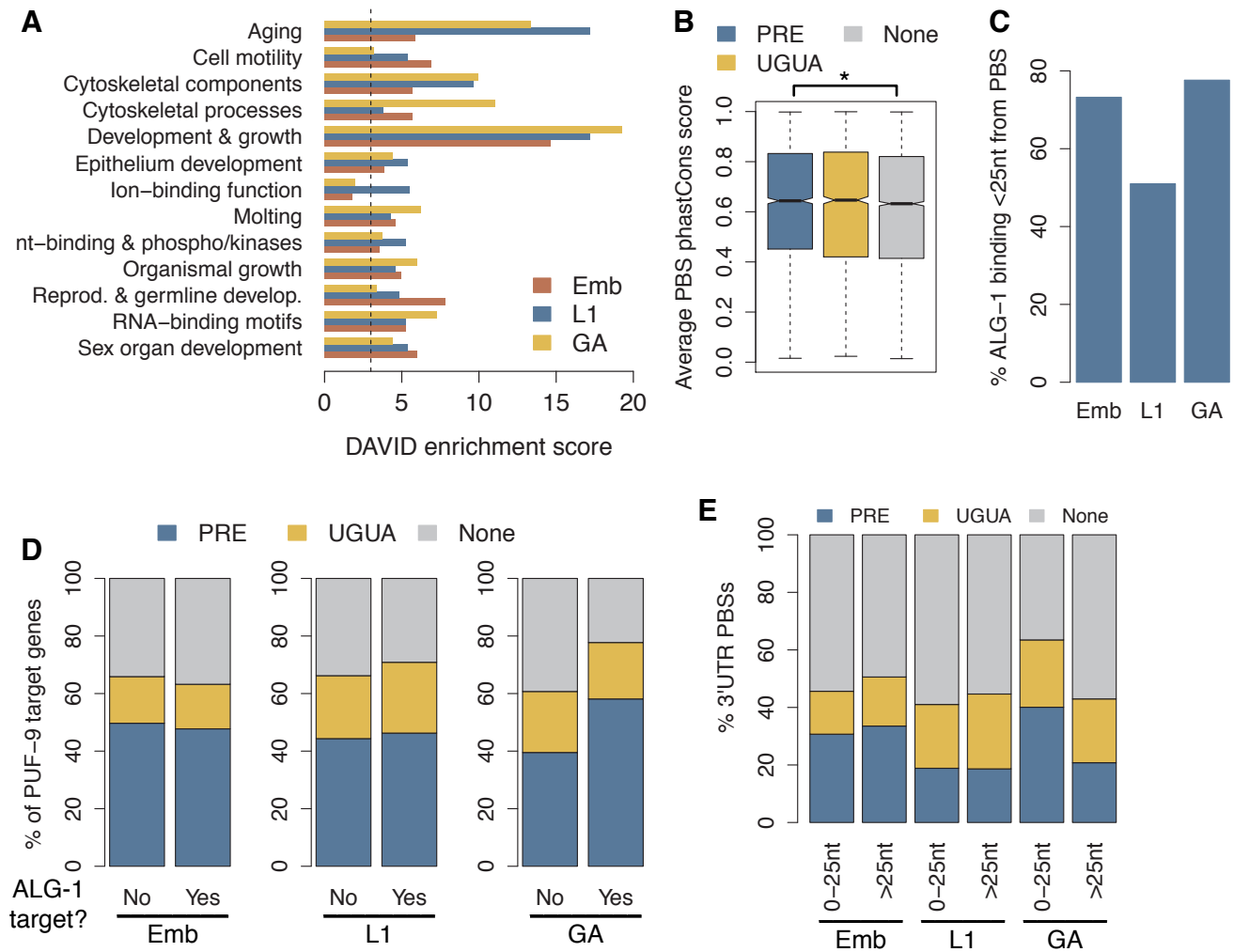


Figure S5.1: (A) PUF-9 targets are enriched for functional categories (including GO terms and KEGG pathways) related to growth and development, aging, cytoskeleton, reproduction, and germline development as determined using the DAVID Bioinformatics Resource (Huang et al., 2009a,b). (B) PBS with a full PRE show significantly higher conservation than PBSs with no motif. $*p < 0.001$. (C) A majority of ALG-1 binding at 3'UTRs occurs within 25 nucleotides of a PBS. (D) Only in gravid adult animals do PUF-9/ALG-1 co-targeted genes exhibit a higher proportion of PBSs with a full PRE than genes targeted only by PUF-9. (E) Only in gravid adult animals do PBSs within 25nt of an ALG-1 binding site show enrichment for presence of a full PRE.

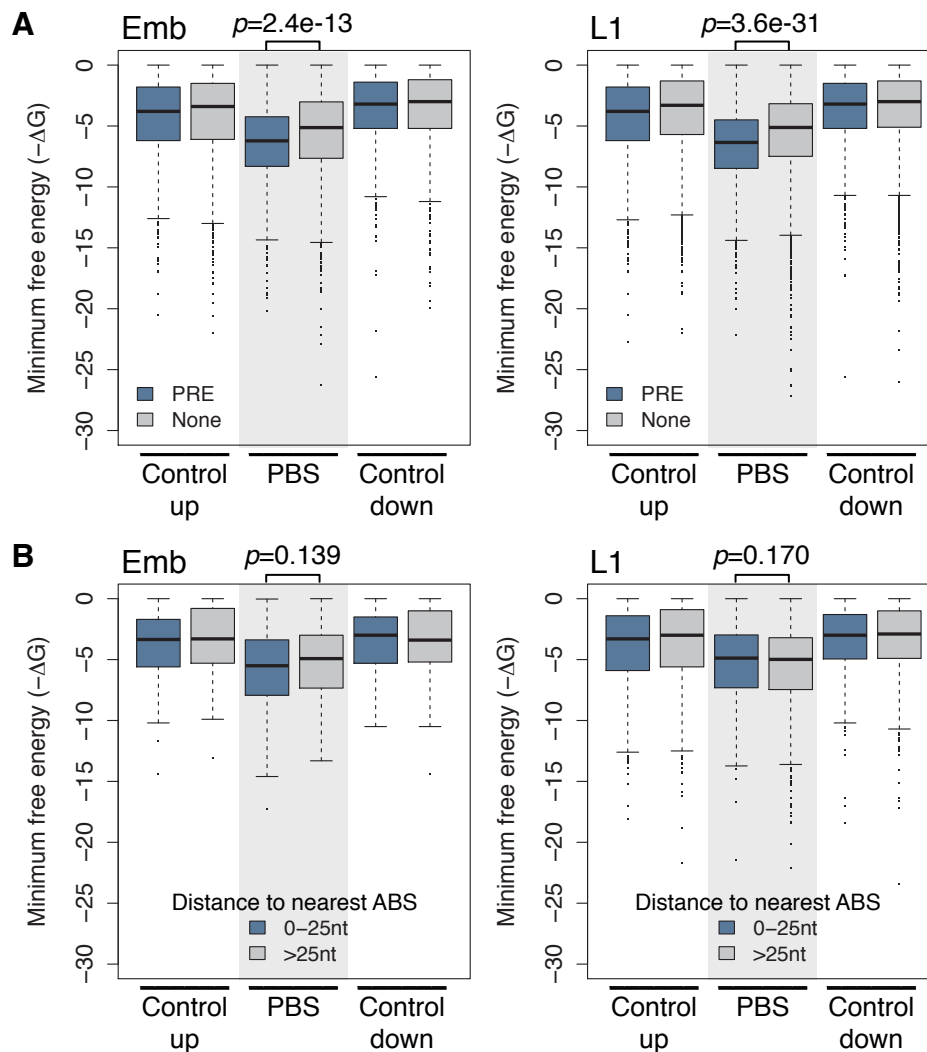
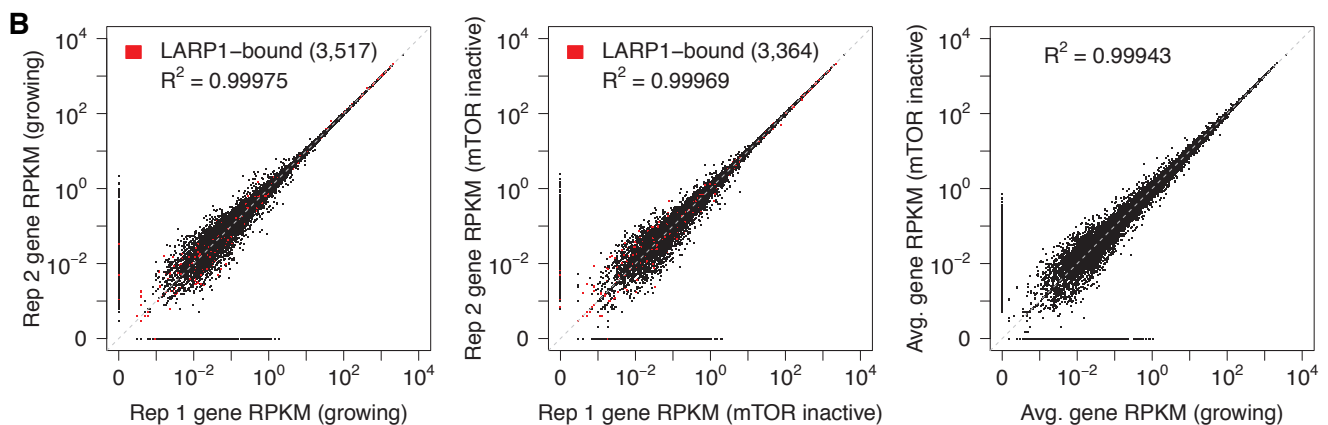
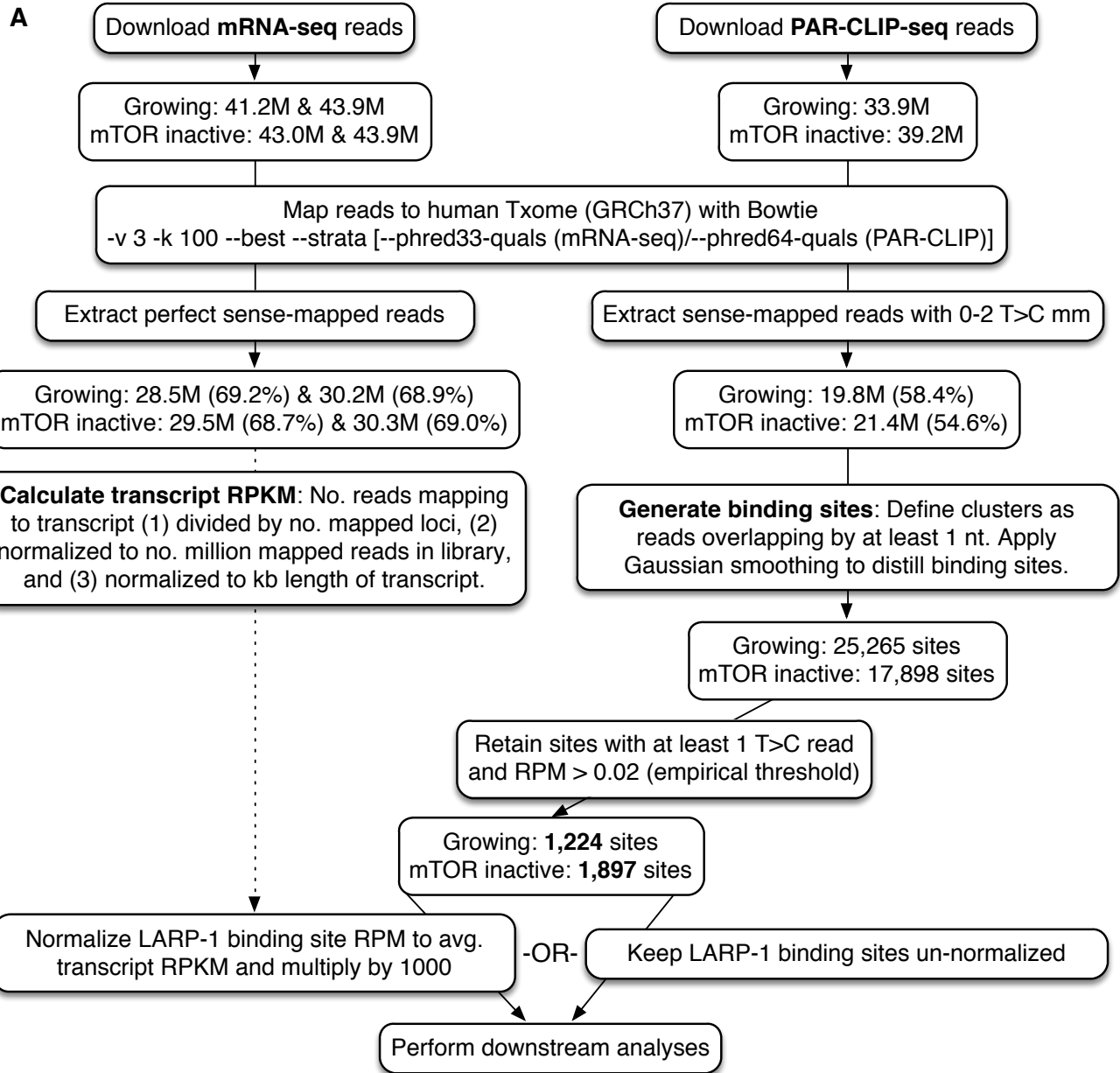


Figure S5.2: (A) PBSs with or without a full PRE in embryo and L1 animals are more structured than RNA regions immediately upstream or downstream. RNA structure stability is predicted by RNAfold and plotted are the minimum free energy scores for the best predicted structure of each PBS and control RNA. PBSs with a full PRE are significantly more structured than PBSs with no motif. (B) Same as A but showing that PBSs within 25nt of an ALG-1 binding site are not more structured than PBSs located farther away from an ABS in embryo and L1 animals.



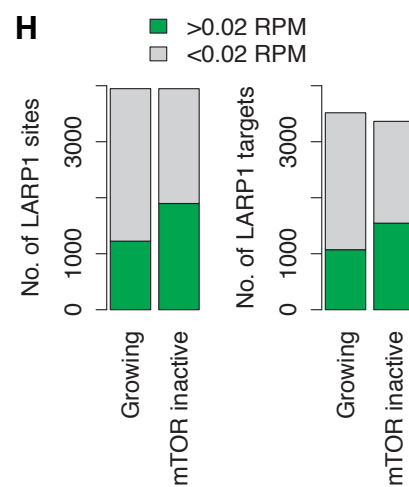
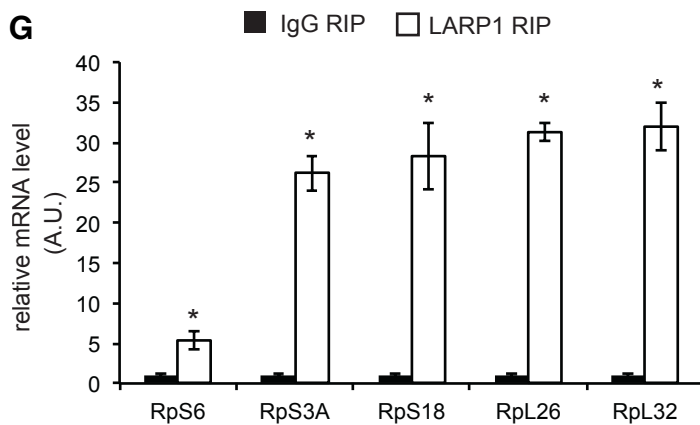
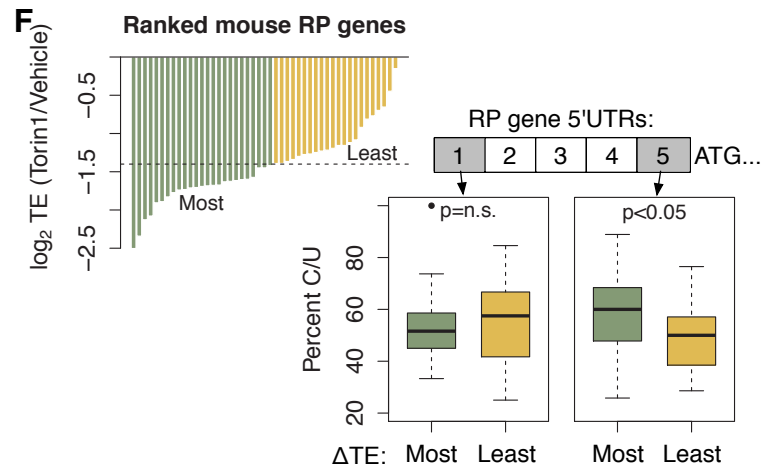
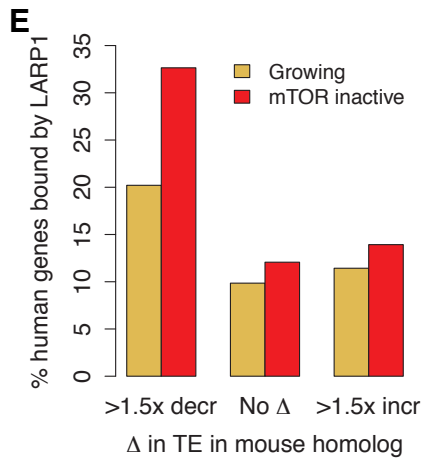
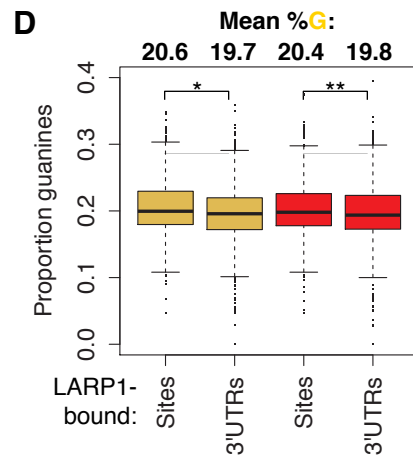
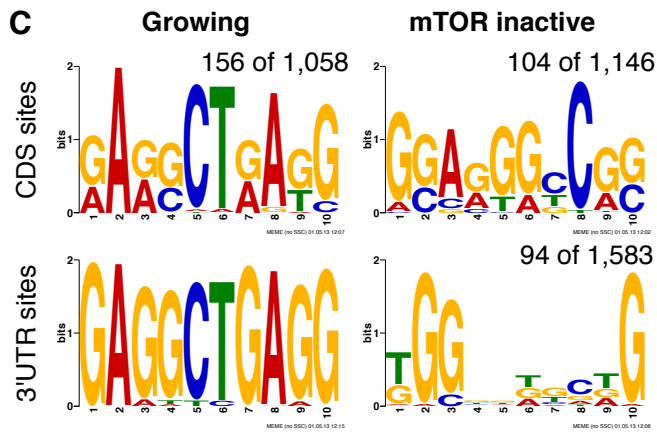


