COMPREHENSIVE ANALYSIS OF EUKARYOTIC SM-ASSOCIATED RIBONUCLEOPROTEIN COMPLEXES

Zhipeng Lu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology (Molecular, Cellular and Deveopmental Biology).

Chapel Hill
2014

Approved by:

A. Gregory Matera

Corbin D. Jones

Alain Laederach

Mark Peifer

William F. Marzluff

ABSTRACT

Zhipeng Lu: **COMPREHENSIVE ANALYSIS OF EUKARYOTIC SM-ASSOCIATED RIBONUCLEOPROTEIN COMPLEXES**
(Under the direction of A. Gregory Matera)

Sm proteins are a family of highly conserved RNA binding proteins present in all three domains of life. These proteins form oligomeric rings and play important roles in many aspects of RNA metabolism. In archaea and bacteria, Sm proteins associate with mRNAs and small RNAs to regulate translation and stability. In eukaryotes, Sm proteins together with their associated RNAs form several distinct complexes to control splicing, histone mRNA processing, and mRNA degradation.

Recent studies suggested that canonical Sm proteins, core components of spliceosomal small nuclear ribonucleoproteins (snRNPs), have functions beyond splicing. The goal of this dissertation is therefore to develop new experimental and computational tools to identify Sm-associated RNPs, and study their structure and function on the molecular, cellular and organismal levels.

To identify Sm-associated RNAs, I developed a multi-targeting RNA immunoprecipitation sequencing (RIP-seq) method (Chapter 2). RIP-seq in *Drosophila* ovaries and human HeLa cells revealed three categories of Sm-associated RNAs: snRNAs, small Cajal body RNAs (scaRNAs) and mRNAs. Specifically, I identified a newly evolved yet highly conserved snRNA, *Like-U (LU)*. More importantly, I found that snRNPs mediate the interaction between Sm proteins and mature mRNAs, suggesting a splicing-independent function for snRNPs.

I developed a computational method, Vicinal, for the accurate determination of ncRNAs ends using chimeric reads from RNA-seq (Chapter 3). Applying Vicinal to hundreds of RNA-seq datasets, I defined the ends of numerous ncRNAs in fly, mouse and human transcriptomes, including the newly identified LU snRNA.

Most snRNAs in higher eukaryotes exist in multi-gene families, however, little is known about their contribution to splicing regulation. In Chapter 4, I analyzed expression of snRNA paralogs during vertebrate and invertebrate development. Surprisingly, I identified a developmental switch in the expression of snRNA paralogs that is conserved in evolution, despite a lack of stable orthologous groups. This work lays the foundation for genetic analysis of snRNA paralog functions.

In Chapter 5, I describe our discovery of SMN bodies in *Drosophila* testes. Our analysis of SMN bodies and U body-like RNPs suggests a concerted pathway for snRNP assembly in the cytoplasm, blockage of which leads to granule formation.

To my parents,

my brother Zhiwei,

and my wife Liang

ACKNOWLEDGMENTS

First and foremost, I must thank my parents. For more than twenty years of my life, they have provided for my brother and me with all they have, despite the hardship of living. They have raised me to be strong and determined. They have always told me that they are happy as long as I live happily, regardless of what I do and what I would achieve. I never worried about failing their expectations during the years away from home, and I keep going, knowing that the caring and love they gave me will not change. I want to thank my brother for the unforgettable childhood memories we had together. Even though we are thousands of miles apart, we are never far away from each other.

I want to thank my advisor, Dr. Greg Matera, for his support and guidance in science. Greg took me in the lab at the time when I had nowhere to go. Greg has given me all the freedom I needed to grow as an independent scientist and has always been very supportive of the projects I worked on. Over the years, I have failed miserably on many experiments, and I have started exciting projects that were originally out of the scope of the lab's expertise. All of these failures and explorations have honed my skills and judgment, and strengthened my aspiration to do great things. None of the achievements I have made would have been possible without Greg's support.

I am grateful to Dr. T. K. Rajendra for all the stimulating and encouraging conversations we had that prepared me well for not only fly genetics, but also scientific research in general. I want to thank Dr. Xiaojun Guan for helping me with the initial RIP-seq analysis presented in Chapter 2 and the analysis of chimeric reads presented in Chapter 3, Casey A. Schmidt for help with some of the experiments in Chapter 2. Ying Wen, Talia L. Hatkevich-O'Donell and John J. Noto helped a lot with an ongoing project on the biology, biogenesis and biotechnology of tricRNAs, which I did not include in this dissertation.

I would also like to thank other previous and current members of the Matera lab who have been great lab mates and friends, Dr. Graydon Gonsalvez, Dr. Jennifer Fuentes, Dr. Mario Izaguirre-Sierra, Dr. Ingo Meier, Dr. Amanda Natalizio, Dr. Kavita Praveen, Mike Meers, Dr. Eric Garcia, Dr. Stephen Klusza, Kelsey Gray, Stephen Cooper, Nathan Spain and Akash Patlolla.

Members of my dissertation committee, Dr. Bill Marzluff, Dr. Corbin Jones, Dr. Mark Peifer, Dr. Jason Lieb and Dr. Alain Laederach have encouraged me and supported me through the rough times when projects were not working well. I am deeply indebted to all these great colleagues that made my life in graduate school easier, and helped me with my postdoc search.

Last, but not least, I want to thank my beautiful, smart and kind wife, Liang, for believing in me and standing by me through the good times and bad. We started graduate school around the same time; we met in graduate school and about to finish at the same time. Together, we went through the years of working hard and enjoying the fun of life, in a land far from our parents. I have never regretted and will never regret having her as the love of my life.

TABLE OF CONTENTS

LIST OF FIGURES xiii

LIST OF TABLES xvi

CHAPTER 1: Introduction 1

1.1 From Central Dogma to RNA Regulation 1

1.2 The Sm family of RNA binding proteins: evolution and features 3

1.3 RNA partners and functions of the bacterial Sm protein Hfq 6

1.4 RNA partners and functions of the archaeal Sm proteins 10

1.5 Overview of known RNA partners and functions of the eukaryotic Sm proteins 10

1.6 Structure and composition of Sm-class snRNPs 15

1.7 snRNP assembly in eukaryotes 18

1.8 The eukaryotic pre-mRNA splicing process 22

1.9 Hints at potential new functions of eukaryotic Sm proteins 24

1.10 Novel functions of spliceosomal snRNPs 26

1.11 Methods for studying DNA-RNA-protein interactions 30

1.12 Abnormal RNA-seq reads, valuable information 32

CHAPTER 2: RIP-seq analysis of eukaryotic Sm containing ribonucleoproteins 36

2.1 Abstract 36

2.2 Background 37

2.3 Results 39

2.3.1 Identification of RNAs that co-purify with eukaryotic Sm proteins 39

2.3.2 Enrichment analysis of Sm RIP-seq experiments 41

2.3.3 A multi-targeting RIP strategy identifies Sm-associated RNAs 46

2.3.4 Sm proteins associate with three major classes of RNAs 50

2.3.5 RIP-seq identifies Sm class snRNAs 50

2.3.6 Sm proteins associate with scaRNAs........58

2.3.7 Sm-associated mRNAs encode mitochondria/translation-related proteins....64

2.3.8 Validation and tissue-specificity of RNA-Sm protein interactions........70

2.3.9 Sm proteins associate with fully spliced and polyadenylated mRNAs........71

2.3.10 Sm protein interaction with mRNAs is mediated by snRNPs........73

2.4 Discussion........81

2.4.1 Sm proteins in scaRNP complexes........81

2.4.2 Splicing-independent, evolutionarily ancient functions for Sm-class snRNPs..82

2.4.3 Technical considerations........84

2.5 Conclusions........85

2.6 Materials and methods........86

2.6.1 Fly strains and cell lines........86

2.6.2 RIP-seq experiment........86

2.6.3 RIP-seq read mapping and quantification........88

2.6.4 Assignment of reads to Drosophila snRNAs........89

2.6.5 *Drosophila* histone mRNA read mapping........89

2.6.6 In situ hybridizations........89

2.6.7 Gaussian mixture modeling........89

2.6.8 Cluster analysis of RIP-seq data........90

2.6.9 Fisher's exact test of the significance of overlap........90

2.6.10 Phylogenetic analysis........90

2.6.11 Meta-gene analysis of read density around splice junctions........91

2.6.12 Meta-gene analysis of read density along the entire gene length........91

2.6.13 Quantitative reverse-transcription PCR........91

2.6.14 CG3776 construct and transfection........92

2.6.15 Measurement of poly(A) tail length........92

2.6.16 Analysis of U1-70 K RIP-seq data........92

2.6.17 RNA secondary structure and base pairing prediction........93

CHAPTER 3: Vicinal: a method for the analysis of chimeric reads from RNA-seq 94

3.1 Abstract 94

3.2 Introduction 95

3.3 Results 96

3.3.1 The Vicinal algorithm 97

3.3.2 Confirmation of known snRNA:U1 and RNaseP:RNA ends using Vicinal ... 100

3.3.3 Vicinal analysis of newly-discovered snRNAs and sno/scaRNAs 103

3.4 Discussion 106

3.5 Materials and Methods 107

3.5.1 Total RNA-seq of pharate adult flies 107

3.5.2 Additional RNA-seq data used 108

3.5.3 Bioinformatic pipeline 109

3.5.4 Implementation of the Vicinal algorithm 110

3.5.5 ncRNA lists and generation of lists with chimeric read numbers. 111

3.5.6 Northern blotting 111

3.5.7 RNA secondary structure prediction 111

CHAPTER 4: Developmental switch of snRNA isoforms is conserved in evolution 119

4.1 Abstract 119

4.2 Introduction 119

4.3 Results and Discussion 122

4.3.1 Generation and identification of appropriate RNA-seq datasets 122

4.3.2 Structural and functional alignment of snRNA isoforms 124

4.3.3 Developmental switching of *Drosophila* snRNA isoform dominance 124

4.3.4 Developmental switching of snRNA isoform dominance is conserved 127

4.4 Conclusions 130

4.5 Materials and Methods 130

4.5.1 RNA-seq data files 130

4.5.2 Conversion of formats 131

4.5.3 Extraction of uniquely mappable reads 131

4.6 Supplementary methods: Assigning RNA-seq reads to snRNA isoforms 132

4.6.1 *Drosophila* snRNAs 132

4.6.2 *Drosophila* Single copy snRNAs 133

4.6.3 *Drosophila* U1 snRNA 134

4.6.4 *Drosophila* U2 snRNA 136

4.6.5 *Drosophila* U4 snRNA 140

4.6.6 *Drosophila* U5 snRNA 142

4.6.7 *Drosophila* U6 snRNA 144

4.6.8 Mouse U1 snRNA 144

4.6.9 Mouse U2 snRNA 147

4.6.10 Mouse U4 snRNA 148

4.6.11 Mouse U5 snRNA 148

CHAPTER 5: SMN body formation caused by a block in snRNP assembly 149

5.1 Abstract 149

5.2 Introduction 150

5.3 Results 152

5.3.1 SMN containing cytoplasmic RNP granules are not ubiquitous 154

5.3.2 The ovarian U bodies are passive snRNP transport particles 155

5.3.3 *Drosophila* testes contain SMN bodies 159

5.3.4 The SMN body is not affected by SMA-associated mutations 162

5.4 Discussion 166

5.4.1 Imbalanced cytoplasmic snRNP assembly causes U/SMN body formation ..166

5.4.2 Cytoplasmic SMN-containing granules 169

5.4.3 Comparison between U bodies and P bodies 170

5.5 Materials and Methods 170

5.5.1 Fly stocks 170

5.5.2 Antibodies and probes 171

5.5.3 Immunofluorescence and in situ hybridizations 171

5.5.4 Live imaging of ovaries and particle tracking 172

CHAPTER 6: Conclusions and future directions 173

6.1 RNA-seq applications and analysis 173

6.2 New types of Sm-class snRNAs 174

6.3 Sm-associated scaRNAs 175

6.4 snRNP-mRNA interactions 176

6.5 Developing new methods to study snRNP-mRNA interactions 177

6.6 Vicinal mapping of abnormal reads. 179

6.7 snRNA functions in alternative splicing 179

6.8 snRNP assembly and RNP granules 180

6.9 Concluding remarks 183

APPENDIX A: Characterization of SmD3 antibodies 184

A.1 Rationale 184

A.2 Results and Discussion 184

APPENDIX B: RIP-seq analysis of Lsm11 in *Drosophila* 188

B.1 Introduction 188

B.2 Materials and Methods 188

B.3 Results and Discussion 189

APPENDIX C: RNA-seq analysis *Drosophila LU* snRNA mutants 194

C.1 Rationale 194

C.2 Rescue of the *LU* mutation line *k05816(10580)* 194

C.3 *LU* mutants and rescue RNA-seq results and discussion 195

APPENDIX D: List of publications and manuscripts 200

APPENDIX E: Defense flyers 201

BIBLOGRAPHY 202

LIST OF FIGURES

1.1 Sunburst diagram of the distribution of Sm proteins. 4

1.2 Structure of the Sm proteins. 5

1.3 Diversity of main Sm rings in all three domains of life. 7

1.4 Well known functions of the bacterial Sm protein Hfq. 9

1.5 Diagram of domains of all fruitfly Sm class proteins. 12

1.6 Consensus Sm sites. 13

1.7 Secondary structure and composition of all spliceosomal and U7 snRNPs. 16

1.8 The biogenesis pathway of spliceosomal snRNPs 20

1.9 A simple diagram of U2-type splicing. 23

1.10 Hints at new functions of the canonical Sm proteins. 25

1.11 New functions for the spliceosomal snRNPs. 28

1.12 All kinds of RNA-seq Reads and their mapping patterns. 33

2.1 RIP-seq experimental analysis strategies. 40

2.2 Quality of the *Drosophila* and human RIP-seq data calculated using FastQC. 44

2.3 RIP-seq data analysis. 45

2.3 RIP-seq data analysis. 46

2.4 Additional Scatterplots and Gaussian mixture modeling plots. 47

2.5 Comparisons among all RIP-seq experiments, excluding ncRNAs. 48

2.6 Enrichment ratios of the consensus set of Sm-associated RNAs. 49

2.7 Three categories of Sm-associated RNAs in Drosophila and human. 51

2.8 Examples of the three categories of Sm-associated RNAs in *Drosophila* and human. . . 53

2.9 Characterization of the *Like-U* (*LU*) snRNA gene.. 57

2.9 Characterization of the *Like-U* (*LU*) snRNA gene.. 58

2.10 Additional characterization of LU snRNA. 59

2.11 Sequence alignment of *D. melanogaster* U1, U2, U4 and U5 paralogs. 61

2.12 Characterization of scaRNA:Prp8.. 62

2.12 Characterization of scaRNA:Prp8.. 64

2.13 Enrichment, structure, and phylogeny of SHAN scaRNA. 65
2.13 Enrichment, structure, and phylogeny of SHAN scaRNA. 66
2.14 CG4692 mRNA localizes along oocyte cortex. 67
2.15 Enrichment ratios for *Drosophila* and human replication-dependent histone mRNAs. 69
2.15 Enrichment ratios for *Drosophila* and human replication-dependent histone mRNAs. 70
2.16 snRNPs associate with mature mRNAs in S2 cells. 70
2.17 RNA-Sm association is cell type-specific and not due to re-assortment. 72
2.18 Sm proteins associate with mature mRNAs. 74
2.19 Analysis of the polyadenylation of Sm-associated mRNAs. 75
2.20 Sm-associated mRNAs are not TMG-capped. 75
2.21 U1 snRNP binds mature mRNAs. 77
2.21 U1 snRNP binds mature mRNAs. 78
2.22 Additional predicted snRNP-mRNA base pairings. 79

3.1 Vicinal pipeline and possible mechanisms for the generation of chimeric reads. 98
3.2 Vicinal analysis of known fly ncRNAs, snRNA:U1 (A-E) and RNaseP:RNA (F-L). 101
3.2 Vicinal analysis of known fly ncRNAs, snRNA:U1 (A-E) and RNaseP:RNA (F-L). 102
3.3 Vicinal analysis of snRNA:Like-U (LU) (A-D) and scaRNA:Prp8 (E-G). 105
3.3 Vicinal analysis of snRNA:Like-U (LU) (A-D) and scaRNA:Prp8 (E-G). 106
3.4 Vicinal analysis of additional fly ncRNAs. 113
3.4 Vicinal analysis of additional fly ncRNAs. 114
3.4 Vicinal analysis of additional fly ncRNAs. 115
3.4 Vicinal analysis of additional fly ncRNAs. 116
3.5 Vicinal analysis of mouse and human ncRNAs 117
3.5 Vicinal analysis of mouse and human ncRNAs 118

4.1 Alignment of *Drosophila* snRNA paralogs. 123
4.2 Alignment of mouse snRNA paralogs. 125
4.3 Expression profiles of snRNA isoforms during development. 128
4.3 Expression profiles of snRNA isoforms during development. 129

4.4 *Drosophila* U2 snRNA variations. 138

5.1 VFP-tagged Sm proteins are properly localized in U bodies in the female germline...153

5.2 Follicle cells contain U bodies. 154

5.3 Tgs1 is localized to U bodies. 156

5.4 U bodies appear in egg chambers around stage 5. 157

5.5 U bodies are non-essential passive snRNP transport particles. 158

5.6 SMN bodies are present only in primary spermatocytes..............................160

5.7 SMN bodies contain SMN complex, Sm proteins but not snRNPs. 161

5.8 SMN bodies are different from other RNP granules..................................162

5.9 SMN localization in SMN bodies is not affected by SMN mutations. 163

5.10 SMN localization in SMN bodies does not require RG box methylation..............164

5.11 U bodies are lost in *Ago2* mutant ovaries. 165

5.12 SMN bodies are lost in *Ago2*, but not affected in *r2d2* or *dcr-2* mutants.............165

A.1 Characterization of the KSm4 antibody 185

A.2 Localization of SmD3 in *Drosophila* egg chambers..................................186

A.3 Colocalization of anti-SmD3 and actin...187

B.1 Analysis of Lsm11 RIP-seq data. 189

B.2 Examples of Lsm11-associated RNAs. 190

C.1 Example mRNAs with altered expression in the *LU* mutant. 197

C.1 Example mRNAs with altered expression in the *LU* mutant. 198

C.1 Example mRNAs with altered expression in the *LU* mutant. 199

E.1 Defense party flyers made by Kelsey and Mike 201

LIST OF TABLES

1.1 Biochemical methodologies for studying DNA-RNA-Protein interactions. 31

2.1 Details about the RIP-seq and RIP-qRT-PCR experiments. 41

2.2 RIP-seq library statistics. 42

2.3 Mappable and unmappable read statistics in random hexamer primed libraries. 43

2.4 Comparison of oligo(dT) and random hexamer primed libraries. 49

2.5 Enrichment ratios of *Drosophila* Sm-associated RNAs 54

2.5 Enrichment ratios of *Drosophila* Sm-associated RNAs 55

2.6 Assignment of unique reads to *Drosophila* snRNA paralogs 56

2.7 Enrichment ratios of human Sm-associated RNAs. 63

2.8 List of primers and oligos used in this study. 87

3.1 RNA-seq datasets used in the study. 99

4.1 RNA-seq datasets used in this study. 121

4.2 *Drosophila* Sm class snRNAs. 132

5.1 Formation of U body like structures by blocking cytoplasmic snRNP assembly. 167

6.1 Available *Drosophila* snRNA mutants 181

B.1 List of RNAs associated with Lsm11. 192

B.1 List of RNAs associated with Lsm11. 193

CHAPTER 1: Introduction

1.1 From Central Dogma to RNA Regulation

The central dogma explains the information flow in all life forms between the three message-carrying polymers, DNA, RNA and protein (Crick 1958, Crick (1970)). Although an over-simplification, it is the unifying scaffold on which evolution works. The genetic information carried by these molecules provides the blueprint for making an organism. However, the information alone is not enough for producing the diversity of cell types and life forms. Many layers of regulation are required for the proper transfer of information at specific time points and in specific places. These regulatory mechanisms control DNA replication, RNA transcription, RNA processing, transport, degradation, protein translation, transport and degradation, and so on (Anko and Neugebauer, 2012).

As the hub of the central dogma and potentially the starting genetic material of life, RNA molecules are the most diverse and versatile. Not only can RNA function as the carrier of genetic information, but also they can catalyze chemical reactions like proteins do (see (Doudna and Cech, 2002) for a review). RNAs almost never exist as naked molecules inside the cell. Most of the time, RNAs are in complex with other RNAs, proteins, DNA, and small molecules. RNA-RNA interactions are commonly mediated by base paring, whereas protein-RNA interactions involve both RNA backbone and base interactions. RNA-containing complexes range from simple assemblies of a few molecules to RNP granules that are easily visible under light microscope, and these membraneless organized entities and membrane-bound organelles are the basis for subcellular compartmentalization that imposes order to the protoplasm. The components of the RNA containing complexes regulate many aspects of RNA metabolism, from production to destruction. The identification of molecular interactions is key to understanding the function and regulation of RNA.

RNA binding proteins are the most important regulators of RNA metabolism, and many of them are highly conserved, consistent with their co-evolution with their RNA partners. Higher eukaryotic genomes encode over 1100 RNA binding proteins as revealed by recent comprehensive experimental identifications (Castello *et al.*, 2012; Baltz *et al.*, 2012). These proteins not only are the basic components of RNP complexes, they integrate the biological signals and environmental stimuli to control RNA metabolism. The majority of RNA binding proteins contain characteristic globular modular RNA binding domains, including the RNA recognition motif (RRM), heterologous nuclear RNP K homology domain (KH), zinc finger (ZNF), etc. (Lunde *et al.*, 2007). A significant number of RBPs contain repetitive sequences enriched in several kinds of amino acids, including Gly, Arg, Lys, Tyr, Ser, and these amino acids typically form disordered motifs like RGG, YGG, RS and poly(K). There are also other RBPs, over 350 of them, that do not contain obvious structures that can be predicted to bind RNA (Castello *et al.*, 2012).

In addition to the RNA binding domains and motifs, RBPs also contain other kinds of functional domains involved in various cellular processes. One prominent example is the recent identification of dozens of intermediary metabolism enzymes as RNA binding proteins, suggesting a link between intermediary metabolism with RNA biology and posttranscriptional regulation (Castello *et al.*, 2012; Baltz *et al.*, 2012). A REM (RNA-Enzyme-Metabolite) hypothesis was recently proposed to explain the link (Hentze and Preiss, 2010). The more widely appreciated functions of RBPs include protein interaction, RNA catalysis, signal perception and localization.

Given the fundamental functions of RBPs, it is perhaps not surprising that many RBPs are absolutely essential for life and mutations in RBPs cause many severe human disorders, like various cancers, neurological disorders, and muscular atrophies (Lukong *et al.*, 2008; Castello *et al.*, 2013). For example, mutations in *FMR* genes cause fragile X mental retardation (FXS); mutations in or loss of *SMN1* cause spinal muscular atrophy (SMA), and mutations in *TDP-43* cause amyotrophic lateral sclerosis (ALS). Several types of common autoimmune diseases are caused by the generation of autoantibodies to RBPs. For example, some systemic lupus erythematosus (SLE) patients produce antibodies to Sm proteins (discussed in the next section); some cases of the paraneoplastic neurologic Hu syndrome are caused by autoantibodies to the

Hu protein; the Nova-1 RBP is an autoantigen in paraneoplastic opsoclonus myoclonus ataxia (POMA). Therefore, understanding the functions of RBPs is not only important for studying basic mechanisms of gene regulation, but also important for the treatment of many kinds of human diseases.

1.2 The Sm family of RNA binding proteins: evolution and features

The Sm proteins are a large family of RNA binding proteins conserved in all three domains of life (Valentin-Hansen *et al.*, 2004; Salgado-Garrido *et al.*, 1999), (Figure 1.1. Note: the Sm family is also called Lsm/LSm family, or Sm-Lsm-Hfq family, but I refer to it as the Sm family throughout the dissertation for consistency and simplicity). At least one copy of an Sm protein family gene is present in each of the sequenced organismal genomes, and the conservation is obvious on both the primary sequence and higher order structures (Figure 1.2). It is most likely that an Sm protein gene was present in the last universal ancestor of all life (Achsel *et al.*, 2001). This ancestral Sm protein gene is one of the around 60 genes common to all cellular life (Koonin, 2003; Anantharaman *et al.*, 2002). This gene was passed on to all descendants, the current extant species, modified by mutations, gene duplications and divergence, and occasionally horizontal gene transfers.

Members of the Sm family were initially discovered separately in bacteria and eukaryotes. The eukaryotic Sm proteins were first discovered in the 1960s. A patient named Stephanie Smith, who was diagnosed of systemic lupus erythmatosus (SLE), produced autoimmune antibodies that recognized a group of small nuclear proteins called Smith (Sm) antigens (Tan and Kunkel, 1966). The nature of these antigens was not known until the late 70s, when the Steitz lab found that these antigens were components of a set of small nuclear ribonucleoproteins complexes (snRNPs) (Lerner *et al.*, 1980). snRNPs have since been identified as the key components of the spliceosome (details will be discussed in the following sections).

In bacteria, the Hfq protein (also known as HF-I protein) was discovered in 1968 as an *E. coli* host factor that is required for the replication of bacteriophage Qbeta (Franze de Fernandez *et al.*, 1968). Studies since the 1990s showed that Hfq bind a class of small RNAs in bacteria to regulate their stability. In addition, it has been shown that Hfq is required for base pairing

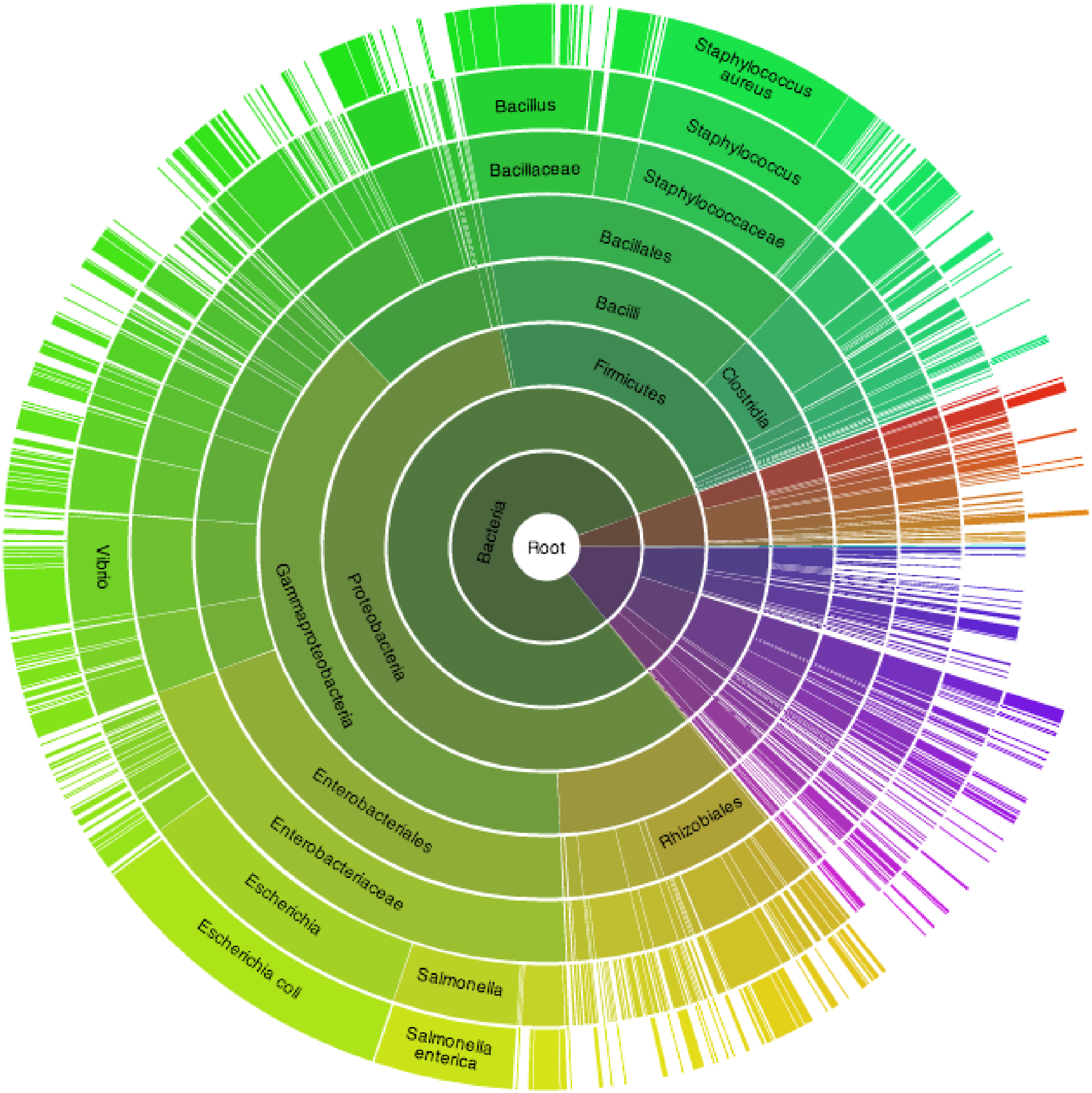


Figure 1.1: **Sunburst diagram of the distribution of Sm proteins.** Sm proteins are in all sequenced species on earth, at least one gene copy in each species. It is one of the around 60 proteins common to all cellular life on earch. The diagram was scaled by the number of species sequenced in each taxonomic rank. Red: archaea; green: bacteria; purple: eukarya. The Sm proteins are defined by their Sm domains (also known as LSM domain, PF01423, including additional LSM like domains). Data are extracted from http://pfam.sanger.ac.uk/.