Figure S5.3: (A) Flowchart of PAR-CLIP-seq and mRNA-seq data processing. (B) Replicate analysis for mRNA-seq data. (C) MEME-derived motifs searching within 3'UTR LARP-1 binding sites. Maximum motif width of 10 (-maxw 10) and zero or one motifs per sequence (-zoops) parameters were used. (D) LARP1-bound sites on 3'UTRs are only slightly, but significantly, enriched for guanines compared to the rest of the 3'UTR sequence. Welch's two-tailed t -test: $*p=4.1e-8$ and $**p=1.8e-4$. (E) Translational efficiency (TE) measurements for 4,840 genes in mouse cells upon Torin1 or Vehicle control treatment (Thoreen et al., 2012) were downloaded. The NCBI HomoloGene database (Release 68) was used to identify human homologs for 4,351 of these genes. Mouse genes were grouped based on their change in TE measured. Plotted are the % of human homologs of the mouse genes in each group that are bound by LARP1 in our study. (F) The 5'UTRs of 47 RP-encoding mouse mRNAs were divided into 5 equal sections, and the percent of pyrimidines in each section for each 5'UTR was computed. Mouse genes were then divided into two groups based on the change in TE measured after treatment with Torin1: genes that were affected the most (>1.4 -fold change in TE, "Most") and genes that were affected the least (<1.4 -fold change in TE, "Least"). We saw no difference in pyrimidine content at the 5'-most section of the 5'UTRs. The 3'-most section of 5'UTRs of the most-affected genes contained a significantly higher proportion of pyrimidines compared to the least affected genes (Welch's two-tailed t -test: $p=0.036$). (G) Select RP-encoding mRNAs immunopurified with LARP1. (H) Approximately 4,000 LARP1 binding sites were identified with T-to-C mismatches. Plotted are the numbers of these sites (and corresponding genes) that pass the empirical 0.02 RPM threshold.

CHAPTER VI

Perspectives on future investigations of PTGR

6.1 Exploring the *S. cerevisiae* RBPome

In Chapter Three, I described global and dynamic characteristics of transcriptome-wide binding patterns of RBPs in budding yeast. Additional studies were subsequently published that identified transcriptome-wide RBP footprints in human embryonic kidney (HEK) (Baltz et al., 2012; Schueler et al., 2014), HeLa (Silverman et al., 2014), and mammary epithelial adenocarcinoma cell lines (Schueler et al., 2014). To complement these studies, identification of all proteins capable of interacting with the transcriptome (the mRBPome) or RNA in general (the RBPome) is of paramount importance.

Novel mRBPs identified in yeast. In an effort to expand the set of known and experimentally validated RBPs in yeast, we conducted a pilot experiment combining photoactivatable ribonucleoside-enhanced crosslinking with oligo(dT) affinity purification and tandem mass spectrometry (PAR-CL-MS) to systematically capture *in vivo* mRBP:mRNA interactions in yeast under normal growth conditions (Fig. 6.1A). By comparing average protein spectral counts between crosslinked and non-crosslinked (control) samples, we identified 259 mRBPs, of which 125 contain no canonical RBDs (Fig. 6.1B). After accounting for known yeast RBPs and yeast homologs of mammalian RBPs, we identified 87 novel yeast mRBPs that function in a wide array of diverse pathways including those related to metabolism, intracellular trafficking, and chromatin remodeling (Table 6.1). Less than 7% of

our 259 mRBPs were identified in *in vitro* protein microarray experiments, highlighting the severe discrepancies that exist between interrogating RBP:RNA interactions in an artificial setting versus a biologically meaningful context. Conversely, 20% of our 259 mRBPs were identified *in vivo* in stress-induced RNP granules and $\sim 40\%$ have human homologs identified *in vivo* in HeLa and HEK cell lines, which supports using *in vivo* methods for interrogating RBP:RNA interactions and emphasizes the evolutionarily conserved role for RBPs across the eukaryotic kingdom.

Although PAR-CL-MS captures many RBPs with annotated RBDs or evidence of binding RNA in other global studies, we still fall short of the approximately 800 RBPs predicted to be encoded by the yeast genome. We likely did not capture non-polyadenylated RNA-binding proteins, low abundance RBPs, or RBPs whose interaction with RNA is extremely transient or context-dependent. For example, we might miss RBP:RNA interactions that only occur in response to environmental stress, such as those identified in glucose-starvation-induced stress RNP granules (Mitchell et al., 2013). Thus, future experiments are required - both in normal growing and stress conditions - to more fully characterize the RBPome of budding yeast.

Characteristics of novel RBPs point to “dual function” properties. Understanding the differences between canonical RBPs and proteins with secondary or context-dependent RNA-binding roles represents a major step towards untangling complex networks of RBP:RNA interactions that underlie PTGR. What differentiates canonical RBPs from proteins that are only now being identified as RNA-interacting? Are intrinsic features shared between or unique to the two classes of RBPs? We can begin to answer these questions by characterizing features of the 87 novel yeast mRBPs we identified which contain no known RBDs and have not been shown to bind RNA in other assays or by computational prediction.

Our novel mRBPs are enriched for GO terms relating to metabolic and energy-related pro-

cesses, adding support to the REM network hypothesis proposing a functional connection between RNA biology, intermediary metabolism, and post-transcriptional gene regulation (Fig. 6.2A). We identified non-RBD Pfam domains enriched among our novel mRBPs including ABC transport domains, enzymatic and metabolic domains, and nucleotide cofactor-binding domains (Fig. 6.2B). Two novel mRBPs, alcohol dehydrogenase regulator Adr1p (Znf-H2C2) and GTPase-activating protein Sec23p (Znf-Sec23_Sec24), contain non-canonical RBDs from the zinc finger family of DNA-binding domains, further expanding the list of DNA-binding zinc finger domains that can interact with RNA. This observation suggests an intriguing mechanism in which DNA-binding zinc finger domains might be able to recognize highly structured dsRNA motifs rather than ssRNA primary sequences. Heat shock proteins (HSPs) can recognize and bind to AU-rich elements in 3'UTRs to stabilize transcripts within stress RNP granules (Buchan and Parker, 2009; Laroia et al., 1999). We identified four novel mRBPs with heat shock domains, suggesting that this regulation occurs in yeast. By identifying mRBPs in yeast subjected to temperature stresses, we may uncover additional HSPs that contribute to stress RNP granule formation. Like HSPs, many novel RBPs have clear molecular functions and also bind RNA. Discovering how these activities are integrated and regulated remains a major future challenge.

Intrinsically disordered (ID) and low complexity (LC) domains are key emerging feature of RBPs. We observe more disordered regions at the N-termini of novel mRBPs compared to the proteome (Fig. 6.2C). ID domains occur at both termini for all mRBPs, confirming that the termini of proteins are more flexible for adapting to various interfaces when coming in contact with ligands such as other proteins, nucleic acids, or signaling molecules (Teilum et al., 2009). The lack of disordered residues at the C-termini of novel mRBPs is indicative of a model whereby the protein region responsible for binding RNA is a highly disordered, N-terminal domain, while the highly structured region(s) confer additional molecular function(s).

Within ID regions of our novel mRBPs, we identified many LC regions consisting of short, repetitive amino acid sequences. Enriched in the LC regions were (1) alanine and glycine, amino acids with small non-polar side chains that are less likely to constrain a peptide structure, (2) basic amino acids lysine and arginine, which form favorable electrostatic interactions with negatively charged RNA, and (3) tyrosine, of which the aromatic ring can form stacking interactions with unpaired nucleotides (Jones et al., 2001). As expected, order-promoting residues (Trp, Cys, Phe, Ile, Tyr, Val, Leu) were among the least represented amino acids in ID regions, while disorder-promoting residues (Ser, Ala, Pro, Arg, Glu, Lys, Gln) were among the most represented (Magidovich et al., 2007). These observations confirm that novel mRBPs, which do not contain canonical RBDs, conform to sequence-level patterns within ID and LC domains, which are proposed to interact with RNAs. Integration of these sequence patterns into RBP-prediction algorithms could further help to refine not only which proteins might bind RNA, but also what protein domains might confer RNA-binding functionality.

While many protein structural domain databases, such as SCOP (Structural Classification of Proteins) (Murzin et al., 1995) and CATH (Protein Structure and Classification Database) (Orengo et al., 1997), rely on secondary and tertiary structures to define functional domains, it is clear that lack of a defined structure within an ID/LC domain is itself a functional unit. ID /LC domains can indirectly be found within the RCSB Protein Data Bank (PDB), an online repository for three-dimensional protein structure data determined by X-ray crystallography, nuclear magnetic resonance (NMR) imaging, cryo-electron microscopy, and structure modeling (Berman et al., 2000). Within X-ray crystallography data, for example, regions of a crystallized protein or peptide that are flexible cannot be resolved from diffraction patterns since the flexible regions produce multiple signals that will not converge on a single predicted structure within acceptable error (Oroguchi et al., 2011). These regions are often left out of the final model submitted to the PDB. Thus, lack of

a resolved structure for certain protein/peptide regions gives some evidence to the region being either uncrystallizable or a flexible, intrinsically disordered region. Future work on protein domain classifications should take into consideration that ID/LC domains might be a distinct functional as well as structural, domain: functional in that these domains are necessary for interfacing with substrates and structural in that these domains are naturally unstructured until they come into contact with a substrate, at which point they adopt the necessary structure to enable substrate binding. In addition, LC domains are necessary and sufficient for forming higher-order phase transitions that promote the re-organization and concentration of cellular constituents. Therefore, how this RBP- and LC domain-mediated remodeling occurs and mediates PTGR will be a major future goal of the field.

Cytoskeletal components are intriguing RBPs. As mentioned in the Introduction, mRNA localization has long been proposed as an important function in large and polarized cells to enable quicker protein synthesis at sites where the protein is needed. While this phenomenon has been well-characterized for a few mRNAs in some specialized cells types (*e.g.* oocytes/embryos, neurons), mounting evidence suggests that localized mRNA translation is actually a pervasive PTGR mechanisms and might be the rule rather than exception (Lécuyer et al., 2007; Sharp et al., 2011; Du et al., 2008; Blower et al., 2007). For example, the cellular cytoskeletal network provides critical sites for aggregation of mRNAs and translation components and is thought to promote translation. Intriguingly, it could also provide a system whereby mRNAs could be physically transported through the cell with the help of other RBPs (Jansen, 1999). Microtubules themselves can bind RNA in specific contexts. For example, the microtubule scaffolding protein APC binds RNAs encoding proteins with microtubule organization and cell motility function, suggesting a model for APC to localize translation of its own subunits and cofactors (Preitner et al., 2014). During cell division, microtubule-bound mRNAs appear to be translationally repressed, suggesting that the mi-

totic spindle might be functioning to actively or passively direct asymmetric and symmetric segregation of developmentally important mRNAs during cell division.

Within our yeast mRBPome, we identified microtubule- and cytoskeleton-related factors providing evidence for roles in post-transcriptionally regulating sets of mRNAs. The next clear steps to explore the role of cytoskeletal components in PTGR are to characterize the global sets of RNA substrates bound by these factors to answer the following: Are the mRNA substrates being actively translated or sequestered into silencing/degradation granules? What RNA substrates are bound during different cell cycle phases, and do substrates change when cells undergo mitosis? Are the substrates that are bound by cytoskeletal components present diffusely throughout the cytoplasm or in granule-like structures? What happens to mRNA segregation patterns when microtubule RBDs cannot function? By answering these and other questions, we can more thoroughly understand how and why the cytoskeletal network binds RNA and potentially mediates a novel form of post-transcriptional gene regulation.

6.2 RNA binding, glycolytic enzymes, and metabolic flux

Among the novel mRBPs identified in our pilot yeast study were glycolytic enzymes including Pfk2p, the β subunit of phosphofructokinase (Pfk). Pfk2p, along with the Pfk α subunit Pfk1p, are conserved from bacteria to humans. Pfk is one of the most crucial enzymes in the glycolytic pathway, responsible for catalyzing an early and rate-limiting step to phosphorylate fructose-6-phosphate. Mutations in human Pfk expressed specifically in

Acknowledgments:

Ting Han made the initial Pfk2p granule formation observations. Ting Han and Meiyang Jin performed the experiments associated with this study, for which a manuscript is in preparation.

muscle tissue causes glycogen storage disease VII, a disorder in which muscle cells cannot catabolize glycogen, thus hindering muscle cell function (Nichols et al., 1996; Vasconcelos et al., 1995; Raben et al., 1995). In collaboration with Dan Klionsky's lab, we observed that in yeast, Pfk2p is normally localized diffusely throughout the cytoplasm (data not shown). When yeast cells are starved of oxygen (hypoxic conditions) for >six hours, Pfk2p localizes to a single foci within each yeast cell, reminiscent of stress granules. This amazing adaptation response to hypoxia begs the questions: How is Pfk2p aggregation accomplished? What is the survival advantage of aggregating all Pfk2p to a single granule? How does aggregation of glycolytic enzymes affect glycolysis? The answer to the first question might be linked to the RNA-binding ability of Pfk2p, which can be explored by identifying the direct mRNA target sites of Pfk2p. Using PAR-CLIP-seq, we identified 525 direct binding sites of Pfk2p on 395 genes *in vivo* (Fig. 6.3). Pfk2p preferentially binds 3'UTRs of genes enriched for GO terms relating to metabolic processes; in fact, Pfk2p binds 12 of 22 glycolytic enzyme-encoding mRNAs. Searching 3'UTR Pfk2p binding sites revealed an AU-rich sequence in all 165 sites, similar to AU-rich elements (AREs) that are required for regulating the stability of transcripts in RNP granules (Wilusz et al., 2001; Wilson and Brewer, 1999).

Many questions remain regarding the role of RNA-binding for the metabolic function of Pfk2p. What are the fates of the RNA targets bound by Pfk2p under hypoxic conditions? What is the molecular mechanism that drives Pfk2p's specificity for its RNA substrates? Does hypoxia-induced granule formation depend on Pfk2p's ability to bind to RNA and/or its ID domain? How is glycolytic flux affected when Pfk2p is aggregated into a granule-like structure? How are other glycolytic enzymes affected by hypoxia or Pfk2p aggregation? Finally, are the mechanisms and components governing this phenomenon conserved in other eukaryotes? Answers to these questions will help define the molecular mechanisms governing Pfk2p (and other non-canonical RBPs) granule formation, and how RNP granule structures

might govern PTGR of the mRNAs targeted to these granules. Insights gained from this study could translate to studies of solid tumors where the local microenvironments are often hypoxic as the tumor grows and interior cells are located farther away from a blood supply. To adapt to hypoxia, tumor cells enhance glycolysis (Vander Heiden et al., 2009; Hanahan and Weinberg, 2011) given that the TCA cycle is inhibited by insufficient oxygen (reviewed in Chen et al., 2015). This phenomenon, called the Warburg effect, raises the question of how upregulated glycolysis is achieved selectively in tumor cells. An intriguing hypothesis is that components required for glycolysis - such as glycolytic enzymes and nucleotide cofactors - are sequestered to subcellular regions creating a high local concentration of glycolysis components and thus promoting higher flux through the pathway and generating higher levels of ATP to compensate for the lack of the TCA cycle.