interactions between small RNAs (sRNAs) and mRNAs (Zhang *et al.*, 2002, 1998; Moller *et al.*, 2002a). Hfq sRNP binding to mRNAs can regulate multiple aspects of RNA metabolism, including translation and degradation, etc. (Figures 1.3 and 1.4). However, the phylogenetic relationship between bacterial Hfq and eukaryotic Sm proteins was not known until early 2000, when several groups found that Hfq contains the conserved Sm secondary structures, forms a doughnut/toroid shaped complex, and has RNA binding specificity similar to eukaryotic Sm proteins (Moller *et al.*, 2002a; Zhang *et al.*, 2002) (Figure 1.2) (details will be discussed in the following sections). The completion of many genomes in three domains of life, and the earlier studies on eukaryotic and bacterial Sm like proteins finally led to the realization that the Sm proteins are a highly conserved family of RNA binding proteins.

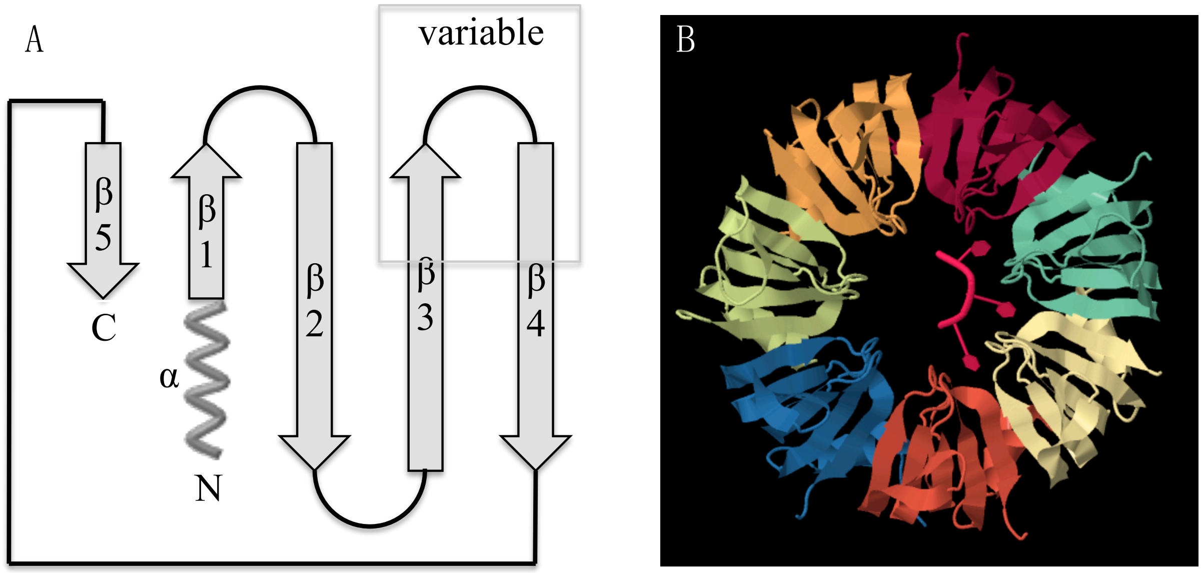


Figure 1.2: **Structure of the Sm proteins.** (A). secondary structure model of the Sm domain, showing the 1 alpha helix and 5 beta sheets. The hinge between beta sheets 3 and 4 are variable among Sm proteins in different species. (B). Crystal structure of an archaeal heptameric Sm protein (AF-Sm1) from Archaeoglobus fulgidus (PDB: 1I5L) complexed with short poly-U RNA. Each subunit of the homoheptamer is colored differently for better visualization (Toro *et al.*, 2001).

The Sm proteins are unique, without any structural similarity to any other known RNA binding domains, like RRM, KH and ZNF. Even though some Sm proteins do contain RG boxes that are known to be involved in RNA binding in other proteins, they are not essential for RNA binding in the context of known Sm proteins (Figure 1.5). Instead, the RNA binding capacity of Sm proteins comes from a toroid-shaped oligomer, and at least four regions on the ring can contribute to RNA binding in various contexts, the central hole, the proximal face, the distal face and the rim (Zhang *et al.*, 2013; Leung *et al.*, 2011; Pomeranz Krummel *et al.*, 2009;

Kambach *et al.*, 1999). The binding mechanisms differ among the various Sm rings. In bacteria, all four regions are known to contribute to RNA binding, with the central hole, distal face and the rim mainly contributing to small RNA binding, while the distal face mainly contributing to mRNA binding. In archaea and eukaryotes, the main region for RNA binding is the central hole (Leung *et al.*, 2011; Pomeranz Krummel *et al.*, 2009; Toro *et al.*, 2002; Urlaub *et al.*, 2001) (it is still a question since less is known about archaeal Sm proteins). Figure 1.2B shows the structure of an archaeal Sm homoheptamer complex together with an oligoU RNA binding to the central hole. The interaction shown in this figure is primarily mediated by stacking interactions between the RNA bases and the amino acid residues.

Sm protein complexes exist in different flavors, either as homo-oligomers or hetero-oligomers (Figure 1.3). Most bacterial and archaeal genomes encode one Sm protein, and only a few of them encode two or three Sm proteins. In most of the cases that have been studied, these Sm proteins form homohexamers and homoheptamers. After divergence of the archaea-eukarya lineage, the eukaryotic Sm protein genes have undergone extensive duplication and divergence. Most eukaryotic genomes encode more than 10, or even 20 distinct Sm class proteins (for example in *Drosophila melanogaster*, Figure 1.5). Most of these eukaryotic Sm proteins form several different rings, while some of them are not known to form ring shaped complexes (Lsm12, Lsm14, Lsm16 and Ataxin-2 in Figure 1.5) (see detailed descriptions of each of these complexes in the following sections).

1.3 RNA partners and functions of the bacterial Sm protein Hfq

Functional characterizations of eukaryotic Sm proteins have mostly been focused on the best-studied complexes and functions, like the canonical Sm ring and Lsm2-8 ring in spliceosomal snRNPs, the Lsm10-11 ring in U7 snRNP and the Lsm1-7 ring in mRNA degradation. This is not a surprising situation, considering the Matthew effect, where more people study the well-studied problems because we know the importance of these problems. Here I briefly review the functions of bacterial Hfq. Given the evolutionary conservation of Sm proteins and diverse categories of RNA partners of eukaryotic Sm proteins, it is very likely that the eukaryotic Sm proteins and Sm-containing complexes have more unknown functions, and lessons learned from

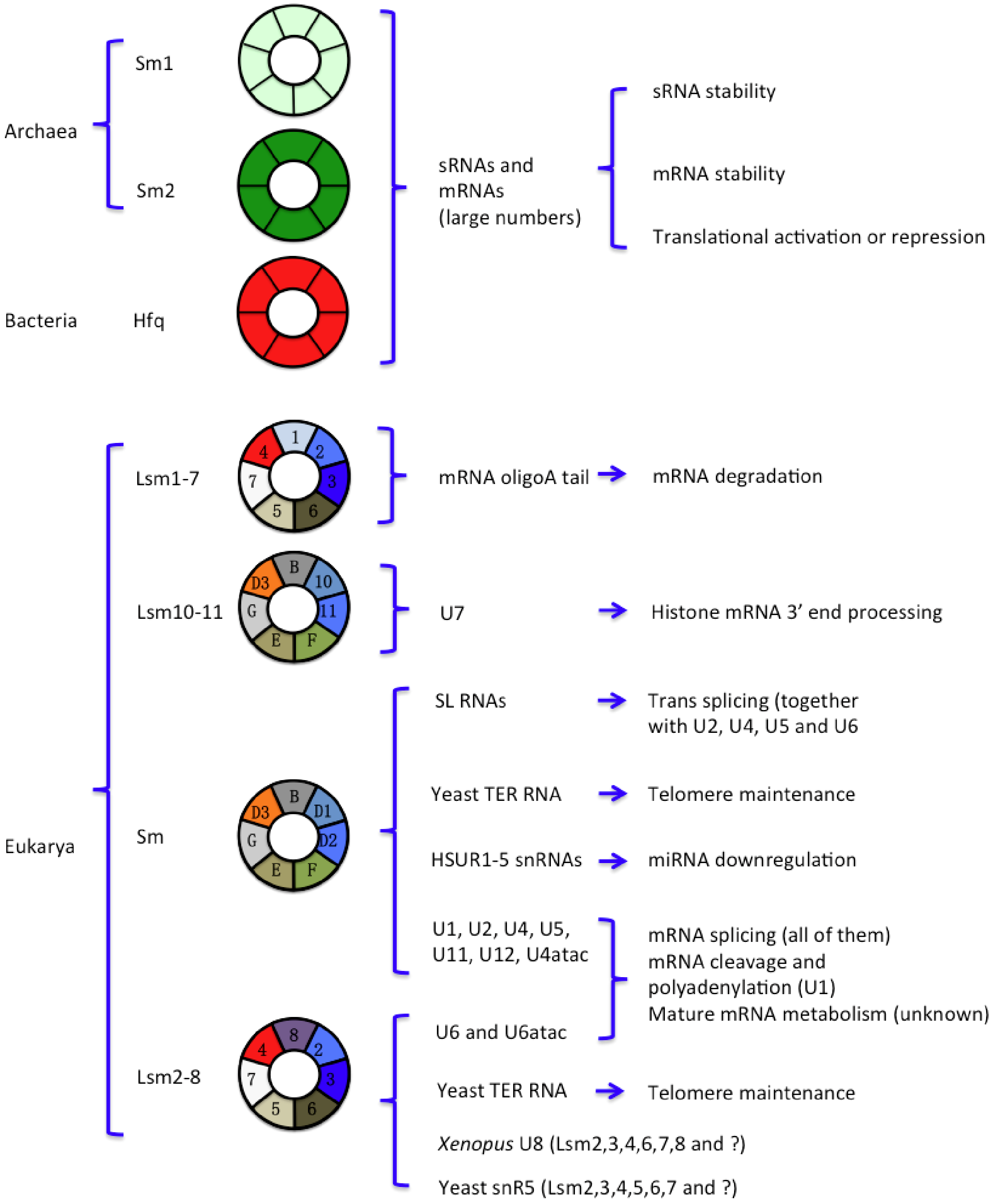

Archaea
Sm1
Sm2
Bacteria
Hfq
sRNAs and mRNAs (large numbers)
sRNA stability
mRNA stability
Translational activation or repression
Eukarya
Lsm1-7
1
2
3
4
5
6
7
mRNA oligoA tail
mRNA degradation
Lsm10-11
B
10
11
D3
G
E
F
U7
Histone mRNA 3' end processing
Sm
B
D1
D2
D3
G
E
F
SL RNAs
Trans splicing (together with U2, U4, U5 and U6
Yeast TER RNA
Telomere maintenance
HSUR1-5 snRNAs
miRNA downregulation
U1, U2, U4, U5, U11, U12, U4atac
mRNA splicing (all of them)
mRNA cleavage and polyadenylation (U1)
Mature mRNA metabolism (unknown)
Lsm2-8
8
2
3
4
5
6
7
U6 and U6atac
Yeast TER RNA
Telomere maintenance
Xenopus U8 (Lsm2,3,4,6,7,8 and ?)
Yeast snR5 (Lsm2,3,4,5,6,7 and ?)

Figure 1.3: **Diversity of main Sm rings in all three domains of life.** The RNAs bound by these Sm rings and the functions of Sm-containing RNP complexes are summarized on the right. Other Sm rings that are not well studied are not presented here. One of them is the archaeal SmAP3, which forms a complex of 14 subunits (Mura *et al.*, 2003). Some archaeal Sm pentamers and octamers have also been observed (Mura *et al.*, 2013). Variants of eukaryotic Sm rings also exist, for example, SmD3 is replaced by SSm4 in trypanosome U4 snRNP (Jae *et al.*, 2010).

bacterial Hfq can be applied to studies on eukaryotic Sm proteins (Figure 1.4).

Many well-studied bacterial species encode dozens and maybe hundreds of Hfq-associated sRNAs 50-250 nt long (for example, (Zhang *et al.*, 2003; Sittka *et al.*, 2008)). The sRNAs usually have one to a few stem loops. The wide variety of sRNAs regulate many different subsets of mRNAs through base pairing. Some of the prominent cellular processes regulated by Hfq-sRNA complexes include quorum sensing (Lenz *et al.*, 2004; Bardill *et al.*, 2011), synthesis of outer membrane proteins (Vogel and Papenfort, 2006; Song and Wai, 2009), virulence (Vogel, 2009), response to a variety of cellular stresses, like osmotic stress, cold shock, iron depletion, SOS response, sugar stress and nitrogen deprivation (Benjamin *et al.*, 2010; Ionescu *et al.*, 2010; Repoila *et al.*, 2003; Delihas and Forst, 2001).