6.3 Characterizing tissue-specific, multi-system RBP:RNA interactomes

To date, experimental characterization of an RBPome in a complex, multi-tissue organism does not exist. RBPs can be expressed or function very differently across cell- or tissue-types, and differences in an RBP's role in one tissue versus another would be masked in whole-organism studies. RBPs specific for the germline include piRNA factors (*e.g.* PIWI and TUDOR proteins) and regulators of translation (*e.g.* DAZL-family and nanos zinc finger proteins) (Yeo, 2014). Dysregulation of genes encoding these RBPs can lead to defects in germline development and fertility (Ishizu et al., 2012; Thomson and Lin, 2009; Brook et al., 2009; Klattenhoff and Theurkauf, 2008). Additionally, dysregulation of brain-specific RBPs, including the NOVA-family of alternative splicing regulators and members of the RBFOX and ELAVL families, can lead to defects in brain development and neurological function (Li

et al., 2007). Murine Rbfox1, which contains an RRM RBD, is expressed in brain but also in heart and skeletal muscle tissue where it is required for specific splicing events of select muscle-specific target mRNAs (Jin et al., 2003). Given the developmental importance of these few examples of tissue-specific RBPs, it would be prudent to not only identify tissue-specific RBPomes in complex organisms, but also identify the tissue-specific RNA targets of these RBPs to create a foundation for establishing networks of RBP:RNA interactions governing tissue-specific PTGR.

Novel RBPs in nematode tissues. We have begun to explore tissue-specific RBPomes by identifying the *in vivo* mRBPs that function in *C. elegans* both in whole animals as well as in *glp-4* mutant animals that lack a germline, thus representing *C. elegans* somatic tissues. Preliminary results uncover 795 mRBPs in whole animals and 509 mRBPs in germline-less animals suggesting that there are a few hundred mRBPs specific to the *C. elegans* germline. Our results confirm many known germline-specific mRBPs. For example, MEX-5 and OMA-2, both of which contain CCCH-type zinc-finger RBDs, are expressed in oocytes and early embryos, and have vital RNA-binding-dependent roles in regulating soma/germline asymmetry in developing embryos (MEX-5, (Oldenbroek et al., 2012)) and oocyte maturation (OMA-2, (Kaymak and Ryder, 2013)). Future experiments to identify additional tissue-specific RBPomes would benefit from the *C. elegans* model system, which has fewer tissue types than vertebrates but clearly defined germline, embryonic, muscle, neuronal, and intestinal tissues characteristic of all metazoans.

Identification of tissue-specific RBPomes will be important for examining pathologically-relevant systems such as normal and solid tumor tissues in cancer patients and neuronal tissues in neuropathy patients. Dysregulation of PTGR pathways is one of the hallmarks of cancer progression and metastatic processes; therefore, characterization of the RBPs that might be misregulated and the RBP:RNA interactions that might be altered in this context

is a key step to uncovering mechanisms driving cancer progression. In addition to dysregulated PTGR pathways, errors resulting in aberrant and insoluble protein aggregates is a common observation across many neuropathies such as amyotrophic lateral sclerosis (ALS) (Furukawa et al., 2006) and Alzheimer’s disease (Waring and Rosenberg, 2008). Our identification of proteins with prionogenic properties among our novel yeast mRBPs points to a role for RNAs in the formation of protein aggregates, which highly resemble RNP granules. Future explorations in this area of RBP biology could lead to the development of therapies for cancer and other human disorders and diseases caused by aberrant or disrupted RNP complex formations.

6.4 Bioinformatic considerations for RBP-related sequencing data

While postulating on the future directions of the field of small RNAs and RBPs in the context of PTGR pathways is important, it is equally important to identify future directions for the computational aspects required for studying PTGR in the context of deep sequencing data analysis. In this final section, I will outline what I believe to be important key steps for the field of bioinformatics in approaching the biological questions presented above.

Improved alignment algorithms for PAR-CLIP-seq-derived sequencing data.

One of the major computational considerations for Chapters Three and Four of this dissertation was accounting for T-to-C mismatches in deep sequencing reads generated by the PAR-CLIP-seq and gPAR-CLIP-seq protocols. While accounting for these mismatches is straightforward, problems can arise when considering whether a read has better alignment at a locus where it aligns with 1 non-T-to-C mismatch or at a different locus where it aligns with 1 T-to-C mismatch. Computationally, there should be no preference for ranking one of these alignments over the other as long as the nucleotides at both mismatch positions

in the read are of comparable quality. Technically, however, we expect to observe a T-to-C mismatch with much higher probability than any other mismatch type. In fact, we and other groups have observed 1 T-to-C event in >70% of reads (Freeberg et al., 2013; Hafner et al., 2010b), supporting choosing the T-to-C-containing alignment over other alignments.

One possible solution to this problem is to alter read alignment algorithms to weight T-to-C or other photoactivatable ribonucleoside analog-induced transition events (*e.g.* 6-thioguanosine produces G-to-A mismatches in aligned reads) differently. This approach could be modeled after how mismatches that occur at the 3' end of an Illumina read (where sequence quality is lower) is given less influence to the overall mappability of the read compared to a mismatch at the 5' end of a read (where sequence quality is higher). Rates of 4sU incorporation into nascent transcripts vary (we estimated the rate to be around 1 in 40 Us in yeast) but can be modified based on the concentration of the nucleotide analog introduced into the culture medium. Exact incorporation rates can be measured biochemically for a specific experiment and could be used as a parameter to the aligner. Alternatively, a post-alignment tool could be developed that takes a BAM or SAM alignment file and converts alignment quality scores according to ribonucleoside incorporation rates. This approach would require reporting all alignments in the BAM/SAM output file in case quality correction alters the ranking of best alignments. PAR-CLIP-based methodologies are becoming more popular as the choice for identification of direct RBP targets; therefore, accommodating the particularities of reads resulting from this methodology requires further exploration and attention.

Global RBP:RNA networks. Given that tagged versions of nearly all proteins in yeast are available, it is now feasible to complete large-scale studies to identify RNA target sites of almost the entire RBPome using high-throughput configurations of PAR-CLIP-seq. Even with multiplexing up to 16 libraries in a single Illumina HiSeq flow cell lane, coverage of

RBP targets will be sufficiently high given that an RBP only directly interacts with a small percentage of the entire transcriptome. After computationally identifying with nucleotide-resolution the sites of RBP:RNA interactions, RBP substrate lists can be used to generate interaction networks and predict molecular functions for uncharacterized RBPs based on similar characteristics of their RNA targets with known RBPs. Building RBP:RNA networks could be accomplished by constructing post-transcriptional regulatory logic networks using techniques such as machine learning and mutual information. The Saeed Tavazoie lab has developed methods for elucidating transcriptional regulatory logic rules from profiling genome-wide protein:DNA interactions (Goodarzi et al., 2012). These methods could be used as a model for constructing informative PTGR logic networks from transcriptome-wide protein:RNA interactions under a variety of intracellular and extracellular conditions.

We have performed a pilot experiment of massively parallel PAR-CLIP-seq to identify the targets of 43 yeast RBPs including 26 known and 17 novel mRBPs. While investigation of these datasets is ongoing, we observed a curious phenomenon in which some ($\sim 1,000$) regions of the transcriptome appear to be direct targets of multiple and seemingly unrelated RBPs at a frequency higher than expected by chance. This observation intriguingly parallels results by the ENCODE and modENCODE groups that some regions of the human, worm, and fly genomes are bound by multiple transcription factors at frequencies higher than expected by chance (modENCODE Consortium et al., 2010; Kvon et al., 2012; Nègre et al., 2011; Gerstein et al., 2010; Moorman et al., 2006). These high-occupancy target (HOT) DNA regions appear to have biological roles, suggesting that our high-occupancy targets on the transcriptome (HOT RNAs) could also be functional, perhaps by serving as general scaffolding components for multiple RBP:RNA interactions. Alternatively, HOT RNAs could potentially be artifacts of the PAR-CLIP-seq protocol representing inherent noise or non-specificity in the system.

Making the distinction between biologically meaningful RBP:RNA binding and noise is important for future studies employing PAR-CLIP-seq methods; however, proving or disproving functional significance for HOT RNAs required additional experimental study. Are HOT RNA regions enriched at specific genic regions? Do they contain subtle primary sequence or secondary structural motifs? Are HOT RNA regions conserved? Are there mRBPs that tend to bind more often to HOT RNA regions than others? Are HOT RNA regions typically bound by the same mRBPs? What is the biological consequence of disrupting HOT RNA regions or the RBPs that bind them? Answering some of these questions could help determine if HOT RNAs are biologically meaningful features or if they are just artifacts. If HOT RNAs are not real, results of studies addressing these questions would provide valuable information about noise in CLIP-seq data and potentially aid in filtering frequently captured RNA sequences from future CLIP-seq experiments, which would be designated as NOT RNAs for noisy occupancy of targets. If HOT RNAs are real, results of future studies could be used to augment RBP:RNA interaction networks and classify RBPs as either distinct or general RNA binders.

The ultimate goal of global RBP:RNA studies would be a rich database of *cis*-regulatory elements located on every noncoding RNA and protein-coding mRNA, the RBPs capable of interacting with each element, the strength or probability of each RBP:RNA interaction under a variety of physiological states (*e.g.* tissue type, cell cycle stage) or environmental conditions (*e.g.* normal growth, nutrient starvation, drug treatment), and functional outcomes of the RBP:RNA interactions from high-throughput assays measuring RNA stability, localization, and translational efficiency. This database could be populated with experimentally validated data where possible but also with computationally predicted data based on features like RBP recognition motifs. Additional data could be supplied to indicate whether multiple RBPs binding to the same target are antagonistic, synergistic, or redundant. Once

developed, this database could be probed in a variety of ways to develop testable hypotheses and elucidate general trends. As one example, measures of RBP expression changes (either directly by protein abundance or indirectly by gene expression) could be obtained for a given sample under a set of conditions and uploaded into the RBP:RNA network database, which would determine the set of RBP:RNA interactions, and thus PTGR mechanisms, affected by the change in RBP expression. This type of system could be incorporated into a pre-existing data analysis platform such as Cytoscape, an open source platform for data integration, network building, and analysis.

Although based on tenets of the Central Dogma, our understanding of post-transcriptional gene regulatory processes is continuously in flux as new evidence comes to light to either support or disprove current corollaries. Seemingly isolated observations can quickly become the rule rather than the exception - as in the case of microtubule RBPs - as more targeted experiments are performed. Proteins previously thought to only function in non-PTGR-related pathways, for example Pfk2p, are now being observed to take on RNA-binding-dependent PTGR roles that are intimately linked to and can feed back into their original functions. The importance of proposing and subsequently establishing genetic and physical interactions with post-transcriptional gene regulation across all other cellular processes cannot be understated: these interactions are foundational for maintaining life. Furthermore, questions of how these interactions change and are regulated to maintain cellular homeostasis in the face of intracellular cues and extracellular conditions will no doubt propel future studies in post-transcriptional gene regulation.

Table 6.1: Examples of novel yeast RBPs and their top mRNA targets by PAR-CLIP-seq.

RBP	Function	Top RNA targets
Pfk2p	β subunit of heterooctameric phosphofructokinase involved in glycolysis	Glycolytic enzymes (<i>e.g. ADH1</i>)
Sup35p	Translation termination factor whose altered protein conformation creates the [PSI(+)] prion that alters translational fidelity	Translation elongation factor (<i>e.g. TEF1</i>)
Rsc3p	Component of the RSC chromatin remodeling complex that is required for maintenance of proper ploidy and regulation of ribosomal protein genes and the cell wall/stress response	Translation components, stress response components (<i>e.g. TMA19</i>)
Apl5p	Subunit of clathrin associated protein complex that functions in transport of alkaline phosphatase to vacuole	ATPases (<i>e.g. KAR2</i>)
Imh1p	Mediates vesicular transport between an endosomal compartment and the Golgi	Membrane-localized proteins (<i>e.g. RGT2</i>)
Ecm25p	Protein of unknown function; contains GTPase-activator protein domain for Rho-like GTPases	Top target is <i>ASC1</i> , the β subunit of yeast G-proteins which binds to Gpa1p (α subunit with GTPase activity)

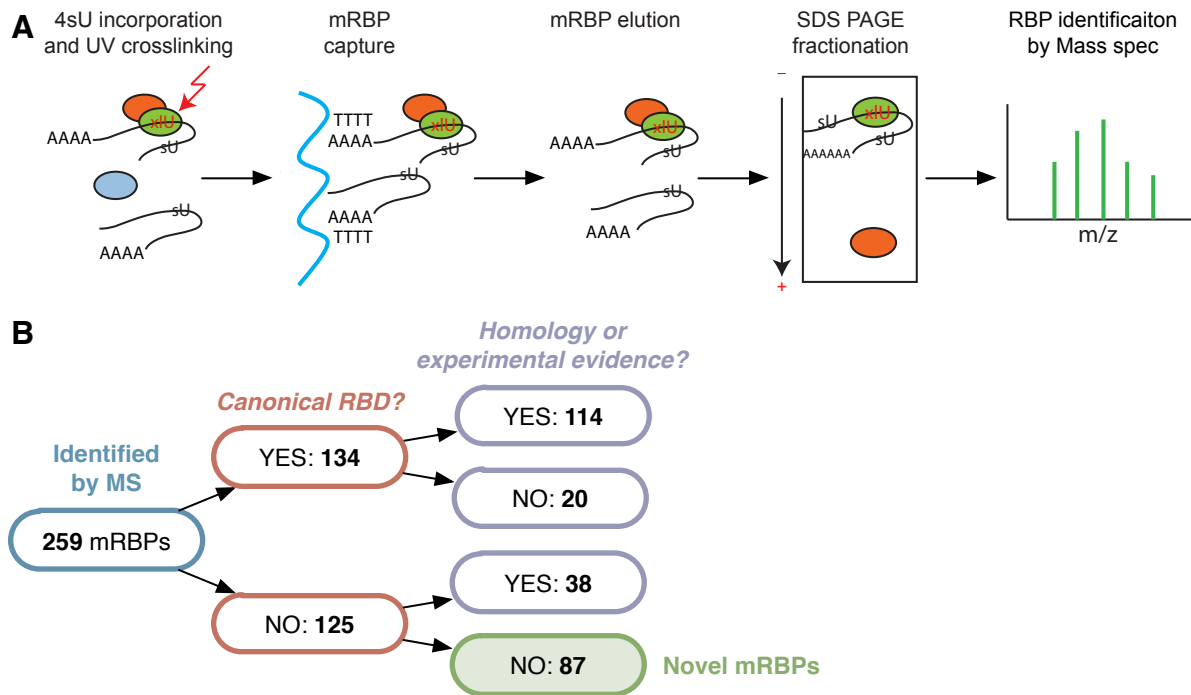


Figure 6.1: (A) PAR-CL-MS protocol defines *in vivo* mRBPs in yeast. (B) 259 yeast mRBPs can be classified by presence of a canonical RBD (Pfam/SMART) and by identification in additional yeast studies or as homologs of RBPs identified in human cell lines.