At least five different kinds of functions have been described for the bacterial Sm protein Hfq on the molecular level, and the functions are usually related to translation and RNA stability (Vogel and Luisi, 2011) (Figure 1.4). Hfq-sRNA complexes could base pair with mRNAs close to or on the ribosome binding sites (RBS), and this association prevents ribosomes from binding and therefore represses translation (Figure 1.4A) (Mizuno *et al.*, 1984; Bouvier *et al.*, 2008; Chen *et al.*, 2004; Udekwu *et al.*, 2005; Argaman and Altuvia, 2000; Huntzinger *et al.*, 2005; Maki *et al.*, 2008; Moller *et al.*, 2002b). On the other hand, mRNAs by themselves could form stable secondary structures that mask the RBS, whereas Hfq-sRNA binding would remodel the secondary structures to allow ribosome binding (Figure 1.4B) (Wang *et al.*, 2013; Ruiz and Silhavy, 2003; Brescia *et al.*, 2003).

Many of the Hfq-associated sRNAs are not stable by themselves, and Hfq binding could stabilize them (Figure 1.4C) (Masse *et al.*, 2003; Folichon *et al.*, 2003; Moll *et al.*, 2003). This is similar to the stabilization effect of Sm proteins on spliceosomal snRNAs in eukaryotes (Roy *et al.*, 1995; Noble and Guthrie, 1996; Bordonne and Tarassov, 1996; Rymond, 1993). Hfq-sRNA binding to the mRNAs, either at the RBS or coding region, could recruit RNaseE to degrade mRNAs (Figure 1.4D) (Masse *et al.*, 2003; Huntzinger *et al.*, 2005; Afonyushkin *et al.*, 2005; Vogel *et al.*, 2004). The Hfq ring can bind some mRNAs directly to recruit poly(A) polymerases to add adenylate tails to mRNAs. The added poly(A) tails could further induce exonuclease-dependent degradation (Figure 1.4E) (Hajnsdorf and Regnier, 2000; Mohanty *et al.*,

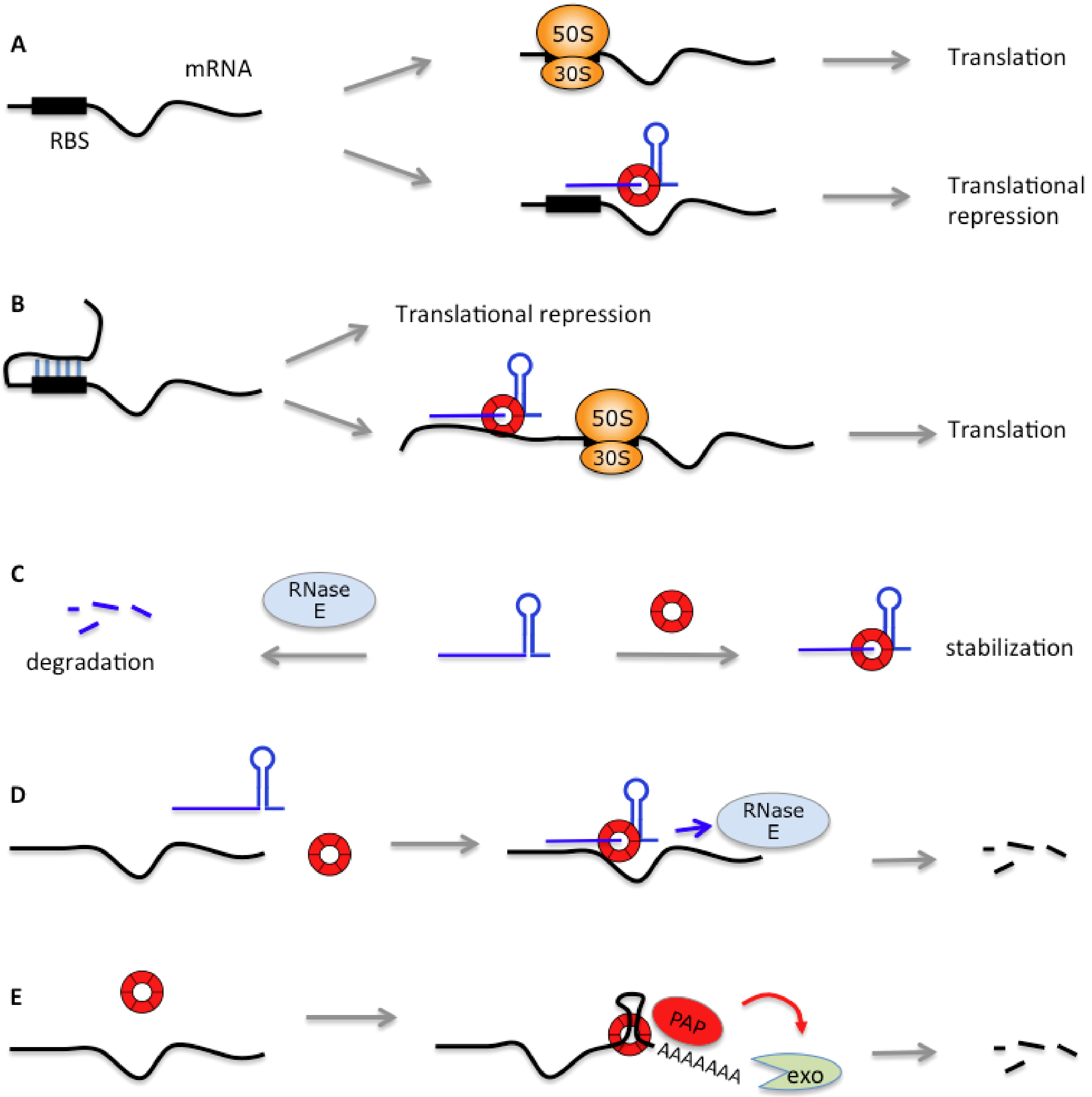


Figure 1.4: **Well known functions of the bacterial Sm protein Hfq.** (A). Hfq sRNP binding to RBS (ribosomal binding site) could prevent ribosome binding to mRNAs and therefore repress translation. (B). The RBS of certain mRNAs are masked by local secondary structures, which could be opened by Hfq sRNPs for translation. (C). Naked small RNAs are prone to degradation by RNase E, while Hfq binding stabilizes small RNAs. (D). Hfq sRNPs binding to mRNAs could recruit RNase E to degrade the bound mRNAs. (E) Hfq could directly bind mRNAs to recruit poly(A) polymerase to polyadenylate mRNAs, which are then degraded by exonucleases (exo) from the 3' end. The exonucleases could be polynucleotide phosphorylase, RNase R or RNase II.

2004; Regnier and Hajnsdorf, 2013).

1.4 RNA partners and functions of the archaeal Sm proteins

Compared to the many studies performed on bacterial and eukaryotic Sm proteins, much less is done on the archaeal Sm proteins. However several unexpected Sm rings were discovered in archaea, suggesting further complexity of Sm rings (Figure 1.3). Typical archaeal genomes encode 1-3 Sm proteins and they form homo-oligomeric rings. A recent study showed that the Sm protein in an archaeon *Haloferax volcanii* forms homoheptameric ring and this ring binds numerous small RNAs and C/D box snoRNAs (Fischer *et al.*, 2010). This study suggests that archaeal Sm proteins may function in a similar way to the bacterial Hfq.

1.5 Overview of known RNA partners and functions of the eukaryotic Sm proteins

Compared to prokaryotes, eukaryotes are much more complicated at both the molecular level and the organismal level. Concomitant with the evolution of eukaryotes, the number of Sm proteins in eukaryotes has also greatly increased through gene duplication and functional divergence. Most eukaryotic genomes encode more than 10 distinct Sm proteins and some of these Sm proteins form several different rings (Figure 1.3). The canonical Sm and Lsm proteins form at least four types of rings: the canonical Sm ring, the Lsm1-7 ring, the Lsm2-8 ring and the Lsm10-11 ring.

The canonical Sm ring consists of SmB, D1, D2, D3, E, F and G, which are the founding members of the Sm family (Figures 1.3 and 1.5). These proteins are all very small, with the biggest one SmB less than 30kD. The smallest ones are less than 10kD, for example SmE, SmF and SmG. All the Sm proteins have the Sm domain; in addition, three of the canonical Sm proteins, SmB, SmD1 and SmD3 also contain an RG box at the C terminus with varying copies of the RG dipeptides. The RG box is known to have RNA-binding activity in many proteins, but it does not seem to contribute to RNA binding in the canonical Sm ring. The Lsm4 protein, which is a component of the Lsm1-7 and Lsm2-8 rings, also contains a RG box.

The RG box is present in many other RNA binding proteins, for example: FMRP, Coilin etc., and the RG box may in some cases interact with the Tudor domain of other proteins. The arginine residues in RG boxes are frequently symmetrically dimethylated (sDMA) by protein arginine methyltransferases (PRMTs). These modifications can modulate the interaction between RG box and the Tudor domain, or RG box and RNA (Brahms *et al.*, 2001). For example, the methylation of RG boxes in Coilin and Sm proteins promotes interaction with SMN (the effect may be minor in certain cases), whereas methylation of the RG box in FMRP affects its binding to RNA (Blackwell *et al.*, 2010; Hebert *et al.*, 2002).

The canonical Sm ring binds several different kinds of small RNAs directly via the Sm site, a consensus RNA sequence in the form of RRUUUUURR, where R stands for A or G (Figure 1.6). The Sm site sometimes has variations, which makes *de novo* prediction of binding partners essentially impossible. The Sm RNA motif threads through the central hole of the Sm ring, with seven of the bases contacting each of the seven Sm proteins through stacking interactions (Leung *et al.*, 2011; Pomeranz Krummel *et al.*, 2009). The canonical Sm ring associates with several types of small RNAs directly, for example, the spliceosomal snRNAs (U1, U2, U4, U5, U11, U12 and U4atac), trans-splicing spliced leader (SL) snRNAs, yeast telomerase RNA, and the *Herpesvirus saimiri* U RNAs (HSURs) (Figure 1.3). A more detailed description of the structure and function of spliceosomal snRNAs and snRNPs will be presented in the following sections.

The SL snRNAs is the trans-spliceosomal counterpart of U1 snRNA, except that SL RNAs are consumed by the splicing reaction and become part of the mature mRNA. SL RNAs exist in a set of remotely related animal clades, including urochordates, nematodes, flatworms, and hydra, as well as in Euglenozoa and dinoflagellates (Derelle *et al.*, 2010). The trans-splicing reaction transfers part of the SL RNA, including the TMG cap, to the mRNA, and this additional sequence on the mRNA can promote translation (Lall *et al.*, 2004).

HSURs were discovered in the 1980s by the Steitz lab (Lee *et al.*, 1988; Wassarman *et al.*, 1989; Albrecht and Fleckenstein, 1992). HSURs are transcribed by pol II with promoters similar to the U snRNAs, and they contain the typical TMG cap, two stem-loops flanking the canonical Sm site. However, little was known about their functions until a few years ago, when Cazalla et al. showed that one of the HSURs could down-regulate a host miRNA and therefore manipulate

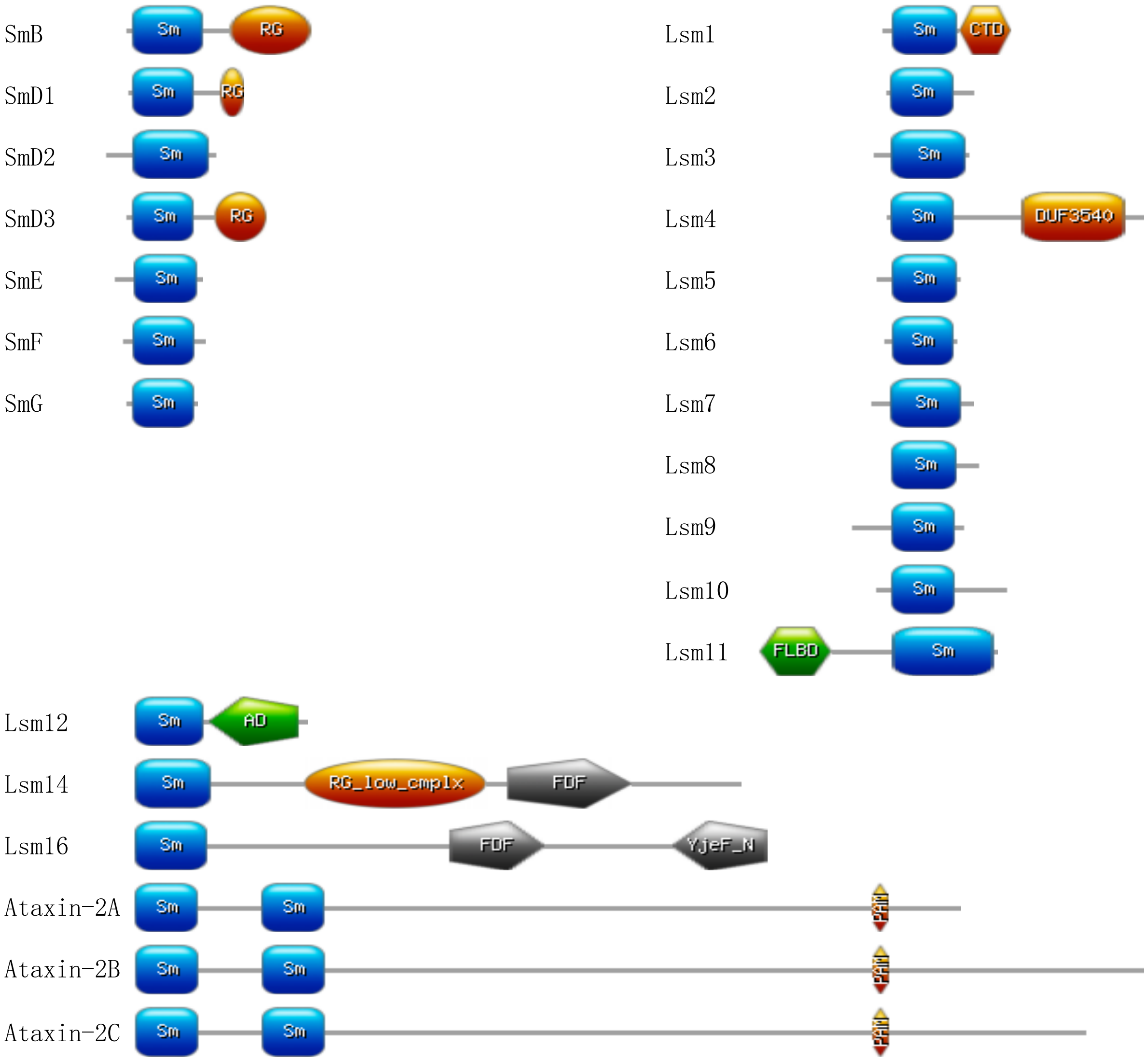


Figure 1.5: **Diagram of domains of all fruitfly Sm class proteins.** Sm domains in Ataxin-2 protein isoforms are sub-classified as Sm-ATX (first one) and LsmAD (second) domains in Pfam, but we refer to them generally as Sm domains. RG: arginine-glycine rich domain; DUF3540, domain of unknown function (recently Lyons et al. showed that this region is required for interaction with SLBP and 3'hExo); FLBD: FLASH binding domain; RG_low_cmplx: RG motifs scattered in a low complexity region. PAM: PolyA binding motif.

host cell gene expression (Cazalla *et al.*, 2010).