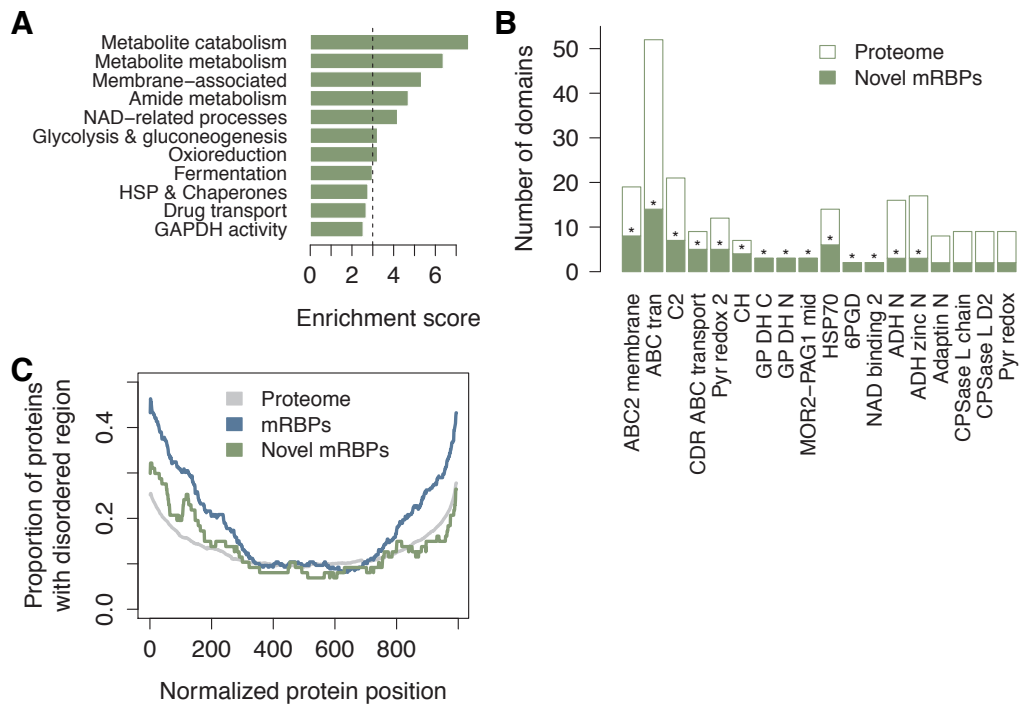


Figure 6.2: (A) Enriched GO terms were calculated among novel mRBPs using the topGO R Bioconductor package (Alexa et al., 2006). Enrichment score is the $-\log_{10}$ of the Bonferroni-adjusted p -value. (B) Enriched Pfam protein domains were calculated among novel mRBPs using a Fisher's exact test. $*p < 0.01$. (C) Amino acid disordered probabilities were calculated using IUPred (Dosztányi et al., 2005a,b). Residues were considered intrinsically disordered if they had a disorder probability > 0.5 compared to the proteome.

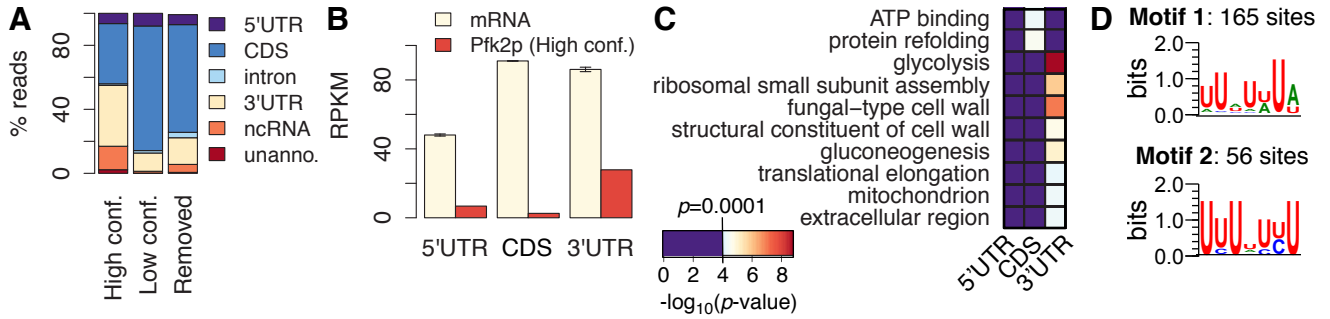


Figure 6.3: (A) Percent of total Pfk2p PAR-CLIP-seq reads mapping to indicated genic regions. (B) Number of reads normalized to total length of the given region in kb from high-confidence binding sites. (C) Enriched GO terms among genes containing at least one high-confidence Pfk2p site among different genic regions. p -values from Fisher's exact tests. (D) Identified sequence motifs among 3'UTR Pfk2p binding sites.

REFERENCES

- Alexa A., Rahnenführer J. and Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–7, 2006.
- Anders S. and Huber W. Differential expression analysis for sequence count data. *Gen Biol*, 11(10):R106, 2010.
- Anderson P. and Kedersha N. RNA granules: post-transcriptional and epigenetic modulators of gene expression. *Nat Rev Mol Cell Biol*, 10(6):430–6, 2009.
- Aoki K., Adachi S., Homoto M., Kusano H., Koike K. et al. LARP1 specifically recognizes the 3' terminus of poly(A) mRNA. *FEBS Lett*, 587(14):2173–8, 2013.
- Aravin A., Gaidatzis D., Pfeffer S., Lagos-Quintana M., Landgraf P. et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099):203–207, 2006.
- Aravin A. A., Hannon G. J. and Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–4, 2007.
- Archer S. K., Luu V.-D., de Queiroz R. A., Brems S. and Clayton C. Trypanosoma brucei PUF9 regulates mRNAs for proteins involved in replicative processes over the cell cycle. *PLoS Pathog*, 5(8):e1000565, 2009.
- Ariz M., Mainpal R. and Subramaniam K. C. elegans RNA-binding proteins PUF-8 and MEX-3 function redundantly to promote germline stem cell mitosis. *Dev Biol*, 326(2):295–304, 2009.
- Aronov S., Aranda G., Behar L. and Ginzburg I. Visualization of translated tau protein in the axons of neuronal P19 cells and characterization of tau RNP granules. *J Cell Sci*, 115(Pt 19):3817–27, 2002.
- Arvand A. and Denny C. T. Biology of EWS/ETS fusions in Ewing's family tumors. *Oncogene*, 20(40):5747–54, 2001.
- Ashburner M., Ball C. A., Blake J. A., Botstein D., Butler H. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9, 2000.

- Ashe A., Sapetschnig A., Weick E. M., Mitchell J., Bagijn M. P. et al. piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. *Cell*, 150(1):88–99, 2012.
- Ashe M. P., De Long S. K. and Sachs A. B. Glucose depletion rapidly inhibits translation initiation in yeast. *Mol Biol Cell*, 11(3):833–48, 2000.
- Auweter S. D., Oberstrass F. C. and Allain F. H.-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res*, 34(17):4943–59, 2006.
- Axtell M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA*, 19(6):740–51, 2013.
- Baejen C., Torkler P., Gressel S., Essig K., Söding J. et al. Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol Cell*, 55(5):745–57, 2014.
- Baek D., Villén J., Shin C., Camargo F. D., Gygi S. P. et al. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.
- Bagijn M. P., Goldstein L. D., Sapetschnig A., Weick E. M., Bouasker S. et al. Function, Targets, and Evolution of *Caenorhabditis elegans* piRNAs. *Science*, 2012.
- Bailey T. L., Boden M., Buske F. A., Frith M., Grant C. E. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res*, 37(Web Server issue):W202–8, 2009.
- Bailey T. L. and Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2:28–36, 1994.
- Bailey T. L. and Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14(1):48–54, 1998.
- Balagopal V. and Parker R. Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Curr Opin Cell Biol*, 21(3):403–408, 2009.
- Baltz A. G., Munschauer M., Schwanhäusser B., Vasile A., Murakawa Y. et al. The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, 46(5):674–90, 2012.
- Barreau C., Paillard L. and Osborne H. B. AU-rich elements and associated factors: are there unifying principles? *Nucleic Acids Res*, 33(22):7138–50, 2005.
- Barret A., Rawlings N. D. and Woessner J. F. *Handbook of Proteolytic Enzymes*. San Diego: Academic Press, 1998.
- Barrett T., Troup D. B., Wilhite S. E., Ledoux P., Evangelista C. et al. NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*, 39(Database issue):D1005–10, 2011.

- Bartel D. P. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–33, 2009.
- Batista P. J., Ruby J. G., Claycomb J. M., Chiang R., Fahlgren N. et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol Cell*, 31(1):67–78, 2008.
- Bayfield M. A., Yang R. and Maraia R. J. Conserved and divergent features of the structure and function of La and La-related proteins (LARPs). *Biochim Biophys Acta*, 1799(5-6):365–78, 2010.
- Bazzini A. A., Lee M. T. and Giraldez A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science*, 336(6078):233–7, 2012.
- Berglund J. A., Chua K., Abovich N., Reed R. and Rosbash M. The splicing factor BBP interacts specifically with the pre-mRNA branchpoint sequence UACUAAC. *Cell*, 89(5):781–7, 1997.
- Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N. et al. The Protein Data Bank. *Nucleic Acids Res*, 28(1):235–42, 2000.
- Bernhart S. H., Hofacker I. L. and Stadler P. F. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–5, 2006.
- Bessereau J. L. Transposons in *C. elegans*. *WormBook : the online review of C. elegans biology*, pages 1–13, 2006.
- Biasini M., Bienert S., Waterhouse A., Arnold K., Studer G. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, 42(Web Server issue):W252–8, 2014.
- Billi A. C., Alessi A. F., Khivansara V., Han T., Freeberg M. et al. The *Caenorhabditis elegans* HEN1 Ortholog, HENN-1, Methylates and Stabilizes Select Subclasses of Germline Small RNAs. *PLoS Genet*, 8(4):e1002617, 2012a.
- Billi A. C., Freeberg M. A. and Kim J. K. piRNAs and siRNAs collaborate in *Caenorhabditis elegans* genome defense. *Gen Biol*, 13(7):164, 2012b.
- Blankenberg D., Von Kuster G., Coraor N., Ananda G., Lazarus R. et al. Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol*, Chapter 19:Unit 19.10.1–21, 2010.
- Blower M. D. Molecular insights into intracellular RNA localization. *Int Rev Cell Mol Biol*, 302:1–39, 2013.
- Blower M. D., Feric E., Weis K. and Heald R. Genome-wide analysis demonstrates conserved localization of messenger RNAs to mitotic microtubules. *J Cell Biol*, 179(7):1365–73, 2007.

- Bousquet-Antonelli C. and Deragon J.-M. A comprehensive analysis of the La-motif protein superfamily. *RNA*, 15(5):750–64, 2009.
- Brasch-Andersen C., Tan Q., Børghlum A. D., Haagerup A., Larsen T. R. et al. Significant linkage to chromosome 12q24.32-q24.33 and identification of SFRS8 as a possible asthma susceptibility gene. *Thorax*, 61(10):874–9, 2006.
- Brauer M. J., Saldanha A. J., Dolinski K. and Botstein D. Homeostatic adjustment and metabolic remodeling in glucose-limited yeast cultures. *Mol Biol Cell*, 16(5):2503–17, 2005.
- Brennecke J., Aravin A. A., Stark A., Dus M., Kellis M. et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell*, 128(6):1089–1103, 2007.
- Brennecke J., Malone C. D., Aravin A. A., Sachidanandam R., Stark A. et al. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, 322(5906):1387–1392, 2008.
- Brook M., Smith J. W. S. and Gray N. K. The DAZL and PABP families: RNA-binding proteins with interrelated roles in translational control in oocytes. *Reproduction*, 137(4):595–617, 2009.
- Brown L. Y. and Brown S. A. Alanine tracts: the expanding story of human illness and trinucleotide repeats. *Trends Genet*, 20(1):51–8, 2004.
- Brümmer A., Kishore S., Subasic D., Hengartner M. and Zavolan M. Modeling the binding specificity of the RNA-binding protein GLD-1 suggests a function of coding region-located sites in translational repression. *RNA*, 19(10):1317–26, 2013.
- Buchan J. R. and Parker R. Eukaryotic stress granules: the ins and outs of translation. *Mol Cell*, 36(6):932–41, 2009.
- Buckanovich R. J., Yang Y. Y. and Darnell R. B. The onconeural antigen Nova-1 is a neuron-specific RNA-binding protein, the activity of which is inhibited by paraneoplastic antibodies. *J Neurosci*, 16(3):1114–22, 1996.
- Burrows C., Abd Latip N., Lam S.-J., Carpenter L., Sawicka K. et al. The RNA binding protein Larp1 regulates cell division, apoptosis and cell migration. *Nucleic Acids Res*, 38(16):5542–53, 2010.
- Byrne K. P. and Wolfe K. H. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. *Genome Res*, 15(10):1456–61, 2005.
- Campbell Z. T., Bhimsaria D., Valley C. T., Rodriguez-Martinez J. A., Menichelli E. et al. Cooperativity in RNA-protein interactions: global analysis of RNA binding specificity. *Cell Rep*, 1(5):570–81, 2012.

- Campbell Z. T., Valley C. T. and Wickens M. A protein-RNA specificity code enables targeted activation of an endogenous human transcript. *Nat Struct Mol Biol*, 21(8):732–8, 2014.
- Carmell M. A., Girard A., van de Kant H. J., Bourc'his D., Bestor T. H. et al. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell*, 12(4):503–514, 2007.
- Carmody S. R. and Wente S. R. mRNA nuclear export at a glance. *J Cell Sci*, 122(Pt 12):1933–7, 2009.
- Castello A., Fischer B., Eichelbaum K., Horos R., Beckmann B. M. et al. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–406, 2012.
- Cecere G., Zheng G. X., Mansisidor A. R., Klymko K. E. and Grishok A. Promoters recognized by forkhead proteins exist for individual 21U-RNAs. *Mol Cell*, 47(5):734–745, 2012.
- Celniker S. E., Dillon L. A., Gerstein M. B., Gunsalus K. C., Henikoff S. et al. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, 2009.
- Chambeyron S., Popkova A., Payen-Groschene G., Brun C., Laouini D. et al. piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. *Proc Natl Acad Sci U S A*, 105(39):14964–14969, 2008.
- Chartrand P., Meng X. H., Singer R. H. and Long R. M. Structural elements required for the localization of ASH1 mRNA and of a green fluorescent protein reporter particle in vivo. *Curr Biol*, 9(6):333–6, 1999.
- Chen K. and Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet*, 8(2):93–103, 2007.
- Chen P.-S., Su J.-L. and Hung M.-C. Dysregulation of microRNAs in cancer. *J Biomed Sci*, 19:90, 2012.
- Chen W. and Moore M. J. The spliceosome: disorder and dynamics defined. *Curr Opin Struct Biol*, 24:141–9, 2014.
- Chen X., Qian Y. and Wu S. The Warburg effect: Evolving interpretations of an established concept. *Free Radic Biol Med*, 79C:253–263, 2015.
- Chi S. W., Zang J. B., Mele A. and Darnell R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86, 2009.
- Chou C.-H., Lin F.-M., Chou M.-T., Hsu S.-D., Chang T.-H. et al. A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing. *BMC Genomics*, 14 Suppl 1:S2, 2013.

- Chu Y. and Corey D. R. RNA sequencing: platform selection, experimental design, and data interpretation. *Nucleic Acid Ther*, 22(4):271–4, 2012.
- Cieřła J. Metabolic enzymes that bind RNA: yet another level of cellular regulatory network? *Acta Biochim Pol*, 53(1):11–32, 2006.
- Ciosk R., DePalma M. and Priess J. R. Translational regulators maintain totipotency in the *Caenorhabditis elegans* germline. *Science*, 311(5762):851–3, 2006.
- Coletta A., Pinney J. W., Solís D. Y. W., Marsh J., Pettifer S. R. et al. Low-complexity regions within protein sequences have position-dependent roles. *BMC Syst Biol*, 4:43, 2010.
- Comoglio F., Sievers C. and Paro R. Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, 16(1):32, 2015.
- Conine C. C., Batista P. J., Gu W., Claycomb J. M., Chaves D. A. et al. Argonautes ALG-3 and ALG-4 are required for spermatogenesis-specific 26G-RNAs and thermotolerant sperm in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*, 107(8):3588–3593, 2010.
- Cooper T. A., Wan L. and Dreyfuss G. RNA and disease. *Cell*, 136(4):777–93, 2009.
- Corcoran D. L., Georgiev S., Mukherjee N., Gottwein E., Skalsky R. L. et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Gen Biol*, 12(8):R79, 2011.
- Cordero F., Beccuti M., Arigoni M., Donatelli S. and Calogero R. A. Optimizing a massive parallel sequencing workflow for quantitative miRNA expression analysis. *PLoS One*, 7(2):e31630, 2012.
- Core L. J., Waterfall J. J. and Lis J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322(5909):1845–8, 2008.
- Cox D. N., Chao A., Baker J., Chang L., Qiao D. et al. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Gen Dev*, 12(23):3715–3727, 1998.
- Creamer T. J., Darby M. M., Jamonnak N., Schaughency P., Hao H. et al. Transcriptome-wide binding sites for components of the *Saccharomyces cerevisiae* non-poly(A) termination pathway: Nrd1, Nab3, and Sen1. *PLoS Genet*, 7(10):e1002329, 2011.
- Crooks G. E., Hon G., Chandonia J.-M. and Brenner S. E. WebLogo: a sequence logo generator. *Gen Res*, 14(6):1188–1190, 2004.
- Darnell J. C., Mostovetsky O. and Darnell R. B. FMRP RNA targets: identification and validation. *Genes Brain Behav*, 4(6):341–9, 2005.