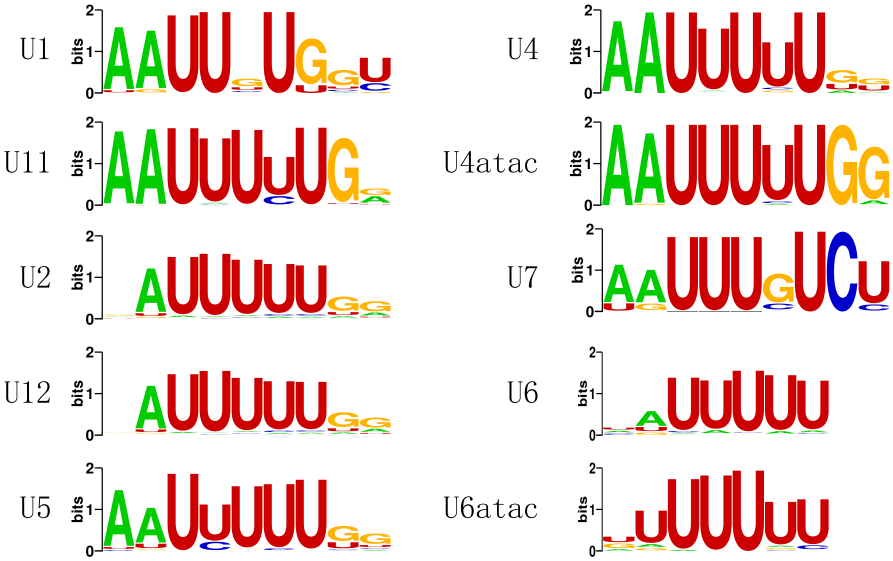


Figure 1.6: **Consensus Sm sites.** Consensus Sm sites were calculated from the seed collection of snRNAs in Rfam (http://rfam.sanger.ac.uk/) using the Weblogo application (http://weblogo.berkeley.edu/logo.cgi) (Burge *et al.*, 2013; Schneider and Stephens, 1990; Crooks *et al.*, 2004)). Number of snRNAs used in the calculation are as follows. U1: 100, U11: 72, U2: 208, U12: 208, U5: 180, U4: 171, U4atac: 61, U7: 56, U6: 188, U6atac: 62.

Yeast telomerase RNA (TER) is unique among the telomerase RNAs in different species and among the Sm-associated RNAs. In many yeast species, but not other more distantly related species, the telomerase RNA contains an Sm site and can be bound by the canonical Sm ring directly (Seto *et al.*, 1999; Leonardi *et al.*, 2008). More interestingly, this binding is required for the cleavage of the TER precursor by the spliceosome to generate mature TER RNA. After the cleavage, the Sm ring falls off the 3' end of the mature TER RNA and is replaced by the Lsm2-8 ring, where the Lsm2-8 ring protects the 3' end of the mature TER RNA and recruit the protein subunits of the telomerase. SmB and SmD3 have been shown to interact with several human snoRNAs directly, however, it is still controversial as to how they interact with snoRNAs, and it is not known whether other canonical Sm proteins also interact with these snoRNAs (Fu and Collins, 2006).

Substitution of SmD1 and SmD2 by Lsm10 and Lsm11 in the canonical Sm ring generates the Lsm10-11 ring. This ring specifically binds the U7 snRNA to form U7 snRNP. A single stranded region in the U7 snRNP base pairs with the histone downstream element (HDE) of the pre-mRNAs of replication-dependent histones, which are probably the only class of eukaryotic mRNAs that do not contain a poly(A) tail (Mowry and Steitz, 1987; Spycher *et al.*, 1994; Williams and Marzluff, 1995). Instead, replication-dependent histone mRNAs end with a stemloop structure, which is bound by the stemloop binding protein (SLBP) and an exonuclease 3'hExo. U7 snRNP binding to histone pre-mRNAs further recruits other mRNA cleavage and processing factors to cut upstream of the HDE to produce mature histone mRNAs. The unique Lsm10 and Lsm 11 proteins play important roles in binding some of the histone processing factors likes FLASH (Yang *et al.*, 2009; Burch *et al.*, 2011).

The best-known RNA partners of the Lsm2-8 ring are the U6 and U6atac spliceosomal snRNAs. The Lsm2-8 ring (in the order Lsm3-2-8-4-7-5-6) binds the 3' end oligo-U sequence motif, which is different from the other spliceosomal snRNAs (Figures 1.3 and 1.6) (Achsel *et al.*, 1999; Mayes *et al.*, 1999; Vidal *et al.*, 1999; Zhou *et al.*, 2014). The four uridine nucleotides at the 3' end are recognized modularly by the Lsm3, Lsm2, Lsm8 and Lsm4 proteins, respectively. However, it is not known why these two snRNAs require a different set of Sm proteins. Several cases of poorly characterized Lsm rings were reported that might be Lsm2-8 ring. Tomasevic et al. reported that in *Xenopus*, a ring that contains at least Lsm2, 3, 4, 6, 7 and 8 bind U8 snoRNA (Tomasevic and Peculis, 2002). In yeast *S. cerevisiae*, a Lsm2-7 ring binds the snoRNA snR5, however, it is not known whether Lsm1 or Lsm8 is also in this complex (Fernandez *et al.*, 2004).

A single subunit substitution of the Lsm2-8 ring, where Lsm1 replaces Lsm8, generates a different complex, Lsm1-7, with drastically different target specificity and function. Lsm1-7 ring binds the oligoadenylated mRNAs at the 3' end oligoA tails, and the binding promotes decapping of the mRNAs and subsequent 5'-3' degradation. The Lsm1-7 complex has also been shown to bind the oligoU tail of replication-dependent histone mRNAs for rapid degradation (Lyons *et al.*, 2014).

While most of the Sm proteins contain only the Sm domain and some of them contain the RG box, a few of them contain extra domains and they are not known to be involved in Sm ring formation (Figure 1.5) (Albrecht and Lengauer, 2004). Compared to the better known canonical Sm proteins and the Lsm1-8 and Lsm10-11 Lsm proteins, these Sm proteins are usually much bigger and contain extra domains and have functions, some of which are not well understood (Figure 1.5). Specifically, Lsm14, also known as Lsm15 and Trailer hitch, is part of a large RNP complex that also contains Me31B and Cup. The Lsm14 complex binds several mRNAs and localizes to ER exit sites to regulate protein trafficking (Wilhelm *et al.*, 2005). Lsm16, also known as enhancer of decapping protein 3 (EDC3), can stimulate mRNA decapping and degradation (Kshirsagar and Parker, 2004; Fenger-Gron *et al.*, 2005; Tritschler *et al.*, 2007). Ataxin-2 is a unique Sm-domain containing protein in that it contains two Sm domains. Interestingly, mutations in human ATXN2 cause splinocerebellar ataxia type2 (SCA2). It has been shown recently that Ataxin-2 functions together with FMRP to regulate neuronal mRNA translation and long-term olfactory habituation (Sudhakaran *et al.*, 2014).

1.6 Structure and composition of Sm-class snRNPs

The composition of eukaryotic Sm-class snRNPs have been studied extensively by several groups, and now we know almost all the components of each of the complexes (Will and Luhrmann, 2011; Matera and Wang, 2014). Here I will briefly review what is known about the most common Sm-class snRNPs, the canonical Sm ring containing snRNPs, the Lsm2-8 bound U6/U6atac snRNPs and the Lsm10-11 bound U7 snRNP (Figure 1.7).

All the single snRNPs are organized in a similar manner. Each snRNP contains an snRNA, an Sm ring, and several snRNP-specific proteins. The compositions of these complexes are summarized in Figure 1.7. Even though some snRNPs exist as stable single snRNPs in vivo, like U1, U2, U5, U11, U12 and U7, other exist as di- or tri-snRNPs in certain situations or exclusively. For example, the extensive base pairing interactions for U4/U6 or U4atac/U6atac made them exist almost exclusively as di-snRNPs in vivo. U11 and U12 snRNPs can also exist as di-snRNPs prior to assembly of the minor spliceosome. U4/U6 and U4atac/U6atac di-snRNPs can interact with U5 snRNP via protein-protein interactions and exist as stable tri-snRNPs in vivo. These di- and tri-snRNPs can be purified and their compositions determined. Dynamic inter-snRNP interactions also exist among other combinations of snRNPs during spliceosome assembly, but most of them are more transient and hard to capture by native purification without crosslinking. Several studies also reported the existence of penta-snRNPs,

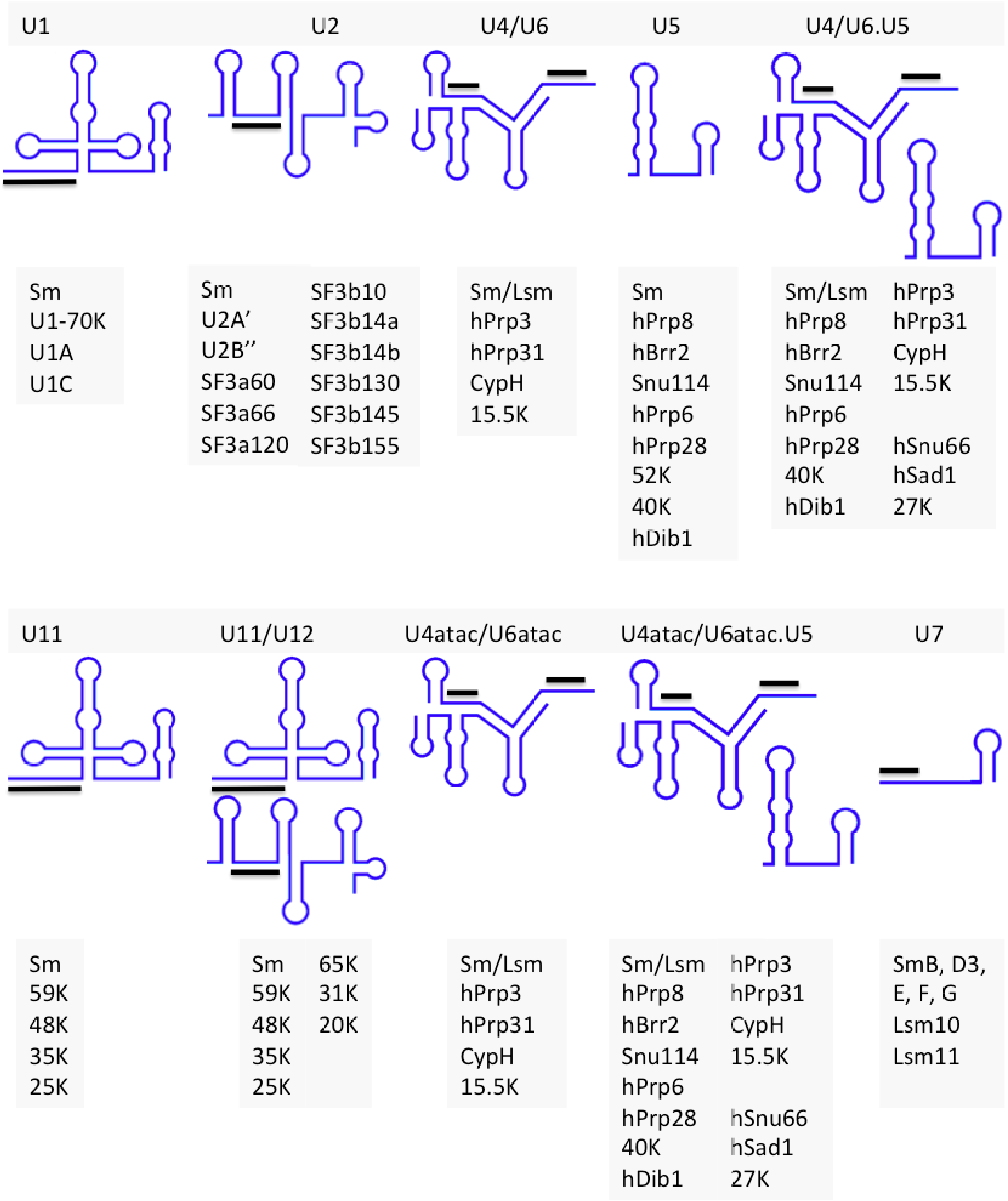


Figure 1.7: **Secondary structure and core protein composition of all spliceosomal and U7 snRNPs.** The thick black lines represent regions that could base-pair with other RNAs, not including regions that participate in intramolecular or between-snRNA base pairings.

suggesting an alternative spliceosome assembly pathway that is different from the step-wise assembly pathway (details about the spliceosome assembly pathway will be presented in the following sections) (Stevens *et al.*, 2002).

The Sm-class snRNPs contain unique 5' cap structures. The pol II transcribed snRNAs (U1, U2, U4, U5, U11, U12, U4atac and U7) all have the same typical trimethyl-guanosine (TMG) cap. This cap comes from hypermethylation of m7G cap. The hypermethylation of m7G cap is catalyzed by trimethylguanosine synthase 1 (Tgs1) in the cytoplasm. The TMG cap confers specific protein binding activities. For example snurportin binds TMG cap to promote snRNP import into the nucleus (reference). U6 and U6atac snRNPs are transcribed by pol III and have a unique non-nucleotide cap, gamma-monomethyl-phosphate.

As expected for macromolecular complexes, extensive intermolecular contacts exist among the snRNP components. Major progress has been made recently in understanding some aspects of these interactions, however, most of them remain elusive up to now. The most salient feature in snRNP structure and composition is the existence of the Sm sites, which dictates specific Sm ring binding (Figure 1.6). The Sm site bound by the canonical Sm ring usually contains two purines at each end and five consecutive uredines in the middle. The Sm site bound by Lsm2-8 ring in U6 and U6atac snRNAs is usually a stretch of 4-5 uridines at the 3' end of the snRNAs. The Lsm10-11 ring binds a variant of the canonical Sm site. The Sm rings and the snRNAs form the core of the snRNPs, on which other components assemble. However, this does not indicate the Sm rings are always loaded the first. The order of the assembly is not entirely known (see the next section 'snRNP assembly in eukaryotes' for more details).

Recent studies using crystallography have started to reveal the detailed organizations of the individual snRNPs. Kiyoshi Nagai's lab has generated crystal structures of U1 and U4 snRNPs that contain the snRNAs (partial), canonical Sm ring and, for U1, snRNP-specific proteins (Leung *et al.*, 2011; Pomeranz Krummel *et al.*, 2009). Lin and Xu have recently solved the crystal structure of the SF3b complex, part of the U2 snRNP (Lin and Xu, 2012). Crystallization of large macromolecular complexes, like the spliceosome, is very difficult. Instead, the Luhrmann lab have used cryo-electron microscopy to study the structure of higher order snRNP complexes and spliceosomes at different stages of splicing (for review see Stark and Luhrmann (2006)).

Both the RNA sequences and the proteins mediate the interactions among components in the snRNPs and with mRNAs. Compared to the proteins, more is known about how the RNA sequences bind their targets (the black lines in Figure 1.7). U1 and U11 snRNP 5' splice site recognition sequence can base pair with the 5' splice site of pre-mRNAs. Specific sequences in U2 and U12 snRNA base pair with the branch point sequence (BPS) in the introns of pre-mRNAs. Besides the sequences in U4/U6 and U4atac/U6atac RNAs that mediate base pairing interactions between the di-snRNPs, there are also single stranded regions that could potentially base pair with other RNA species. U5 snRNA has an invariant loop in the 5' end stemloop that base pairs with the exon-intron junction during spliceosome remodeling. All these sequence elements can also be utilized to base pair with other kinds of RNAs and regulate other aspects of RNA metabolism. (See Figure 1.8 for the known base-pairing interactions among snRNPs and with pre-mRNAs. See also Chapter 2 for description of the new mode of snRNP-mRNA interaction involving the U1 5' splice site recognition sequence and Chapter 6 for perspectives on how these base pairing regions can be used to bind different subsets of mRNAs to regulate their metabolism)

In addition to the sequence and protein components of the snRNPs, snRNAs also contain large numbers of RNA modifications and the functions of these modifications are not well known (see review by Karijolich and Yu (2010)). Besides the TMG and gamma-monomethyl-guanosine caps, the well known modifications include pseudouridylation and 2'-O-methylation, which are guided by box H/ACA and box C/D scaRNPs, respectively.

1.7 snRNP assembly in eukaryotes

For small RNAs that are usually shorter than 300 nucleotides, the biogenesis of snRNAs and snRNPs is surprisingly complicated (Note that in many yeast species, including *S. pombe* and *S. cerevisiae*, U2 snRNA can be over 1kb due to an hypervariable insertion in the middle of the RNA). Assembly of most of the known spliceosomal snRNPs (U1, U2, U4, U5, U11, U12 and U4atac) and U7 snRNP can be divided into three phases, the first nuclear phase, the cytoplasmic phase and the second nuclear phase (Figure 1.8). The assembly of U6 and U6atac snRNPs occurs entirely inside the nucleus (Hamm and Mattaj, 1989; Vankan *et al.*, 1990; Terns *et al.*, 1993; Boelens *et al.*, 1995; Pante *et al.*, 1997; Spiller *et al.*, 2007).

After pol II type snRNAs are transcribed by pol II in the nucleus they are first m7G capped and then their 3' ends processed by the integrator complex (Figure 1.8A). After preliminary processing, snRNAs are bound by the cap binding complex (CBC), Ars2 and Phax (phosphorylated adapter for snRNA export) (Ohno *et al.*, 2000; Hallais *et al.*, 2013). Certain snRNP-specific proteins associate with the snRNAs prior to export, for example, U1A, while others are assembled onto snRNPs in the cytoplasm or after reimport into the nucleus (Terns *et al.*, 1993; Kambach and Mattaj, 1994). The pre-export complex transits through the Cajal body and is exported to the cytoplasm with the help of Crm1 and Ran-GTP (Suzuki *et al.*, 2010; Ohno *et al.*, 2000).

The Cajal body, initially discovered by Santiago Ramon y Cajal over 100 years ago, also known as the nucleolar accessory body or coiled body, is a unique subnuclear RNP granule present in many proliferative cells and neurons (see review in Gall (2000); Morris *et al.* (2008)). The Cajal body has been implicated in assembly and maturation of many RNP complexes, including the snRNPs, RNA polymerase, telomerase etc. (Darzacq *et al.*, 2002). Cajal bodies are molecularly defined by the presence of coilin, and in most cases are associated with actively transcribing snRNA genes (Matera, 1999). In addition to Coilin and snRNPs, Cajal bodies also contain the SMN complex, snoRNPs (small nucleolar RNPs) and scaRNPs (small Cajal-body specific RNPs, including the telomerase complex). In some cell types (e.g. HeLa cells), SMN and some of its associated proteins form snRNP-free granules tightly associated with Cajal bodies, and are called gems (for Gemini of the Cajal bodies) (Carvalho *et al.*, 1999; Matera, 1999). Histone processing factors such as U7 snRNP, FLASH etc. often form another kind of nuclear RNP granule at the histone loci called the histone locus body (HLB). HLB and Cajal bodies sometimes are very close to each other, or completely overlap (for review see Nizami *et al.* (2010); Matera *et al.* (2009)). Both the outbound and inbound routes of snRNP transport go through the Cajal bodies, however, it is not entirely clear how Cajal bodies affect snRNP assembly. The widely accepted theory is that concentrating factors in a small volume accelerates chemical reactions, which in this case, results in more efficient snRNP assembly.

After pre-mature snRNPs are exported to the cytoplasm, several steps of maturation and

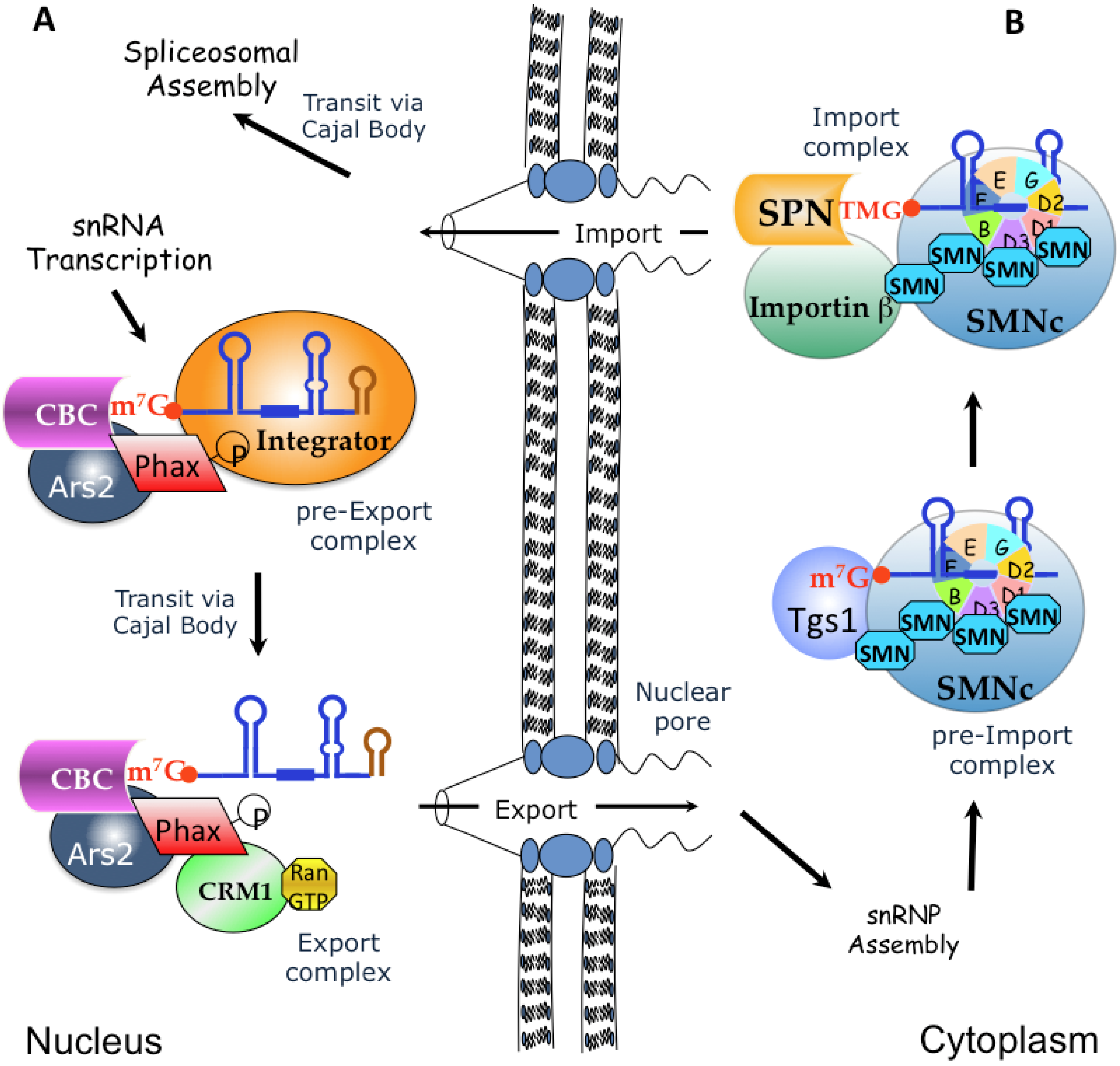


Figure 1.8: **The biogenesis pathway of spliceosomal snRNPs** (excluding U6 and U6atac, which are assembled through a different pathway localized in the nucleus). (A). The nuclear phase of snRNP assembly, from snRNA transcription to the assembly of pre-export and export complexes. (B). The cytoplasmic phase of snRNP assembly, where SMN assembles Sm proteins onto snRNAs, Tgs1 hypermethylates the m7G cap. After re-import, snRNPs are assembled into spliceosomes. CBC: cap binding complex. Phax: phosphorylated adapter for RNA export. CRM1: chromosome region maintenance 1, also known as exporting or Xpo1. Tgs1: trimethylguanosine synthase 1. SMN: survival of motor neuron. SMNc: SMN complex. SPN: snurportin. Adapted from Matera and Wang 2014.

assembly ensue (Figure 1.8B), and these steps are catalyzed and coordinated by the SMN complex (SMNc) (Fischer *et al.*, 1997; Massenet *et al.*, 2002; Pellizzoni *et al.*, 2002). The major proteins assembled onto the snRNAs are the Sm proteins. Three of the seven canonical Sm proteins, as mentioned above, are symmetrically dimethylated at the arginine residues (sDMA) in the RG boxes (Brahms *et al.*, 2001). sDMA modification is required for efficient assembly of snRNPs by the SMN complex in human cells, but not in *Drosophila* cells. The methylation of Sm proteins is catalyzed mainly by the type II protein arginine methyltransferase (PRMTs) PRMT5, together with other proteins, pICln and WDR77/MEP50.

The SMN complex consists of multiple copies of SMN, Gemin2-8 and Unrip. Mutations and loss of the human *SMN1* gene are known to cause a severe human disease, SMA (Spinal Muscular Atrophy). The SMN protein can oligomerize to form the scaffold for the whole SMN complex that contains other proteins. The best-studied functions of the SMN complex is in the assembly of the snRNPs. Recent crystallographic studies have started to reveal certain details of the assembly pathway. Gemin2, a conserved member of the SMN complex binds directly to five of the seven Sm proteins, the SmD1-SmD2-SmF-SmE-SmG pentamer and holds them in a semistable state for subsequent snRNA loading and ring closure (Zhang *et al.*, 2011). The chaperon protein pICln, which is a component of the PRMT5 complex, mimics SmB-SmD3 dimer structure in vivo that stabilizes the pentamer before Gemin2 binding (Grimm *et al.*, 2013; Chari *et al.*, 2008). The Tudor domain of SMN contains an Sm fold, and is thought to also have a SmB-SmD3 mimetic role during Sm core assembly (Grimm *et al.*, 2013).

After Sm proteins are assembled onto snRNAs, the m7G cap is hypermethylated to TMG by Tgs1, which is recruited by the SMN complex. Sm core assembly and TMG capping are two important signals for snRNP import back into the nuclei. snRNP import is mainly mediated by the snRNP import adapter Snurportin, which binds the TMG cap directly. Importin beta (Moleskin in flies) are the import acceptors for the Snurportin-snRNP complex. Interestingly, in addition to the nuclear Cajal bodies where many steps of snRNP assembly take place, cytoplasmic RNP granules call U bodies, have also been discovered that are related to snRNP assembly (Liu and Gall, 2007). However, little is known about which steps are organized in U bodies, and whether U bodies actually facilitate snRNP assembly (see Chapter 5 for more results and discussion).

After snRNPs are imported into the nucleus together with the SMN complex, they localize temporarily to the Cajal bodies, where SMN complex dissociates from snRNPs, more snRNP-specific proteins are assembled, and box C/D and box H/ACA scaRNPs guide snRNA 2'-O-methylation and pseudouridylation respectively. snRNAs are extensively modified, and some of these modifications are required for splicing activity (for review see Karijolich and Yu (2010)).