- Darnell J. C., Van Driesche S. J., Zhang C., Hung K. Y. S., Mele A. et al. FMRP stalls ribosomal translocation on mRNAs linked to synaptic function and autism. *Cell*, 146(2):247–61, 2011.
- Darnell R. B. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA*, 1(2):266–86, 2010a.
- Darnell R. B. RNA regulation in neurologic disease and cancer. *Cancer Res Treat*, 42(3):125–9, 2010b.
- Darty K., Denise A. and Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–5, 2009.
- Das P. P., Bagijn M. P., Goldstein L. D., Woolford J. R., Lehrbach N. J. et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. *Mol Cell*, 31(1):79–90, 2008.
- Datla U. S., Scovill N. C., Brokamp A. J., Kim E., Asch A. S. et al. Role of PUF-8/PUF protein in stem cell control, sperm-oocyte decision and cell fate reprogramming. *J Cell Physiol*, 229(10):1306–11, 2014.
- Davis C. A. and Ares, Jr M. Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*, 103(9):3262–7, 2006.
- de Haro M., Al-Ramahi I., Jones K. R., Holth J. K., Timchenko L. T. et al. Smaug/SAMD4A restores translational activity of CUGBP1 and suppresses CUG-induced myopathy. *PLoS Genet*, 9(4):e1003445, 2013.
- de Wit E., Linsen S. E., Cuppen E. and Berezikov E. Repertoire and evolution of miRNA genes in four divergent nematode species. *Gen Res*, 19(11):2064–2074, 2009.
- Deng Y., Singer R. H. and Gu W. Translation of ASH1 mRNA is repressed by Puf6p-Fun12p/eIF5B interaction and released by CK2 phosphorylation. *Genes Dev*, 22(8):1037–50, 2008.
- DePristo M. A., Banks E., Poplin R., Garimella K. V., Maguire J. R. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–8, 2011.
- Dey S., Tameire F. and Koumenis C. PERK-ing up autophagy during MYC-induced tumorigenesis. *Autophagy*, 9(4):612–4, 2013.
- Dezso Z., Nikolsky Y., Sviridov E., Shi W., Serebriyskaya T. et al. A comprehensive functional analysis of tissue specificity of human gene expression. *BMC Biol*, 6:49, 2008.
- Dicthenberg J. B., Swanger S. A., Antar L. N., Singer R. H. and Bassell G. J. A direct role for FMRP in activity-dependent dendritic mRNA transport links filopodial-spine morphogenesis to fragile X syndrome. *Dev Cell*, 14(6):926–39, 2008.

- Dieterich C. and Stadler P. F. Computational biology of RNA interactions. *Wiley Interdiscip Rev RNA*, 4(1):107–20, 2013.
- Dominissini D., Moshitch-Moshkovitz S., Schwartz S., Salmon-Divon M., Ungar L. et al. Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, 485(7397):201–6, 2012.
- Dosztányi Z., Csizmok V., Tompa P. and Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, 21(16):3433–4, 2005a.
- Dosztányi Z., Csizmók V., Tompa P. and Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*, 347(4):827–39, 2005b.
- Dreyfuss G., Adam S. A. and Choi Y. D. Physical change in cytoplasmic messenger ribonucleoproteins in cells treated with inhibitors of mRNA transcription. *Mol Cell Biol*, 4(3):415–23, 1984.
- Dreyfuss G., Kim V. N. and Kataoka N. Messenger-RNA-binding proteins and the messages they carry. *Nat Rev Mol Cell Biol*, 3(3):195–205, 2002.
- Du T.-G., Jellbauer S., Müller M., Schmid M., Niessing D. et al. Nuclear transit of the RNA-binding protein She2 is required for translational control of localized ASH1 mRNA. *EMBO Rep*, 9(8):781–7, 2008.
- Du T.-G., Schmid M. and Jansen R.-P. Why cells move messages: the biological functions of mRNA localization. *Semin Cell Dev Biol*, 18(2):171–7, 2007.
- Dubnau J., Chiang A.-S., Grady L., Barditch J., Gossweiler S. et al. The stauufen/pumilio pathway is involved in *Drosophila* long-term memory. *Curr Biol*, 13(4):286–96, 2003.
- Dyson H. J. and Wright P. E. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6(3):197–208, 2005.
- Ebert M. S. and Sharp P. A. Roles for microRNAs in conferring robustness to biological processes. *Cell*, 149(3):515–24, 2012.
- Edgar R., Domrachev M. and Lash A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–10, 2002.
- Edwards T. A., Pyle S. E., Wharton R. P. and Aggarwal A. K. Structure of Pumilio reveals similarity between RNA and peptide binding motifs. *Cell*, 105(2):281–9, 2001.
- Erhard F., Dölken L., Jaskiewicz L. and Zimmer R. PARma: identification of microRNA target sites in AGO-PAR-CLIP data. *Genome Biol*, 14(7):R79, 2013.

- Eulalio A., Behm-Ansmant I., Schweizer D. and Izaurralde E. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Mol Cell Biol*, 27(11):3970–81, 2007.
- Farazi T. A., Ten Hoeve J. J., Brown M., Mihailovic A., Horlings H. M. et al. Identification of distinct miRNA target regulation between breast cancer molecular subtypes using AGO2-PAR-CLIP and patient datasets. *Genome Biol*, 15(1):R9, 2014.
- Favre A., Moreno G., Salet C. and Vinzens F. 4-Thiouridine incorporation into the RNA of monkey kidney cells (CV-1) triggers near-UV light long-term inhibition of DNA, RNA and protein synthesis. *Photochem Photobiol*, 58(5):689–94, 1993.
- Fazlollahi M., Lee E., Muroff I., Lu X.-J., Gomez-Alcala P. et al. Harnessing natural sequence variation to dissect posttranscriptional regulatory networks in yeast. *G3 (Bethesda)*, 4(8):1539–53, 2014.
- Flicek P., Amode M. R., Barrell D., Beal K., Brent S. et al. Ensembl 2012. *Nucleic Acids Res*, 40(Database issue):D84–90, 2012.
- Foat B. C., Houshmandi S. S., Olivas W. M. and Bussemaker H. J. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. *Proc Natl Acad Sci U S A*, 102(49):17675–80, 2005.
- Freeberg M. A., Han T., Moresco J. J., Kong A., Yang Y.-C. et al. Pervasive and dynamic protein binding sites of the mRNA transcriptome in *Saccharomyces cerevisiae*. *Genome Biol*, 14(2):R13, 2013.
- Friedman R. C., Farh K. K.-H., Burge C. B. and Bartel D. P. Most mammalian mRNAs are conserved targets of microRNAs. *Gen Res*, 19(1):92–105, 2009.
- Frokjaer-Jensen C., Davis M. W., Hopkins C. E., Newman B. J., Thummel J. M. et al. Single-copy insertion of transgenes in *Caenorhabditis elegans*. *Nat Genet*, 40(11):1375–1383, 2008.
- Fu Y., Dominissini D., Rechavi G. and He C. Gene expression regulation mediated through reversible m⁶A RNA methylation. *Nat Rev Genet*, 15(5):293–306, 2014.
- Furukawa Y., Fu R., Deng H.-X., Siddique T. and O’Halloran T. V. Disulfide cross-linked protein represents a significant fraction of ALS-associated Cu, Zn-superoxide dismutase aggregates in spinal cords of model mice. *Proc Natl Acad Sci U S A*, 103(18):7148–53, 2006.
- Gadir N., Haim-Vilmovsky L., Kraut-Cohen J. and Gerst J. E. Localization of mRNAs coding for mitochondrial proteins in the yeast *Saccharomyces cerevisiae*. *RNA*, 17(8):1551–65, 2011.
- Gallego C., Garí E., Colomina N., Herrero E. and Aldea M. The Cln3 cyclin is down-regulated by translational repression and degradation during the G1 arrest caused by nitrogen deprivation in budding yeast. *EMBO J*, 16(23):7196–206, 1997.

- Gao G. and Dudley, Jr S. C. RBM25/LUC7L3 function in cardiac sodium channel splicing regulation of human heart failure. *Trends Cardiovasc Med*, 23(1):5–8, 2013.
- Garber M., Grabherr M. G., Guttman M. and Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods*, 8(6):469–77, 2011.
- Garneau N. L., Wilusz J. and Wilusz C. J. The highways and byways of mRNA decay. *Nat Rev Mol Cell Biol*, 8(2):113–26, 2007.
- Garrey S. M., Voelker R. and Berglund J. A. An extended RNA binding site for the yeast branch point-binding protein and the role of its zinc knuckle domains in RNA binding. *J Biol Chem*, 281(37):27443–53, 2006.
- Gasch A. P., Spellman P. T., Kao C. M., Carmel-Harel O., Eisen M. B. et al. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell*, 11(12):4241–57, 2000.
- Gatchel J. R. and Zoghbi H. Y. Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet*, 6(10):743–55, 2005.
- Gavin A.-C., Bösch M., Krause R., Grandi P., Marzioch M. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–7, 2002.
- Gent J. I., Lamm A. T., Pavelec D. M., Maniar J. M., Parameswaran P. et al. Distinct phases of siRNA synthesis in an endogenous RNAi pathway in *C. elegans* soma. *Mol Cell*, 37(5):679–689, 2010.
- Gent J. I., Schvarzstein M., Villeneuve A. M., Gu S. G., Jantsch V. et al. A *Caenorhabditis elegans* RNA-directed RNA polymerase in sperm development and endogenous RNA interference. *Genetics*, 183(4):1297–1314, 2009.
- Gerber A. P., Herschlag D. and Brown P. O. Extensive association of functionally and cytologically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol*, 2(3):E79, 2004.
- Gerber A. P., Luschnig S., Krasnow M. A., Brown P. O. and Herschlag D. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A*, 103(12):4487–92, 2006.
- German M. A., Luo S., Schroth G., Meyers B. C. and Green P. J. Construction of Parallel Analysis of RNA Ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. *Nat Protoc*, 4(3):356–62, 2009.
- Gerstberger S., Hafner M., Ascano M. and Tuschl T. Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Adv Exp Med Biol*, 825:1–55, 2014.

- Gerstein M. B., Lu Z. J., Van Nostrand E. L., Cheng C., Arshinoff B. I. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science*, 330(6012):1775–87, 2010.
- Ghaemmaghami S., Huh W.-K., Bower K., Howson R. W., Belle A. et al. Global analysis of protein expression in yeast. *Nature*, 425(6959):737–41, 2003.
- Giardine B., Riemer C., Hardison R. C., Burhans R., Elnitski L. et al. Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15(10):1451–5, 2005.
- Gingold H. and Pilpel Y. Determinants of translation efficiency and accuracy. *Mol Syst Biol*, 7:481, 2011.
- Girard A., Sachidanandam R., Hannon G. J. and Carmell M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099):199–202, 2006.
- Glisovic T., Bachorik J. L., Yong J. and Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett*, 582(14):1977–86, 2008.
- Goecks J., Nekrutenko A., Taylor J. and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.
- Goh W.-S. S., Seah J. W. E., Harrison E. J., Chen C., Hammell C. M. et al. A genome-wide RNAi screen identifies factors required for distinct stages of *C. elegans* piRNA biogenesis. *Gen Dev*, 28(7):797–807, 2014.
- Goldstrohm A. C., Hook B. A., Seay D. J. and Wickens M. PUF proteins bind Pop2p to regulate messenger RNAs. *Nat Struct Mol Biol*, 13(6):533–9, 2006.
- Goldstrohm A. C., Seay D. J., Hook B. A. and Wickens M. PUF protein-mediated deadenylation is catalyzed by Ccr4p. *J Biol Chem*, 282(1):109–14, 2007.
- Gonzalez I., Buonomo S. B., Nasmyth K. and von Ahsen U. ASH1 mRNA localization in yeast involves multiple secondary structural elements and Ash1 protein translation. *Curr Biol*, 9(6):337–40, 1999.
- Goodarzi H., Najafabadi H. S., Oikonomou P., Greco T. M., Fish L. et al. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature*, 485(7397):264–8, 2012.
- Gottwein E., Corcoran D. L., Mukherjee N., Skalsky R. L., Hafner M. et al. Viral microRNA targetome of KSHV-infected primary effusion lymphoma cell lines. *Cell Host Microbe*, 10(5):515–26, 2011.
- Green N. M. Avidin and streptavidin. *Methods Enzymol*, 184:51–67, 1990.
- Greenberg J. R. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res*, 6(2):715–32, 1979.

- Greiner-Stöffele T., Foerster H. H. and Hahn U. Ribonuclease T1 cleaves RNA after guanosines within single-stranded gaps of any length. *Nucleosides Nucleotides Nucleic Acids*, 19(7):1101–9, 2000.
- Grentzinger T., Armenise C., Brun C., Mugat B., Serrano V. et al. piRNA-mediated transgenerational inheritance of an acquired trait. *Gen Res*, 2012.
- Griac P. and Henry S. A. The yeast inositol-sensitive upstream activating sequence, UASINO, responds to nitrogen availability. *Nucleic Acids Res*, 27(9):2043–50, 1999.
- Grigull J., Mnaimneh S., Pootoolal J., Robinson M. D. and Hughes T. R. Genome-wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Mol Cell Biol*, 24(12):5534–47, 2004.
- Grimson A., Srivastava M., Fahey B., Woodcroft B. J., Chiang H. R. et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, 455(7217):1193–1197, 2008.
- Grousl T., Ivanov P., Frýdlová I., Vasicová P., Janda F. et al. Robust heat shock induces eIF2 α -phosphorylation-independent assembly of stress granules containing eIF3 and 40S ribosomal subunits in budding yeast, *Saccharomyces cerevisiae*. *J Cell Sci*, 122(Pt 12):2078–88, 2009.
- Gu S. G., Pak J., Guang S., Maniar J. M., Kennedy S. et al. Amplification of siRNA in *Caenorhabditis elegans* generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint. *Nat Genet*, 44(2):157–164, 2012.
- Gu W., Deng Y., Zenklusen D. and Singer R. H. A new yeast PUF family protein, Puf6p, represses ASH1 mRNA translation and is required for its localization. *Genes Dev*, 18(12):1452–65, 2004.
- Gu W., Shirayama M., Conte D., Vasale J., Batista P. J. et al. Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the *C. elegans* germline. *Mol Cell*, 36(2):231–244, 2009.
- Guipaud O., Guillonnet F., Labas V., Praseuth D., Rossier J. et al. An in vitro enzymatic assay coupled to proteomics analysis reveals a new DNA processing activity for Ewing sarcoma and TAF(II)68 proteins. *Proteomics*, 6(22):5962–72, 2006.
- Gunawardane L. S., Saito K., Nishida K. M., Miyoshi K., Kawamura Y. et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science*, 315(5818):1587–1590, 2007.
- Guo H., Ingolia N. T., Weissman J. S. and Bartel D. P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, 466(7308):835–40, 2010.

- Hackenberg M., Sturm M., Langenberger D., Falcón-Pérez J. M. et al. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*, 37(Web Server issue):W68–76, 2009.
- Hafidh S., Capková V. and Honys D. Safe keeping the message: mRNP complexes tweaking after transcription. *Adv Exp Med Biol*, 722:118–36, 2011.
- Hafner M., Landthaler M., Burger L., Khorshid M., Hausser J. et al. PAR-Clip—a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J Vis Exp*, 41, 2010a.
- Hafner M., Landthaler M., Burger L., Khorshid M., Hausser J. et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–41, 2010b.
- Hall T. M. T. Expanding the RNA-recognition code of PUF proteins. *Nat Struct Mol Biol*, 21(8):653–5, 2014.
- Han T., Manoharan A. P., Harkins T. T., Bouffard P., Fitzpatrick C. et al. 26G endo-siRNAs regulate spermatogenic and zygotic gene expression in *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A*, 106(44):18674–18679, 2009.
- Han T. W., Kato M., Xie S., Wu L. C., Mirzaei H. et al. Cell-free formation of RNA granules: bound RNAs identify features and components of cellular assemblies. *Cell*, 149(4):768–79, 2012.
- Hanahan D. and Weinberg R. A. Hallmarks of cancer: the next generation. *Cell*, 144(5):646–74, 2011.
- Hardcastle T. J. and Kelly K. A. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.
- Hashimoto T. B., Edwards M. D. and Gifford D. K. Universal count correction for high-throughput sequencing. *PLoS Comput Biol*, 10(3):e1003494, 2014.
- Hausser J. and Zavolan M. Identification and consequences of miRNA-target interactions—beyond repression of gene expression. *Nat Rev Genet*, 15(9):599–612, 2014.
- Heiss N. S., Knight S. W., Vulliamy T. J., Klauck S. M., Wiemann S. et al. X-linked dyskeratosis congenita is caused by mutations in a highly conserved gene with putative nucleolar functions. *Nat Genet*, 19(1):32–8, 1998.
- Hendrickson D. G., Hogan D. J., McCullough H. L., Myers J. W., Herschlag D. et al. Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA. *PLoS Biol*, 7(11):e1000238, 2009.
- Hentze M. W. Enzymes as RNA-binding proteins: a role for (di)nucleotide-binding domains? *Trends Biochem Sci*, 19(3):101–3, 1994.