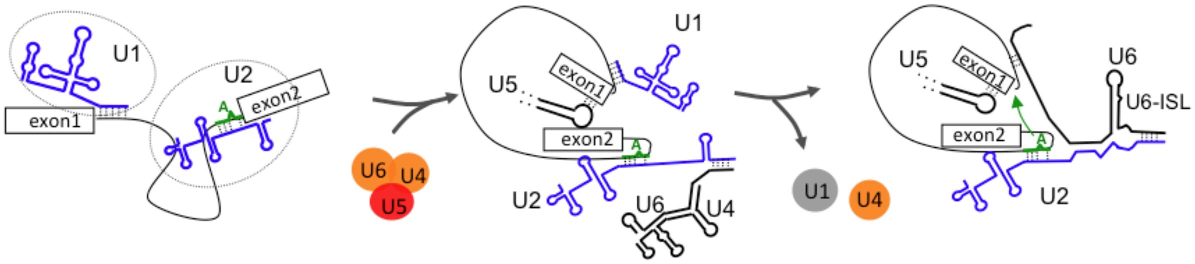


Figure 1.9: **A simple diagram of U2-type splicing.** The splicing process rearranges the RNA-protein interactions in the spliceosome in a highly ordered manner to facilitate the the formation of the catalytic center of the spliceosome. In the first step, U1 and U2 snRNPs form base pairing interactions with the 5' splice site (5'ss) and the intronic branch point sequence (BPS), respectively. U2 binding exposes the the branch point adenosine, indicated by the letter A. The first step forms the pre-spliceosome complex A. Subsequently, U4/U6.U5 tri-snRNP is recruited to the complex A to form the pre-catalytic spliceosome complex B. In complex B, U2 and U6 forms base pairing interactions, releasing U4 from the U4/U6 base paired di-snRNP, while U5 snRNP invariant loop base pairs with sequences in the 5' exon. The 5' end of U6 snRNA then base pairs with the 5'ss to release U1 snRNP from the spliceosome. After U1 and U4 are displaced, complex B* is formed. The complex B* then undergoes extensive rearrangement to bring 5'ss and the branch point adenosine close to each other and facilitate the first trans-esterification reaction. Further rearrangements occur to facilitate the second trans-esterification reaction. The center of U6 forms an intra-molecular stem-loop (U6-ISL), which is necessary for catalyzing the trans-esterification reactions. (Adapted from Matera and Wang 2014.)

1.8 The eukaryotic pre-mRNA splicing process

Since the discovery of Sm containing RNP complexes in eukaryotes, their best-studied function is in pre-mRNA splicing. Most genes in higher eukaryotes are interrupted by intervening sequences, called introns. Although introns are often considered as junk because they are not translated into proteins, studies have shown that they are essential in the process of molecular evolution, and they regulate multiple aspects of gene expression (Roy and Gilbert, 2006).

The presence of introns facilitates exon-duplication and shuffling to create more complicated transcripts and therefore protein structures. Almost all the steps in gene expression are tightly coupled, and therefore it is not surprising that the splicing reaction itself has a role in promoting gene expression, in certain situations (Gruss *et al.*, 1979; Callis *et al.*, 1987; Nott *et al.*, 2004). Another important function of splicing, which is currently under active investigation is that alternative use of exons produce more diverse transcripts than could be encoded by intronless genes. These alternative transcripts can be expressed in a tissue specific manner, at distinct stages of development.

In a classical paper by Joan Steitz and colleagues, they analyzed the sequence features of snRNPs and exon-intron junctions of pre-mRNAs, and found extensive complementarity between snRNAs and pre-mRNAs. This was the first clue that snRNPs are involved in splicing. The mechanism of splicing has since been worked out in great details (fore recent reference see Matera and Wang (2014)). A brief summary of the process is presented here.

For the assembly of the U2 type spliceosome, U1 snRNP first base pairs with the 5' ss (splice site) of the pre-mRNA forming the early complex (complex E, Figure 1.8). Then U2 snRNP binds the branchpoint sequence (BPS) and the interaction between U1 and U2 snRNPs brings together the 5' end and the 3' end of the intron to form complex A. The binding of U1 and U2 snRNPs to pre-mRNA splicing elements are relatively weak, usually involving 6-8 imperfect base pairs. The weak basepairing is thought to avoid having a energetic trap which makes recycling of snRNPs difficult. These weak interactions are further strengthened by accessory proteins, such as SR proteins and U2AF.

U4/U6.U5 tri-snRNP is recruited to complex A with the help of the DExD/H helicase Prp28, forming complex B. The resulting complex B then undergoes a series of rearrangements in composition and conformation to form the active complex B*. The activation of complex B is accompanied by the release of U1 and U4 snRNPs from the spliceosome. At the same time U2 and U6 forms extensive base pairing with each other, and the 5' end of U6 snRNP base pairs with the 5' ss. Complex B* catalyzes the first transesterification reaction, generating complex C, which contains the 5' free exon, the intron-3'-exon lariat with U2, U5 and U6 snRNPs. Complex C is further rearranged to catalyze the second transesterification reaction, releasing the lariat intron and connecting the two exons. U2, U5 and U6 are released from the splicing reaction and recycled. Recent studies showed that U6 snRNA is at the catalytic center of the spliceosome and functions as a ribozyme for both transesterification reactions (Fica *et al.*, 2013).

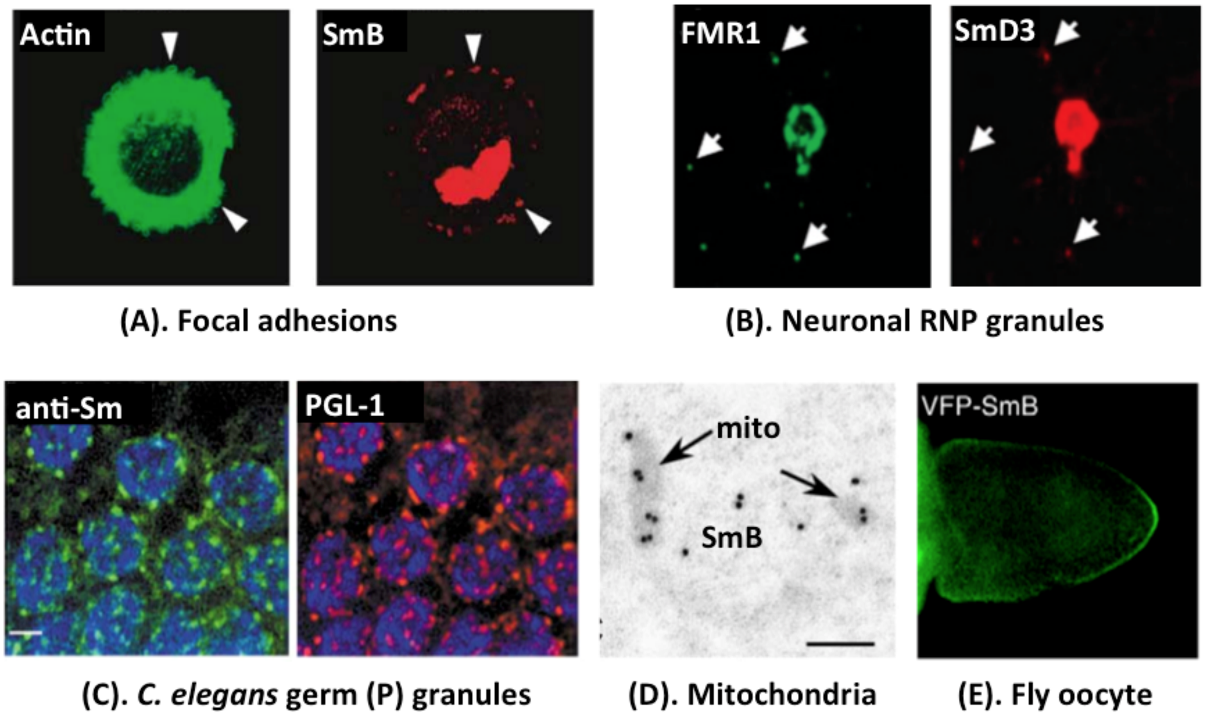
Besides the step-wise assembly pathway, a penta-snRNP has been proposed to form independent of the pre-mRNAs and the penta-snRNP binds the pre-mRNA as one complex. However, this model is not well supported by experimental evidence (Stevens *et al.*, 2002; Malca *et al.*, 2003). The U12 type of splicing is very similar to the U2 type splicing. One of the major difference is that U11/U12 di-snRNP is formed prior to binding the pre-mRNAs.

1.9 Hints at potential new functions of eukaryotic Sm proteins

A number of publications have indicated potentially new functions of the Sm proteins, based on genetic interactions and subcellular localization. Mechanistic insights have been lacking from most of these studies and it is not clear whether these functions involve spliceosomal snRNPs or splicing.

Focal adhesions are subcellular structures that help attach the cells to extracellular matrix and play important roles in cell signaling. An early stage of focal adhesions, spreading initiation centers (SIC) has an unique composition. Quantitative mass spectrometry analysis revealed a set of RBPs concentrating in SIC, including FUS/TLS, hnRNP K and E1, and Sm proteins (de Hoog *et al.*, 2004) (Figure 1.10A). Ribosomal RNAs are also concentrated in the SIC, suggesting that these regions are involved in localized translation and Sm proteins may play a role in this process.

Localized translation is an important mechanism in the polarity of neurons. It has been shown that a specialized RNP granule exists in the neuronal processes in many species, and this granule is implicated in the transport and translational repression of mRNPs (Kiebler and Bassell, 2006; Barbee and Evans, 2006; Cziko *et al.*, 2009) (Figure 1.10B). Neuronal RNP granules contain many RBPs involved in mRNA metabolism, including FMR1 (translational regulation), SMN (snRNP assembly and mRNA transport), Staufen (germline granules), Argonautes (miRNA silencing), Sm proteins etc. A suppressor screen in flies suggest that Sm



(A). Focal adhesions

(B). Neuronal RNP granules

(C). *C. elegans* germ (P) granules

(D). Mitochondria

(E). Fly oocyte

Figure 1.10: **Hints at new functions of the canonical Sm proteins.** (A). Colocalization of Sm proteins and actin in focal adhesions (de Hoog *et al.*, 2004). (B). Colocalization of Sm proteins and FMR1 in neuronal granules (Cziko *et al.*, 2009). (C). Localization of Sm proteins in (C. elegans) germ granules (Barbee *et al.*, 2002; Barbee and Evans, 2006). (D). Localization of Sm proteins at the surface of mitochondria (Bilinski *et al.*, 2004). (E). Localization of Sm proteins along the cortex of fruitfly oocytes (Gonsalvez *et al.*, 2010).

protein genes genetically interact with *Fmr1* (Cziko *et al.*, 2009). The genetic interaction and colocalization of Sm proteins with general mRNA metabolism factors suggest that Sm proteins have additional functions beyond splicing.

Germ granules, also known as P granules, are large RNP aggregates present in the germline cytoplasm of many animal species (Figure 1.10C, D and E) (Barbee *et al.*, 2002; Barbee and Evans, 2006; Updike and Strome, 2010). Alternative forms of P granules exist dynamically at different stages of germline development, as perinuclear nuage, mitochondrial cloud, intermitochondrial cement, and chromatoid bodies etc. These structures often contain large amounts of various RNA and protein components and are involved in regulating the identity and properties of the germline. Surprisingly, several studies have shown that Sm proteins localize to the germline granules in various species including worm, fly and mouse (Figure 1.10C, D and E).

Loss of Sm proteins in *C. elegans* results in defects in the segregation of P granules (Barbee *et al.*, 2002; Barbee and Evans, 2006). Sm proteins localize to the posterior pole of fly oocytes where germline determination factors are concentrated (Gonsalvez *et al.*, 2010). A hypomorphic mutation of SmD (*Smd3pt*) causes mislocalization of oskar mRNA, a factor critical for the determination of germline, and a grandchildless phenotype.

Sm proteins are RNA binding proteins, however, in these examples mentioned above, the RNA targets of Sm proteins have mostly remained elusive (except in the case of fly germline specification). Further studies are required to identify their targets in vivo. Since Sm proteins are essential for cell survival and loss of Sm proteins has pleiotrophic effects, it is hard to attribute any of them to specific interactions outside of the spliceosome. Identifying the RNA partners of Sm proteins would greatly facilitate a more targeted analysis of the interactions.

1.10 Novel functions of spliceosomal snRNPs

A well accepted dogma in prokaryotes is that transcription and translation are coupled. The coupling was best illustrated by the localization of polyribosomes on elongating mRNAs on DNA in a 'Christmas tree' like conformation (Miller *et al.*, 1970). This is in stark contrast to eukaryotes, where the presence of nuclear membranes separates the two processes in time and space. Conceptually, the temporal and spatial separation of the two most basic cellular processes allows for multiple levels of control and regulation, part of the reason for more complex structures of eukaryotic organisms.

Research in the past three decades, however, have gradually revealed that there is another consistent theme to eukaryotic gene expression besides compartmentalization, that is, coupling. The distinct steps of gene expression in an eukaryotic cell have been characterized in detail: transcription, capping, splicing, cleavage and polyadenylation, export, subcellular localization, pioneer round of translation, non-sense mediated decay, stable translation, translational repression, and degradation. All these steps are now known to be coupled in a sequential (sometimes circular) fashion. Coupling of the gene expression events makes the each step more precise and efficient, similar to the idea for transcription/translation coupling in prokaryotes.

A large body of work has been devoted to studying the mechanisms of coupling among the events, and many reviews have been written about these studies (Neugebauer, 2002; Proudfoot *et al.*, 2002; Maniatis and Reed, 2002; Lee and Tarn, 2013; Bentley, 2014). These studies provided a systemic view of how gene expression is coordinated and also provided a framework for understanding new discoveries. It is in this light that I review the novel functions of spliceosomal snRNPs, in addition to the comparison to the versatile functions of prokaryotic Sm containing small RNPs.