- Hentze M. W. and Preiss T. The REM phase of gene regulation. *Trends Biochem Sci*, 35(8):423–6, 2010.
- Heyn H., Ferreira H. J., Bassas L., Bonache S., Sayols S. et al. Epigenetic disruption of the PIWI pathway in human spermatogenic disorders. *PLoS One*, 7(10):e47892, 2012.
- Hillier L. W., Reinke V., Green P., Hirst M., Marra M. A. et al. Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*. *Gen Res*, 19(4):657–666, 2009.
- Hinnebusch A. G. Translational regulation of GCN4 and the general amino acid control of yeast. *Annu Rev Microbiol*, 59:407–50, 2005.
- Hirschfeld M., Zhang B., Jaeger M., Stamm S., Erbes T. et al. Hypoxia-dependent mRNA expression pattern of splicing factor YT521 and its impact on oncological important target gene expression. *Mol Carcinog*, 53(11):883–92, 2014.
- Ho J. J. D. and Marsden P. A. Competition and collaboration between RNA-binding proteins and microRNAs. *Wiley Interdiscip Rev RNA*, 5(1):69–86, 2014.
- Hofacker I., Fontana W., Stadler P., Bonhoeffer L., Tacker M. et al. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie - Chemical Monthly*, 125(2):167–188, 1994. ISSN 0026-9247.
- Hogan D. J., Riordan D. P., Gerber A. P., Herschlag D. and Brown P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*, 6(10):e255, 2008.
- Hohmann S. and Mager W. H. *Yeast stress responses*. Springer, Berlin, 2003. ISBN 3540439269 (hardcover : alk. paper).
- Hook B. A., Goldstrohm A. C., Seay D. J. and Wickens M. Two yeast PUF proteins negatively regulate a single mRNA. *J Biol Chem*, 282(21):15430–8, 2007.
- Houshmandi S. S. and Olivás W. M. Yeast Puf3 mutants reveal the complexity of Puf-RNA binding and identify a loop required for regulation of mRNA decay. *RNA*, 11(11):1655–66, 2005.
- Houwing S., Kamminga L. M., Berezikov E., Cronembold D., Girard A. et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, 129(1):69–82, 2007.
- Huang D. W., Sherman B. T. and Lempicki R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, 2009a.
- Huang D. W., Sherman B. T. and Lempicki R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009b.

- Huh W.-K., Falvo J. V., Gerke L. C., Carroll A. S., Howson R. W. et al. Global analysis of protein localization in budding yeast. *Nature*, 425(6959):686–91, 2003.
- Huisinga K. L., Brower-Toland B. and Elgin S. C. R. The contradictory definitions of heterochromatin: transcription and silencing. *Chromosoma*, 115(2):110–22, 2006.
- Huo Y., Iadevaia V. and Proud C. G. Differing effects of rapamycin and mTOR kinase inhibitors on protein synthesis. *Biochem Soc Trans*, 39(2):446–50, 2011.
- Incarnato D., Neri F., Diamanti D. and Oliviero S. MREdictor: a two-step dynamic interaction model that accounts for mRNA accessibility and Pumilio binding accurately predicts microRNA targets. *Nucleic Acids Res*, 41(18):8421–33, 2013.
- Ingolia N. T., Ghaemmaghami S., Newman J. R. S. and Weissman J. S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23, 2009.
- Ipsaro J. J., Haase A. D., Knott S. R., Joshua-Tor L. and Hannon G. J. The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. *Nature*, 491(7423):279–83, 2012.
- Ishizu H., Siomi H. and Siomi M. C. Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Gen Dev*, 26(21):2361–73, 2012.
- Ivanov P. A., Chudinova E. M. and Nadezhdina E. S. Disruption of microtubules inhibits cytoplasmic ribonucleoprotein stress granule formation. *Exp Cell Res*, 290(2):227–33, 2003.
- Jackson, Jr J. S., Houshmandi S. S., Lopez Leban F. and Olivas W. M. Recruitment of the Puf3 protein to its mRNA target for regulation of mRNA decay in yeast. *RNA*, 10(10):1625–36, 2004.
- Jafarifar F., Yao P., Eswarappa S. M. and Fox P. L. Repression of VEGFA by CA-rich element-binding microRNAs is modulated by hnRNP L. *EMBO J*, 30(7):1324–34, 2011.
- Jamonnak N., Creamer T. J., Darby M. M., Schaugency P., Wheelan S. J. et al. Yeast Nrd1, Nab3, and Sen1 transcriptome-wide binding maps suggest multiple roles in post-transcriptional RNA processing. *RNA*, 17(11):2011–25, 2011.
- Jansen R. P. Rna-cytoskeletal associations. *FASEB J*, 13(3):455–66, 1999.
- Jens M. and Rajewsky N. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nat Rev Genet*, 16(2):113–26, 2015.
- Jensen K. B., Dredge B. K., Stefani G., Zhong R., Buckanovich R. J. et al. Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability. *Neuron*, 25(2):359–71, 2000a.

- Jensen K. B., Musunuru K., Lewis H. A., Burley S. K. and Darnell R. B. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain. *Proc Natl Acad Sci U S A*, 97(11):5740–5, 2000b.
- Jia G., Fu Y., Zhao X., Dai Q., Zheng G. et al. N6-methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nat Chem Biol*, 7(12):885–7, 2011.
- Jiang Q., Wang Y., Hao Y., Juan L., Teng M. et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*, 37(Database issue):D98–104, 2009.
- Jiao Y. and Meyerowitz E. M. Cell-type specific analysis of translating RNAs in developing flowers reveals new levels of control. *Mol Syst Biol*, 6:419, 2010.
- Jin Y., Suzuki H., Maegawa S., Endo H., Sugano S. et al. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. *EMBO J*, 22(4):905–12, 2003.
- Jing Q., Huang S., Guth S., Zarubin T., Motoyama A. et al. Involvement of microRNA in AU-rich element-mediated mRNA instability. *Cell*, 120(5):623–34, 2005.
- Jones S., Daley D. T., Luscombe N. M., Berman H. M. and Thornton J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res*, 29(4):943–54, 2001.
- Joo Y. J., Kim J.-H., Kang U.-B., Yu M.-H. and Kim J. Gcn4p-mediated transcriptional repression of ribosomal protein genes under amino-acid starvation. *EMBO J*, 30(5):859–72, 2011.
- Kamminga L. M., van Wolfswinkel J. C., Luteijn M. J., Kaaij L., Bagijn M. P. et al. Differential impact of the Hen1 homolog HENN-1 on 21U and 26G RNAs in the germline of *Caenorhabditis elegans*. *PLoS Genet*, 8(7), 2012.
- Kasper D. M., Wang G., Gardner K. E., Johnstone T. G. and Reinke V. The *C. elegans* SNAPc component SNPC-4 coats piRNA domains and is globally required for piRNA abundance. *Dev Cell*, 31(2):145–58, 2014.
- Kato M., de Lencastre A., Pincus Z. and Slack F. J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Gen Biol*, 10(5):R54, 2009.
- Kato M., Han T. W., Xie S., Shi K., Du X. et al. Cell-free formation of RNA granules: low complexity sequence domains form dynamic fibers within hydrogels. *Cell*, 149(4):753–67, 2012.
- Kawaoka S., Izumi N., Katsuma S. and Tomari Y. 3' end formation of PIWI-interacting RNAs in vitro. *Mol Cell*, 43(6):1015–1022, 2011.
- Kaye J. A., Rose N. C., Goldsworthy B., Goga A. and L'Etoile N. D. A 3'UTR pumilio-binding element directs translational activation in olfactory sensory neurons. *Neuron*, 61(1):57–70, 2009.

- Kaymak E. and Ryder S. P. RNA recognition by the *Caenorhabditis elegans* oocyte maturation determinant OMA-1. *J Biol Chem*, 288(42):30463–72, 2013.
- Kechavarzi B. and Janga S. C. Dissecting the expression landscape of RNA-binding proteins in human cancers. *Gen Biol*, 15(1):R14, 2014.
- Kedde M., van Kouwenhove M., Zwart W., Oude Vrielink J. A. F., Elkon R. et al. A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. *Nat Cell Biol*, 12(10):1014–20, 2010.
- Kedersha N., Ivanov P. and Anderson P. Stress granules and cell signaling: more than just a passing phase? *Trends Biochem Sci*, 38(10):494–506, 2013.
- Kellis M., Birren B. W. and Lander E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, 428(6983):617–24, 2004.
- Kertesz M., Wan Y., Mazor E., Rinn J. L., Nutter R. C. et al. Genome-wide measurement of rna secondary structure in yeast. *Nature*, 467(7311):103–7, 2010.
- Keshav K. F., Chen C. and Dutta A. Rpa4, a homolog of the 34-kilodalton subunit of the replication protein A complex. *Mol Cell Biol*, 15(6):3119–28, 1995.
- Khorshid M., Rodak C. and Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res*, 39(Database issue):D245–52, 2011.
- Kim H. H., Kuwano Y., Srikantan S., Lee E. K., Martindale J. L. et al. HuR recruits let-7/RISC to repress c-Myc expression. *Gen Dev*, 23(15):1743–8, 2009.
- Kircher M., Stenzel U. and Kelso J. Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Gen Biol*, 10(8):R83, 2009.
- Kirino Y. and Mourelatos Z. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. *Nat Struct Mol Biol*, 14(4):347–348, 2007.
- Klattenhoff C. and Theurkauf W. Biogenesis and germline functions of piRNAs. *Development*, 135(1):3–9, 2008.
- Koh Y. Y., Wang Y., Qiu C., Opperman L., Gross L. et al. Stacking interactions in PUF-RNA complexes. *RNA*, 17(4):718–27, 2011.
- König J., Zarnack K., Rot G., Curk T., Kayikci M. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–15, 2010.
- Kozomara A. and Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res*, 39(Database issue):D152–7, 2011.

- Kozomara A. and Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42(Database issue):D68–73, 2014.
- Kumar P., Anaya J., Mudunuri S. B. and Dutta A. Meta-analysis of tRNA derived RNA fragments reveals that they are evolutionarily conserved and associate with AGO proteins to recognize specific RNA targets. *BMC Biol*, 12:78, 2014.
- Kumar P., Mudunuri S. B., Anaya J. and Dutta A. tRFdb: a database for transfer RNA fragments. *Nucleic Acids Res*, 43(Database issue):D141–5, 2015.
- Kundu P., Fabian M. R., Sonenberg N., Bhattacharyya S. N. and Filipowicz W. HuR protein attenuates miRNA-mediated repression by promoting miRISC dissociation from the target RNA. *Nucleic Acids Res*, 40(11):5088–100, 2012.
- Kuo M.-W., Wang S.-H., Chang J.-C., Chang C.-H., Huang L.-J. et al. A novel puf-A gene predicted from evolutionary analysis is involved in the development of eyes and primordial germ-cells. *PLoS One*, 4(3):e4980, 2009.
- Kuramochi-Miyagawa S., Watanabe T., Gotoh K., Totoki Y., Toyoda A. et al. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Gen Dev*, 22(7):908–917, 2008.
- Kurth H. M. and Mochizuki K. 2'-O-methylation stabilizes Piwi-associated small RNAs and ensures DNA elimination in Tetrahymena. *RNA*, 15(4):675–685, 2009.
- Kvon E. Z., Stampfel G., Yáñez-Cuna J. O., Dickson B. J. and Stark A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Gen Dev*, 26(9):908–13, 2012.
- Kwon I., Kato M., Xiang S., Wu L., Theodoropoulos P. et al. Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains. *Cell*, 155(5):1049–60, 2013a.
- Kwon S. C., Yi H., Eichelbaum K., Föhr S., Fischer B. et al. The RNA-binding protein repertoire of embryonic stem cells. *Nat Struct Mol Biol*, 20(9):1122–30, 2013b.
- Lagos-Quintana M., Rauhut R., Lendeckel W. and Tuschl T. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, 2001.
- Langmead B. and Salzberg S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods*, 9(4):357–9, 2012.
- Langmead B., Trapnell C., Pop M. and Salzberg S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Gen Biol*, 10(3):R25, 2009.
- Larochelle M., Drouin S., Robert F. and Turcotte B. Oxidative stress-activated zinc cluster protein Stb5 has dual activator/repressor functions required for pentose phosphate pathway regulation and NADPH production. *Mol Cell Biol*, 26(17):6690–701, 2006.

- Larويا G., Cuesta R., Brewer G. and Schneider R. J. Control of mRNA decay by heat shock-ubiquitin-proteasome pathway. *Science*, 284(5413):499–502, 1999.
- Lasham A., Print C. G., Woolley A. G., Dunn S. E. and Braithwaite A. W. YB-1: oncoprotein, prognostic marker and therapeutic target? *Biochem J*, 449(1):11–23, 2013.
- Lau N. C., Lim L. P., Weinstein E. G. and Bartel D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science*, 294(5543):858–62, 2001.
- Lau N. C., Seto A. G., Kim J., Kuramochi-Miyagawa S., Nakano T. et al. Characterization of the piRNA complex from rat testes. *Science*, 313(5785):363–367, 2006.
- Le Thomas A., Stuwe E., Li S., Du J., Marinov G. et al. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Gen Dev*, 28(15):1667–80, 2014.
- Lebedeva S., Jens M., Theil K., Schwanhäusser B., Selbach M. et al. Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol Cell*, 43(3):340–52, 2011.
- Lécuyer E., Yoshida H., Parthasarathy N., Alm C., Babak T. et al. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, 131(1):174–87, 2007.
- Lee D., Ohn T., Chiang Y.-C., Quigley G., Yao G. et al. PUF3 acceleration of deadenylation in vivo can operate independently of CCR4 activity, possibly involving effects on the PAB1-mRNP structure. *J Mol Biol*, 399(4):562–75, 2010.
- Lee H. C., Gu W., Shirayama M., Youngman E., Conte D. J. et al. *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*, 150(1):78–87, 2012.
- Lee R. C. and Ambros V. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science*, 294(5543):862–4, 2001.
- Lee R. C., Feinbaum R. L. and Ambros V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75(5):843–54, 1993.
- Lessnick S. L. and Ladanyi M. Molecular pathogenesis of Ewing sarcoma: new therapeutic and transcriptional targets. *Annu Rev Pathol*, 7:145–59, 2012.
- Levin J. Z., Yassour M., Adiconis X., Nusbaum C., Thompson D. A. et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat Methods*, 7(9):709–715, 2010.
- Levine M. and Tjian R. Transcription regulation and animal diversity. *Nature*, 424(6945):147–51, 2003.

- Li C., Vagin V. V., Lee S., Xu J., Ma S. et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell*, 137(3):509–521, 2009a.
- Li H. and Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–60, 2009.
- Li H., Handsaker B., Wysoker A., Fennell T., Ruan J. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–9, 2009b.
- Li Q., Lee J.-A. and Black D. L. Neuronal regulation of alternative pre-mRNA splicing. *Nat Rev Neurosci*, 8(11):819–31, 2007.
- Li S.-C., Tsai K.-W., Pan H.-W., Jeng Y.-M., Ho M.-R. et al. MicroRNA 3' end nucleotide modification patterns and arm selection preference in liver tissues. *BMC Syst Biol*, 6 Suppl 2:S14, 2012.
- Li Y., Zhao D. Y., Greenblatt J. F. and Zhang Z. RIPSeeker: a statistical package for identifying protein-associated transcripts from RIP-seq experiments. *Nucleic Acids Res*, 41(8):e94, 2013.
- Li Z., Lee I., Moradi E., Hung N.-J., Johnson A. W. et al. Rational extension of the ribosome biogenesis pathway using network-guided genetics. *PLoS Biol*, 7(10):e1000213, 2009c.
- Liang J., Wang J., Azfer A., Song W., Tromp G. et al. A novel CCCH-zinc finger protein family regulates proinflammatory activation of macrophages. *J Biol Chem*, 283(10):6337–46, 2008.
- Licatalosi D. D. and Darnell R. B. Splicing regulation in neurologic disease. *Neuron*, 52(1):93–101, 2006.
- Licatalosi D. D., Mele A., Fak J. J., Ule J., Kayikci M. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–9, 2008.
- Lin H. and Spradling A. C. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. *Development*, 124(12):2463–2476, 1997.
- Little M. and Wells C. A clinical overview of WT1 gene mutations. *Hum Mutat*, 9(3):209–25, 1997.
- Liu J., Yue Y., Han D., Wang X., Fu Y. et al. A METTL3-METTL14 complex mediates mammalian nuclear RNA N⁶-adenosine methylation. *Nat Chem Biol*, 10(2):93–5, 2014.
- Long R. M., Gu W., Lorimer E., Singer R. H. and Chartrand P. She2p is a novel RNA-binding protein that recruits the Myo4p-She3p complex to ASH1 mRNA. *EMBO J*, 19(23):6592–601, 2000.

- Lorenz R., Bernhart S. H., Höner Zu Siederdisen C., Tafer H., Flamm C. et al. ViennaRNA Package 2.0. *Algorithms Mol Biol*, 6:26, 2011.
- Lu L., Wang S., Zheng L., Li X., Suswam E. A. et al. Amyotrophic lateral sclerosis-linked mutant SOD1 sequesters Hu antigen R (HuR) and TIA-1-related protein (TIAR): implications for impaired post-transcriptional regulation of vascular endothelial growth factor. *J Biol Chem*, 284(49):33989–98, 2009.
- Lucks J. B., Mortimer S. A., Trapnell C., Luo S., Aviran S. 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*, 108(27):11063–8, 2011.
- Lukong K. E., Chang K.-w., Khandjian E. W. and Richard S. RNA-binding proteins in human genetic disease. *Trends Genet*, 24(8):416–25, 2008.
- Lunde B. M., Moore C. and Varani G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, 8(6):479–90, 2007.
- Luteijn M. J., van Bergeijk P., Kaaij L. J., Almeida M. V., Roovers E. F. et al. Extremely stable Piwi-induced gene silencing in *Caenorhabditis elegans*. *EMBO J*, 31(16):3422–3430, 2012.
- Lynch M. and Conery J. S. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, 2000.
- Maatz H., Jens M., Liss M., Schafer S., Heinig M. et al. RNA-binding protein RBM20 represses splicing to orchestrate cardiac pre-mRNA processing. *J Clin Invest*, 124(8):3419–30, 2014.
- Magidovich E., Orr I., Fass D., Abdu U. and Yifrach O. Intrinsic disorder in the C-terminal domain of the Shaker voltage-activated K⁺ channel modulates its interaction with scaffold proteins. *Proc Natl Acad Sci U S A*, 104(32):13022–7, 2007.
- Maillet L. and Collart M. A. Interaction between Not1p, a component of the Ccr4-not complex, a global regulator of transcription, and Dhh1p, a putative RNA helicase. *J Biol Chem*, 277(4):2835–42, 2002.
- Marioni J. C., Mason C. E., Mane S. M., Stephens M. and Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–17, 2008.
- Martin K. C. and Ephrussi A. mRNA localization: gene expression in the spatial dimension. *Cell*, 136(4):719–30, 2009.
- Matoulkova E., Michalova E., Vojtesek B. and Hrstka R. The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biol*, 9(5):563–76, 2012.

- Mayrand S., Setyono B., Greenberg J. R. and Pederson T. Structure of nuclear ribonucleoprotein: identification of proteins in contact with poly(A)⁺ heterogeneous nuclear RNA in living HeLa cells. *J Cell Biol*, 90(2):380–4, 1981.
- McHugh C. A., Russell P. and Guttman M. Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biol*, 15(1):203, 2014.
- McKenna A., Hanna M., Banks E., Sivachenko A., Cibulskis K. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–303, 2010.
- Meacham F., Boffelli D., Dhahbi J., Martin D. I. K., Singer M. et al. Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics*, 12:451, 2011.
- Menon K. P., Sanyal S., Habara Y., Sanchez R., Wharton R. P. et al. The translational repressor Pumilio regulates presynaptic morphology and controls postsynaptic accumulation of translation factor eIF-4E. *Neuron*, 44(4):663–76, 2004.
- Miles W. O., Tschöp K., Herr A., Ji J.-Y. and Dyson N. J. Pumilio facilitates miRNA regulation of the E2F3 oncogene. *Genes Dev*, 26(4):356–68, 2012.
- Mili S. and Steitz J. A. Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA*, 10(11):1692–4, 2004.
- Miller C., Schwalb B., Maier K., Schulz D., Dümcke S. et al. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol Syst Biol*, 7:458, 2011.
- Miller M. A. and Olivas W. M. Roles of Puf proteins in mRNA degradation and translation. *Wiley Interdiscip Rev RNA*, 2(4):471–92, 2011.
- Miller M. A., Russo J., Fischer A. D., Lopez Leban F. A. and Olivas W. M. Carbon source-dependent alteration of Puf3p activity mediates rapid changes in the stabilities of mRNAs involved in mitochondrial function. *Nucleic Acids Res*, 42(6):3954–70, 2014.
- Miller M. T., Higgin J. J. and Hall T. M. T. Basis of altered RNA-binding specificity by PUF proteins revealed by crystal structures of yeast Puf4p. *Nat Struct Mol Biol*, 15(4):397–402, 2008.
- Minoche A. E., Dohm J. C. and Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Gen Biol*, 12(11):R112, 2011.
- Mitchell S. F., Jain S., She M. and Parker R. Global analysis of yeast mRNPs. *Nat Struct Mol Biol*, 20(1):127–33, 2013.

- Mitchell S. F. and Parker R. Principles and properties of eukaryotic mRNPs. *Mol Cell*, 54(4):547–58, 2014.
- Mochizuki K., Fine N. A., Fujisawa T. and Gorovsky M. A. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell*, 110(6):689–699, 2002.
- modENCODE Consortium, Roy S., Ernst J., Kharchenko P. V., Kheradpour P. et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, 330(6012):1787–97, 2010.
- Montgomery T. A., Rim Y.-S., Zhang C., Downen R. H., Phillips C. M. et al. PIWI Associated siRNAs and piRNAs Specifically Require the Caenorhabditis elegans HEN1 Ortholog henn-1. *PLoS Genet*, 8(4):e1002616, 2012.
- Moore M. J. and Proudfoot N. J. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, 2009.
- Moorman C., Sun L. V., Wang J., de Wit E., Talhout W. et al. Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 103(32):12027–32, 2006.
- Morrison H. *4-thiouridine as an intrinsic photoaffinity probe of nucleic acid structure and interactions*, volume 1. Wiley, New York, 1990.
- Mortazavi A., Williams B. A., McCue K., Schaeffer L. and Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–8, 2008.
- Mukherjee N., Corcoran D. L., Nusbaum J. D., Reid D. W., Georgiev S. et al. Integrative regulatory mapping indicates that the RNA-binding protein HuR couples pre-mRNA processing and mRNA stability. *Mol Cell*, 43(3):327–39, 2011.
- Müller S., Rycak L., Winter P., Kahl G., Koch I. et al. omiRas: a Web server for differential expression analysis of miRNAs derived from small RNA-Seq data. *Bioinformatics*, 29(20):2651–2, 2013.
- Muraro N. I., Weston A. J., Gerber A. P., Luschnig S., Moffat K. G. et al. Pumilio binds para mRNA and requires Nanos and Brat to regulate sodium current in Drosophila motoneurons. *J Neurosci*, 28(9):2099–109, 2008.
- Murzin A. G., Brenner S. E., Hubbard T. and Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*, 247(4):536–40, 1995.
- Nagalakshmi U., Wang Z., Waern K., Shou C., Raha D. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–9, 2008.

- Nakamura K., Oshima T., Morimoto T., Ikeda S., Yoshikawa H. et al. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*, 39(13):e90, 2011.
- Nègre N., Brown C. D., Ma L., Bristow C. A., Miller S. W. et al. A cis-regulatory map of the *Drosophila* genome. *Nature*, 471(7339):527–31, 2011.
- Nichols R. C., Rudolphi O., Ek B., Exelbert R., Plotz P. H. et al. Glycogenesis type VII (Tarui disease) in a Swedish family: two novel mutations in muscle phosphofructokinase gene (PFK-M) resulting in intron retentions. *Am J Hum Genet*, 59(1):59–65, 1996.
- Nishimasu H., Ishizu H., Saito K., Fukuhara S., Kamatani M. K. et al. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. *Nature*, 491(7423):284–7, 2012.
- Nolan T., Hands R. E. and Bustin S. A. Quantification of mRNA using real-time RT-PCR. *Nat Protoc*, 1(3):1559–1582, 2006.
- Nolde M. J., Saka N., Reinert K. L. and Slack F. J. The *Caenorhabditis elegans* pumilio homolog, puf-9, is required for the 3'UTR-mediated repression of the let-7 microRNA target gene, hbl-1. *Dev Biol*, 305(2):551–63, 2007.
- Ohara T., Sakaguchi Y., Suzuki T., Ueda H. and Miyauchi K. The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. *Nat Struct Mol Biol*, 14(4):349–350, 2007.
- Oldenbroek M., Robertson S. M., Guven-Ozkan T., Gore S., Nishi Y. et al. Multiple RNA-binding proteins function combinatorially to control the soma-restricted expression pattern of the E3 ligase subunit ZIF-1. *Dev Biol*, 363(2):388–98, 2012.
- Olivas W. and Parker R. The Puf3 protein is a transcript-specific regulator of mRNA degradation in yeast. *EMBO J*, 19(23):6602–11, 2000.
- Orengo C. A., Michie A. D., Jones S., Jones D. T., Swindells M. B. et al. CATH—a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–108, 1997.
- Oroguchi T., Ikeguchi M. and Sato M. Towards the structural characterization of intrinsically disordered proteins by SAXS and MD simulation. In *Journal of Physics: Conference Series*, volume 272, page 012005. IOP Publishing, 2011.
- Ozsolak F., Kapranov P., Foissac S., Kim S. W., Fishilevich E. et al. Comprehensive polyadenylation site maps in yeast and human reveal pervasive alternative polyadenylation. *Cell*, 143(6):1018–29, 2010.
- Pauley K. M., Eystathioy T., Jakymiw A., Hamel J. C., Fritzler M. J. et al. Formation of GW bodies is a consequence of microRNA genesis. *EMBO Rep*, 7(9):904–10, 2006.
- Paulsen M. T., Veloso A., Prasad J., Bedi K., Ljungman E. A. et al. Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. *Proc Natl Acad Sci U S A*, 110(6):2240–5, 2013.

- Phan A. T., Kuryavyi V., Darnell J. C., Serganov A., Majumdar A. et al. Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat Struct Mol Biol*, 18(7):796–804, 2011.
- Piper M. and Holt C. RNA translation in axons. *Annu Rev Cell Dev Biol*, 20:505–23, 2004.
- Piqué M., López J. M., Foissac S., Guigó R. and Méndez R. A combinatorial code for CPE-mediated translational control. *Cell*, 132(3):434–48, 2008.
- Pollard K. S., Hubisz M. J., Rosenbloom K. R. and Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*, 20(1):110–21, 2010.
- Preitner N., Quan J., Nowakowski D. W., Hancock M. L., Shi J. et al. APC is an RNA-binding protein, and its interactome provides a link to neural development and microtubule assembly. *Cell*, 158(2):368–82, 2014.
- Puton T., Kozłowski L., Tuszyńska I., Rother K. and Bujnicki J. M. Computational methods for prediction of protein-RNA interactions. *J Struct Biol*, 179(3):261–8, 2012.
- Qiu C., Kershner A., Wang Y., Holley C. P., Wilinski D. et al. Divergence of Pumilio/fem-3 mRNA binding factor (PUF) protein specificity through variations in an RNA-binding pocket. *J Biol Chem*, 287(9):6949–57, 2012.
- Qiu C., McCann K. L., Wine R. N., Baserga S. J. and Hall T. M. T. A divergent Pumilio repeat protein family for pre-rRNA processing and mRNA localization. *Proc Natl Acad Sci U S A*, 111(52):18554–9, 2014.
- Quenault T., Lithgow T. and Traven A. PUF proteins: repression, activation and mRNA localization. *Trends Cell Biol*, 21(2):104–12, 2011.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.
- Raben N., Exelbert R., Spiegel R., Sherman J. B., Nakajima H. et al. Functional expression of human mutant phosphofructokinase in yeast: genetic defects in French Canadian and Swiss patients with phosphofructokinase deficiency. *Am J Hum Genet*, 56(1):131–41, 1995.
- Ray D., Kazan H., Cook K. B., Weirauch M. T., Najafabadi H. S. et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–7, 2013.
- Reimand J., Arak T. and Vilo J. g:Profiler—a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res*, 39(Web Server issue):W307–15, 2011.
- Reimand J., Kull M., Peterson H., Hansen J. and Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res*, 35(Web Server issue):W193–200, 2007.

- Reinhart B. J., Slack F. J., Basson M., Pasquinelli A. E., Bettinger J. C. et al. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403(6772):901–6, 2000.
- Reinke V., Gil I. S., Ward S. and Kazmer K. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development*, 131(2):311–23, 2004.
- Reyes-Herrera P. H. and Ficarra E. Computational Methods for CLIP-seq Data Processing. *Bioinform Biol Insights*, 8:199–207, 2014.
- Riordan D. P., Herschlag D. and Brown P. O. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucleic Acids Res*, 39(4):1501–9, 2011.
- Rogelj B., Easton L. E., Bogu G. K., Stanton L. W., Rot G. et al. Widespread binding of FUS along nascent RNA regulates alternative splicing in the brain. *Sci Rep*, 2:603, 2012.
- Rowe W., Kershaw C. J., Castelli L. M., Costello J. L., Ashe M. P. et al. Puf3p induces translational repression of genes linked to oxidative stress. *Nucleic Acids Res*, 42(2):1026–41, 2014.
- Ruby J. G., Jan C., Player C., Axtell M. J., Lee W. et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127(6):1193–1207, 2006.
- Saint-Georges Y., Garcia M., Delaveau T., Jourdren L., Le Crom S. et al. Yeast mitochondrial biogenesis: a role for the PUF RNA-binding protein Puf3p in mRNA localization. *PLoS One*, 3(6):e2293, 2008.
- Sakaguchi A., Sarkies P., Simon M., Doebley A.-L., Goldstein L. D. et al. *Caenorhabditis elegans* RSD-2 and RSD-6 promote germ cell immortality by maintaining small interfering RNA populations. *Proc Natl Acad Sci U S A*, 111(41):E4323–31, 2014.
- Santos A. C. and Lehmann R. Germ cell specification and migration in *Drosophila* and beyond. *Curr Biol*, 14(14):R578–89, 2004.
- Schäffler K., Schulz K., Hirmer A., Wiesner J., Grimm M. et al. A stimulatory role for the La-related protein 4B in translation. *RNA*, 16(8):1488–99, 2010.
- Schattner P., Decatur W. A., Davis C. A., Ares, Jr M., Fournier M. J. et al. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*, 32(14):4281–96, 2004.
- Scheckel C., Gaidatzis D., Wright J. E. and Ciosk R. Genome-wide analysis of GLD-1-mediated mRNA regulation suggests a role in mRNA storage. *PLoS Genet*, 8(5):e1002742, 2012.

- Scherrer T., Mittal N., Janga S. C. and Gerber A. P. A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS One*, 5(11):e15499, 2010.
- Schreve J. L. and Garrett J. M. Yeast Agp2p and Agp3p function as amino acid permeases in poor nutrient conditions. *Biochem Biophys Res Commun*, 313(3):745–51, 2004.
- Schueler M., Munschauer M., Gregersen L. H., Finzel A., Loewer A. et al. Differential protein occupancy profiling of the mRNA transcriptome. *Gen Biol*, 15(1):R15, 2014.
- Seay D., Hook B., Evans K. and Wickens M. A three-hybrid screen identifies mRNAs controlled by a regulatory protein. *RNA*, 12(8):1594–600, 2006.
- Segref A., Sharma K., Doye V., Hellwig A., Huber J. et al. Mex67p, a novel factor for nuclear mRNA export, binds to both poly(A)+ RNA and nuclear pores. *EMBO J*, 16(11):3256–71, 1997.
- Serganov A. and Patel D. J. Towards deciphering the principles underlying an mRNA recognition code. *Curr Opin Struct Biol*, 18(1):120–9, 2008.
- Shahbadian K. and Chartrand P. Control of cytoplasmic mRNA localization. *Cell Mol Life Sci*, 69(4):535–52, 2012.
- Shahbadian K., Jeronimo C., Forget A., Robert F. and Chartrand P. Co-transcriptional recruitment of Puf6 by She2 couples translational repression to mRNA localization. *Nucleic Acids Res*, 42(13):8692–704, 2014.
- Sharp J. A., Plant J. J., Ohsumi T. K., Borowsky M. and Blower M. D. Functional analysis of the microtubule-interacting transcriptome. *Mol Biol Cell*, 22(22):4312–23, 2011.
- Sheth U. and Parker R. Targeting of aberrant mRNAs to cytoplasmic processing bodies. *Cell*, 125(6):1095–109, 2006.
- Shimobayashi M. and Hall M. N. Making new contacts: the mTOR network in metabolism and signalling crosstalk. *Nat Rev Mol Cell Biol*, 15(3):155–62, 2014.
- Shirayama M., Seth M., Lee H. C., Gu W., Ishidate T. et al. piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. *Cell*, 150(1):65–77, 2012.
- Siepel A., Bejerano G., Pedersen J. S., Hinrichs A. S., Hou M. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15(8):1034–50, 2005.
- Sievers C., Schlumpf T., Sawarkar R., Comoglio F. and Paro R. Mixture models and wavelet transforms reveal high confidence RNA-protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res*, 40(20):e160, 2012.

- Sijen T., Steiner F. A., Thijssen K. L. and Plasterk R. H. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science*, 315(5809):244–247, 2007.
- Silverman I. M., Li F., Alexander A., Goff L., Trapnell C. et al. RNase-mediated protein footprint sequencing reveals protein-binding sites throughout the human transcriptome. *Gen Biol*, 15(1):R3, 2014.
- Simpson C. E. and Ashe M. P. Adaptation to stress in yeast: to translate or not? *Biochem Soc Trans*, 40(4):794–9, 2012.
- Siomi M. C., Miyoshi T. and Siomi H. piRNA-mediated silencing in Drosophila germlines. *Semin Cell Dev Biol*, 21(7):754–9, 2010.
- Sonenberg N. and Hinnebusch A. G. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–45, 2009.
- Spike C., Meyer N., Racen E., Orsborn A., Kirchner J. et al. Genetic analysis of the Caenorhabditis elegans GLH family of P-granule proteins. *Genetics*, 178(4):1973–87, 2008a.
- Spike C. A., Bader J., Reinke V. and Strome S. DEPS-1 promotes P-granule assembly and RNA interference in C. elegans germ cells. *Development*, 135(5):983–93, 2008b.
- Stark C., Breitkreutz B.-J., Reguly T., Boucher L., Breitkreutz A. et al. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*, 34(Database issue):D535–9, 2006.
- Steigele S., Huber W., Stocsits C., Stadler P. F. and Nieselt K. Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol*, 5:25, 2007.
- Steward O. and Schuman E. M. Compartmentalized synthesis and degradation of proteins in neurons. *Neuron*, 40(2):347–59, 2003.
- Stoeckius M., Maaskola J., Colombo T., Rahn H. P., Friedlander M. R. et al. Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat Methods*, 6(10):745–751, 2009.
- Subramaniam K. and Seydoux G. Dedifferentiation of primary spermatocytes into germ cell tumors in C. elegans lacking the pumilio-like protein PUF-8. *Curr Biol*, 13(2):134–9, 2003.
- Subtelny A. O., Eichhorn S. W., Chen G. R., Sive H. and Bartel D. P. Poly(A)-tail profiling reveals an embryonic switch in translational control. *Nature*, 508(7494):66–71, 2014.
- Sugimoto Y., König J., Hussain S., Zupan B., Curk T. et al. Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Gen Biol*, 13(8):R67, 2012.

- Suh N., Crittenden S. L., Goldstrohm A., Hook B., Thompson B. et al. FBF and its dual control of *gld-1* expression in the *Caenorhabditis elegans* germline. *Genetics*, 181(4):1249–60, 2009.
- Szcześniak M. W., Deorowicz S., Gapski J., Kaczyński Ł. and Makalowska I. miRNEST database: an integrative approach in microRNA search and annotation. *Nucleic Acids Res*, 40(Database issue):D198–204, 2012.
- Tabara H., Yigit E., Siomi H. and Mello C. C. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in *C. elegans*. *Cell*, 109(7):861–871, 2002.
- Takizawa P. A. and Vale R. D. The myosin motor, Myo4p, binds Ash1 mRNA via the adapter protein, She3p. *Proc Natl Acad Sci U S A*, 97(10):5273–8, 2000.
- Tamburino A. M., Ryder S. P. and Walhout A. J. M. A compendium of *Caenorhabditis elegans* RNA binding proteins predicts extensive regulation at multiple levels. *G3*, 3(2):297–304, 2013.
- Tcherkezian J., Cargnello M., Romeo Y., Huttlin E. L., Lavoie G. et al. Proteomic analysis of cap-dependent translation identifies LARP1 as a key regulator of 5'TOP mRNA translation. *Genes Dev*, 28(4):357–71, 2014.
- Teilum K., Olsen J. G. and Kragelund B. B. Functional aspects of protein flexibility. *Cell Mol Life Sci*, 66(14):2231–47, 2009.
- Tenenbaum S. A., Carson C. C., Lager P. J. and Keene J. D. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A*, 97(26):14085–90, 2000.
- Thedieck K., Holzwarth B., Prentzell M. T., Boehlke C., Kläsener K. et al. Inhibition of mTORC1 by astrin and stress granules prevents apoptosis in cancer cells. *Cell*, 154(4):859–74, 2013.
- Thomson T. and Lin H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu Rev Cell Dev Biol*, 25:355–76, 2009.
- Thoreen C. C., Chantranupong L., Keys H. R., Wang T., Gray N. S. et al. A unifying model for mTORC1-mediated regulation of mRNA translation. *Nature*, 485(7396):109–13, 2012.
- Tompa P. and Csermely P. The role of structural disorder in the function of RNA and protein chaperones. *FASEB J*, 18(11):1169–75, 2004.
- Torchet C., Badis G., Devaux F., Costanzo G., Werner M. et al. The complete set of H/ACA snoRNAs that guide rRNA pseudouridylations in *Saccharomyces cerevisiae*. *RNA*, 11(6):928–38, 2005.

- Trapnell C., Pachter L. and Salzberg S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.
- Trapnell C., Williams B. A., Pertea G., Mortazavi A., Kwan G. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5):511–515, 2010.
- Tsvetanova N. G., Klass D. M., Salzman J. and Brown P. O. Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. *PLoS One*, 5(9), 2010.
- Tuck A. C. and Tollervey D. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell*, 154(5):996–1009, 2013.
- Tucker M., Staples R. R., Valencia-Sanchez M. A., Muhlrud D. and Parker R. Ccr4p is the catalytic subunit of a Ccr4p/Pop2p/Notp mRNA deadenylase complex in *Saccharomyces cerevisiae*. *EMBO J*, 21(6):1427–36, 2002.
- Ulbricht R. J. and Olivas W. M. Puf1p acts in combination with other yeast Puf proteins to control mRNA stability. *RNA*, 14(2):246–62, 2008.
- Ule J., Jensen K., Mele A. and Darnell R. B. CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376–86, 2005.
- Ule J., Jensen K. B., Ruggiu M., Mele A., Ule A. et al. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–5, 2003.
- Uren P. J., Bahrami-Samani E., Burns S. C., Qiao M., Karginov F. V. et al. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, 28(23):3013–20, 2012.
- Urlaub H., Hartmuth K. and Lührmann R. A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods*, 26(2):170–81, 2002.
- Vagin V. V., Sigova A., Li C., Seitz H., Gvozdev V. et al. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, 313(5785):320–324, 2006.
- Valley C. T., Porter D. F., Qiu C., Campbell Z. T., Hall T. M. T. et al. Patterns and plasticity in RNA-protein interactions enable recruitment of multiple proteins through a single site. *Proc Natl Acad Sci U S A*, 109(16):6054–9, 2012.
- Valouev A., Ichikawa J., Tonthat T., Stuart J., Ranade S. et al. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Gen Res*, 18(7):1051–1063, 2008.
- van Gurp T. P., McIntyre L. M. and Verhoeven K. J. F. Consistent errors in first strand cDNA due to random hexamer mispriming. *PLoS One*, 8(12):e85583, 2013.

- Vander Heiden M. G., Cantley L. C. and Thompson C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science*, 324(5930):1029–33, 2009.
- Vasale J. J., Gu W., Thivierge C., Batista P. J., Claycomb J. M. et al. Sequential rounds of RNA-dependent RNA transcription drive endogenous small-RNA biogenesis in the ERGO-1/Argonaute pathway. *Proc Natl Acad Sci U S A*, 107(8):3582–3587, 2010.
- Vasconcelos O., Sivakumar K., Dalakas M. C., Quezado M., Nagle J. et al. Nonsense mutation in the phosphofructokinase muscle subunit gene associated with retention of intron 10 in one of the isolated transcripts in Ashkenazi Jewish patients with Tarui disease. *Proc Natl Acad Sci U S A*, 92(22):10322–6, 1995.
- Venne A. S., Kollipara L. and Zahedi R. P. The next level of complexity: crosstalk of posttranslational modifications. *Proteomics*, 14(4-5):513–24, 2014.
- Vessey J. P., Schoderboeck L., Gingl E., Luzi E., Riefler J. et al. Mammalian Pumilio 2 regulates dendrite morphogenesis and synaptic function. *Proc Natl Acad Sci U S A*, 107(7):3222–7, 2010.
- Vyas V. K., Berkey C. D., Miyao T. and Carlson M. Repressors Nrg1 and Nrg2 regulate a set of stress-responsive genes in *Saccharomyces cerevisiae*. *Eukaryot Cell*, 4(11):1882–91, 2005.
- Wagenmakers A. J., Reinders R. J. and van Venrooij W. J. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur J Biochem*, 112(2):323–30, 1980.
- Wang J.-J., Cui D.-Y., Xiao T., Sun X., Zhang P. et al. The influences of PRG-1 on the expression of small RNAs and mRNAs. *BMC Genomics*, 15:321, 2014a.
- Wang L., Feng Z., Wang X., Wang X. and Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–8, 2010.
- Wang T., Chen B., Kim M., Xie Y. and Xiao G. A model-based approach to identify binding sites in CLIP-Seq data. *PLoS One*, 9(4):e93248, 2014b.
- Wang T., Xie Y. and Xiao G. dCLIP: a computational approach for comparative CLIP-seq analyses. *Gen Biol*, 15(1):R11, 2014c.
- Wang X., Lu Z., Gomez A., Hon G. C., Yue Y. et al. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–20, 2014d.
- Wang X., McLachlan J., Zamore P. D. and Hall T. M. T. Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110(4):501–12, 2002.

- Wang Y., Li Y., Toth J. I., Petroski M. D., Zhang Z. et al. N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat Cell Biol*, 16(2):191–8, 2014e.
- Wang Y., Opperman L., Wickens M. and Hall T. M. T. Structural basis for specific recognition of multiple mRNA targets by a PUF regulatory protein. *Proc Natl Acad Sci U S A*, 106(48):20186–91, 2009a.
- Wang Z., Gerstein M. and Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009b.
- Ward A. J., Rimer M., Killian J. M., Dowling J. J. and Cooper T. A. CUGBP1 overexpression in mouse skeletal muscle reproduces features of myotonic dystrophy type 1. *Hum Mol Genet*, 19(18):3614–22, 2010.
- Waring S. C. and Rosenberg R. N. Genome-wide association studies in Alzheimer disease. *Arch Neurol*, 65(3):329–34, 2008.
- Washietl S., Hofacker I. L. and Stadler P. F. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, 102(7):2454–9, 2005.
- Webb S., Hector R. D., Kudla G. and Granneman S. PAR-CLIP data indicate that Nrd1-Nab3-dependent transcription termination regulates expression of hundreds of protein coding genes in yeast. *Genome Biol*, 15(1):R8, 2014.
- Wei W.-J., Mu S.-R., Heiner M., Fu X., Cao L.-J. et al. YB-1 binds to CAUC motifs and stimulates exon inclusion by enhancing the recruitment of U2AF to weak polypyrimidine tracts. *Nucleic Acids Res*, 40(17):8622–36, 2012.
- Wen J., Parker B. J., Jacobsen A. and Krogh A. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA*, 17(5):820–34, 2011.
- Wickens M., Bernstein D. S., Kimble J. and Parker R. A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet*, 18(3):150–7, 2002.
- Will C. L. and Lührmann R. Spliceosome structure and function. *Cold Spring Harb Perspect Biol*, 3(7), 2011.
- Wilson G. M. and Brewer G. The search for trans-acting factors controlling messenger RNA decay. *Prog Nucleic Acid Res Mol Biol*, 62:257–91, 1999.
- Wilusz C. J., Wormington M. and Peltz S. W. The cap-to-tail guide to mRNA turnover. *Nat Rev Mol Cell Biol*, 2(4):237–46, 2001.
- Wippich F., Bodenmiller B., Trajkovska M. G., Wanka S., Aebersold R. et al. Dual specificity kinase DYRK3 couples stress granule condensation/dissolution to mTORC1 signaling. *Cell*, 152(4):791–805, 2013.

- Woolford, Jr J. L. and Baserga S. J. Ribosome biogenesis in the yeast *Saccharomyces cerevisiae*. *Genetics*, 195(3):643–81, 2013.
- Xie C., Huang L., Xie S., Xie D., Zhang G. et al. LARP1 predict the prognosis for early-stage and AFP-normal hepatocellular carcinoma. *J Transl Med*, 11:272, 2013.
- Xu C., Wang X., Liu K., Roundtree I. A., Tempel W. et al. Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain. *Nat Chem Biol*, 10(11):927–9, 2014.
- Yamanaka S., Siomi M. C. and Siomi H. piRNA clusters and open chromatin structure. *Mob DNA*, 5:22, 2014.
- Yan L., Wu S., Zhang S., Ji G. and Gu A. Genetic variants in telomerase reverse transcriptase (TERT) and telomerase-associated protein 1 (TEP1) and the risk of male infertility. *Gene*, 534(2):139–43, 2014.
- Yang X., Zhang H. and Li L. Global analysis of gene-level microRNA expression in Arabidopsis using deep sequencing data. *Genomics*, 98(1):40–6, 2011.
- Yang Y.-C. T., Di C., Hu B., Zhou M., Liu Y. et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16(1):51, 2015.
- Yang Z. Q., Streicher K. L., Ray M. E., Abrams J. and Ethier S. P. Multiple interacting oncogenes on the 8p11-p12 amplicon in human breast cancer. *Cancer Res*, 66(24):11632–43, 2006.
- Yassour M., Kaplan T., Fraser H. B., Levin J. Z., Pfiffner J. et al. Ab initio construction of a eukaryotic transcriptome by massively parallel mRNA sequencing. *Proc Natl Acad Sci U S A*, 106(9):3264–9, 2009.
- Yeo G. *Systems Biology of RNA Binding Proteins*. Advances in Experimental Medicine and Biology. Springer New York, 2014.
- Yosefzon Y., Koh Y. Y., Chritton J. J., Lande A., Leibovich L. et al. Divergent RNA binding specificity of yeast Puf2p. *RNA*, 17(8):1479–88, 2011.
- Yun J., Wang T. and Xiao G. Bayesian hidden Markov models to identify RNA-protein interaction sites in PAR-CLIP. *Biometrics*, 70(2):430–40, 2014.
- Zhang B., Gallegos M., Puoti A., Durkin E., Fields S. et al. A conserved RNA-binding protein that regulates sexual fates in the *C. elegans* hermaphrodite germ line. *Nature*, 390(6659):477–84, 1997.
- Zhang C. and Darnell R. B. Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat Biotech*, 29(7):607–14, 2011.
- Zheng G., Dahl J. A., Niu Y., Fedorcsak P., Huang C.-M. et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol Cell*, 49(1):18–29, 2013.

Zhou Z., Licklider L. J., Gygi S. P. and Reed R. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–5, 2002.

Zhu D., Stumpf C. R., Krahn J. M., Wickens M. and Hall T. M. T. A 5' cytosine binding pocket in Puf3p specifies regulation of mitochondrial mRNAs. *Proc Natl Acad Sci U S A*, 106(48):20192–7, 2009.

Zipor G., Haim-Vilmovsky L., Gelin-Licht R., Gadir N., Brocard C. et al. Localization of mRNAs coding for peroxisomal proteins in the yeast, *Saccharomyces cerevisiae*. *Proc Natl Acad Sci U S A*, 106(47):19848–53, 2009.

Zisoulis D. G., Lovci M. T., Wilbert M. L., Hutt K. R., Liang T. Y. et al. Comprehensive discovery of endogenous Argonaute binding sites in *Caenorhabditis elegans*. *Nat Struct Mol Biol*, 17(2):173–9, 2010.