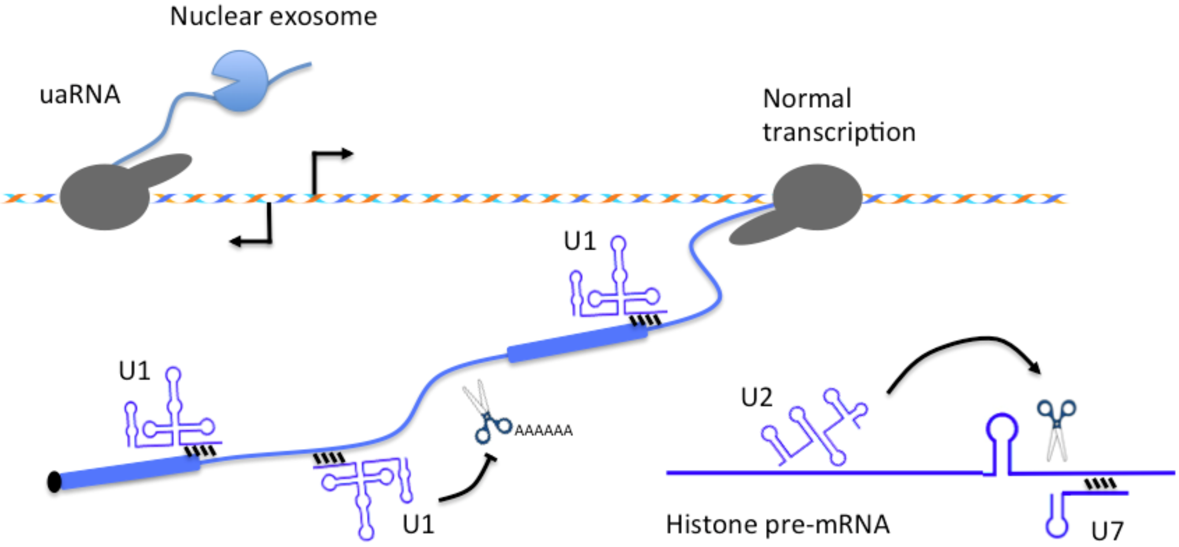


Figure 1.11: **New functions for the spliceosomal snRNPs.** (A) U1 snRNP binding to pre-mRNAs regulates cleavage and polyadenylation. Consequences of this regulation include promoting stabilization of normal transcription relative to upstream transcription, inhibition of premature (or even normal) cleavage and polyadenylation. (B) U2 snRNP promotes histone mRNA processing.

U1 snRNP and polyadenylation. The first indication that U1 snRNP is involved in non-splicing activities came from studies on the U1-A protein (see Figure 1.7 for snRNP compositions) (Gunderson *et al.*, 1994). The 3' end of U1A mRNA forms a stem loop secondary structure, resembling part of the first stem loop in U1 snRNA, that associates with U1A protein directly. U1A protein interact with poly(A) polymerase (PAP) and inhibits its activity. Inhibition of polyadenylation makes U1A mRNA less active in translation. Therefore, U1-A protein negatively regulate the production of U1-A, forming a classical negative feedback loop.

More interestingly, it is shown by several groups that U1 snRNP binding to the 5' splice site of certain mRNAs could also inhibit polyadenylation (Gunderson *et al.*, 1998; Ashe *et al.*, 1997, 2000; Liu *et al.*, 2002) (Figure 1.11. These effects were mediated instead by U1-70K protein of the U1 snRNP complex in a mechanistically similar manner to U1-A negative feedback loop. A structure motif in U1-70K similar to U1-A binds PAP to inhibit its activity. These examples illustrated the coupling of splicing and polyadenylation, which are mediated by the U1 snRNP.

These findings have recently been confirmed and extended by the Dreyfuss lab to the transcriptome scale (Kaida *et al.*, 2010; Berg *et al.*, 2012). Functional knockdown of U1 snRNAs using antisense morpholinos induced pre-mature cleavage and polyadenylation of many transcripts in fly, mouse and human cells. As a result, the average length of the transcripts is reduced. The degenerate recognition of 5' ss in pre-mRNAs by U1 snRNAs predicts large numbers of U1 binding sites on pre-mRNAs, especially in higher eukaryotes, where intronic sequences are usually much longer than exonic sequences. Even though it is unclear whether this mechanism of gene regulation affects the steady state level and length of the mRNAs, it nevertheless suggests a new function for snRNPs beyond splicing, but to some extent coupled to splicing. (Recent studies in our own lab suggest the effect of U1 downregulation on transcriptome wide mRNA length is hardly visible. Garcia et al. unpublished results.)

U1 snRNP and promoter upstream antisense transcripts. Recent transcriptome analysis showed that promoters of many genes are bidirectional (Preker *et al.*, 2008; Seila *et al.*, 2008) (Figure 1.11). Promoter upstream transcription generates unstable RNAs that are targeted by the nuclear exosome complex for degradation. Functions of the upstream antisense RNAs (uaRNAs, also known as promoter upstream transcript, PROMPTs) are still unknown. One theory suggests that these transcripts may lead to new gene origination during evolution (Wu and Sharp, 2013). Recently, a unique mechanism has been shown to contribute to the instability of uaRNAs (Almada *et al.*, 2013; Ntini *et al.*, 2013). Compared to the normal transcripts produced downstream of the promoters, the uaRNAs are relatively depleted of U1 snRNA binding sites. Whereas U1 snRNP binding on pre-mRNAs protects the mRNAs from pre-mature cleavage and polyadenylation, the lack of U1 snRNP binding sites on uaRNAs predisposes them to polyadenylation-dependent decay.

U2 snRNP and mRNA cleavage and polyadenylation. Together with U1 snRNP, U2 snRNP also participates in the intron/exon definition early in splicing of pre-mRNAs. Not surprisingly, U2 snRNP has also been shown to be coupled to mRNA 3' end processing. Kyburz et al. recently showed that CPSF (cleavage and polyadenylation specificity factor) binds U2 snRNP directly, and the physical interaction is functional in both directions (Kyburz *et al.*, 2006). Presence of U2 snRNP in cleavage and polyadenylation assays significantly increased the efficiency; at the same time, CPSF can also promote the efficiency of splicing.

U2 and U12 snRNP and histone mRNA processing. The replication-dependent histone mRNAs are probably the only class of non-polyadenylated mRNAs in eukaryotes. This unique feature entails a unique 3' end processing reaction that is different from normal polyadenylated mRNAs. As mentioned above, processing of histone mRNA 3' end requires U7 snRNP to base pair with the histone down stream element (HDE), and further recruit cleavage factors to cut between the histone mRNA stemloop and HDE (Figure 1.11). Many of the factors involved in 3' end processing also regulate mRNAs and replication-dependent histone mRNAs. Steitz and colleagues showed that, despite the fact that replication-dependent histone mRNAs do not have introns, U2 and U12 snRNPs can still bind them (Friend *et al.*, 2007). The SF3b complex in U2 snRNP, not basepairing, mediates U2 and U12 snRNPs binding to histone pre-mRNAs. The interaction stimulates efficient 3' end processing. Interestingly, the SF3b complex-mediated interaction can be replaced by basepairing-mediated interaction, without affecting the processing stimulation effect.

Compared to U1 and U2 snRNPs, the functions of the other snRNPs outside of splicing has been less well studied. The major difficulty in studying the non-canonical functions of snRNPs lies in the essentialness of splicing. New methods are needed to identify the binding sites of all the snRNPs on pre- and mature mRNAs. Once the binding sites are identified, it would be possible to study these functions by disrupting the mRNA-snRNA interaction by mutating the mRNAs, instead of the snRNAs.

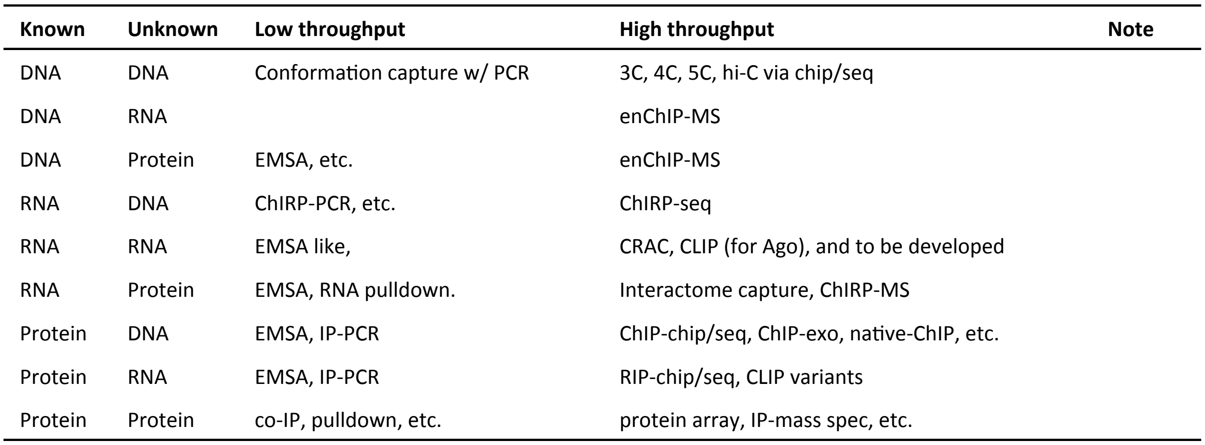
1.11 Methods for studying DNA-RNA-protein interactions

The three information-carrying biomolecules, DNA, RNA and protein, interact extensively to coordinate their behaviors in the cells. The dynamic assembly of DNA-RNA-protein macromolecular complexes is critical for executing hardwired genetic programs and responding to extracellular stimuli. Therefore, comprehensive knowledge of the diversity of molecules and their intricate interactions is the foundation for modern molecular biology.

Technological innovation is a driving force for scientific discoveries. Numerous biochemical techniques have been developed in the past decades to study the interactions among DNA, RNA and proteins. With these methods, we are beginning to have a rough overview of how macromolecular machineries are organized in space and time. The methods generally have the same principles and have gradually evolved from low throughput to high throughput. Here I briefly review the biochemical methods developed so far for studying interactions, with a focus on the high throughput varieties. The commonly used methods are summarized in Table 1.1. Genetic methods for high throughput detection of interactions, like yeast two-hybrid screens, are not discussed here.

Traditional ways to analyze DNA-RNA-protein interactions usually involve the purification of individual components using a variety of chromatography or immunological methods. These methods take advantage of the properties of biomolecules, for example, sequence complementarity in DNA and RNA sequences, antibody-antigen affinities etc. Once one or both molecules we try to assay are purified (not necessarily to homogeneity), several types of methods are used to detect their interactions, for example, electrophoretic mobility shift assay (EMSA), co-immunoprecipitation (co-IP), pull-downs, etc.

EMSA and related methods can be used to detect the interactions between protein and protein, DNA and protein, RNA and protein. Co-IP and related methods can also be used for these purposes. Subsequent detection of molecules could involve a variety of methods, like non-specific dye staining (using Coomasie Blue etc.), western blotting, northern blotting, southern blotting, PCR, etc.



Known	Unknown	Low throughput	High throughput	Note
DNA	DNA	Conformation capture w/ PCR	3C, 4C, 5C, hi-C via chip/seq	
DNA	RNA		enChIP-MS	
DNA	Protein	EMSA, etc.	enChIP-MS	
RNA	DNA	ChIRP-PCR, etc.	ChIRP-seq	
RNA	RNA	EMSA like,	CRAC, CLIP (for Ago), and to be developed	
RNA	Protein	EMSA, RNA pulldown.	Interactome capture, ChIRP-MS	
Protein	DNA	EMSA, IP-PCR	ChIP-chip/seq, ChIP-exo, native-ChIP, etc.	
Protein	RNA	EMSA, IP-PCR	RIP-chip/seq, CLIP variants	
Protein	Protein	co-IP, pulldown, etc.	protein array, IP-mass spec, etc.	

Table 1.1: **Biochemical methodologies for studying DNA-RNA-Protein interactions.** This is not a comprehensive list, as less commonly used methods are not listed here. 3C: chromosome conformation capture (Dekker *et al.*, 2002). 4C: 3C on a chip, or circular 3C (Zhao *et al.*, 2006). 5C: carbon-copy 3C (Dostie and Dekker, 2007). Hi-C: (Belton *et al.*, 2012).EMSA: electrophoretic mobility shift assay. IP: immunoprecipitation. ChIP: chromatin immunoprecipitation (Gilmour and Lis, 1984). enChIP: engineered ChIP (Fujita *et al.*, 2013). ChIP-exo: ChIP with exonuclease treatment (Rhee and Pugh, 2011). RIP: RNA immunoprecipitation (Keene *et al.*, 2006; Zhao *et al.*, 2010; Lu *et al.*, 2014). CLIP: cross-linking immunoprecipitation (Ule *et al.*, 2003). PAR-CLIP: Photoactivatable-Ribonucleoside-Enhanced CLIP (Hafner *et al.*, 2010). MS and Mass spec: mass spectrometry. ChIRP: chromatin isolation by RNA purification (Chu *et al.*, 2012). CHART-seq capture hybridization analysis of RNA targets by sequencing (Simon *et al.*, 2011). CRAC: cross-linking and analysis of cDNAs (Granneman *et al.*, 2009). ATAC-seq: assay for transposase-accessible chromatin using sequencing (Buenrostro *et al.*, 2013).

The major disadvantage of these low throughput methods is that, besides the often time-consuming and labor-intensive procedure, they cannot be used to easily discover new types of interactions. The advent of high throughput methods solved this problem. While the molecular complex purification step remains similar, the methods for detection of potential binding partners have changed. Mass spectrometry has been developed to identify the proteins that interact with specific protein, RNA or DNA in a unbiased way. Microarrays and high throughput sequencing methods have been developed to identify DNA and RNA sequences that interact with other DNA, RNA and proteins. While many possible methods have been developed, however, other approaches remain to be developed, or further optimized; for example, a method for identifying all intermolecular RNA interactions, a method for identifying all protein and RNA components associated with a specific RNA. The development of these methods, while challenging, would greatly advance studies on RNA-RNA and RNA-protein interactions. Further ideas on these directions and potential applications will be discussed in the Conclusions chapter.

Critical to the interpretation of these analyses are the inclusion of proper controls and independent lines of evidence. However, these principles, which are widely applied to research, have been largely neglected in the execution and analysis of many high throughput experiments. Many of the genome-wide studies tend to report a large number of identified molecules, with little confidence. In other words, people have inappropriately placed more emphasis on sensitivity than on specificity (see a nice layman's summary of things in genomics approaches that need special attention here, http://genomeinformatician.blogspot.co.uk/2011/07/10-rules-of-thumb-in-genomics.html). A recent paper actually showed that almost none of the PAR-CLIP datasets have proper controls (Friedersdorf and Keene, 2014). With this in mind, we included proper controls and multiple lines of experiments in our RIP-seq analysis of Sm proteins (in Chapter 2) (Lu *et al.*, 2014).

1.12 Abnormal RNA-seq reads, valuable information

Recent development of massively parallel sequencing methods enabled deep analysis the transcriptome The great power of RNA-seq lies not only in the accurate determination of expression levels, like 'counting beads', but also in the rich information contained in the sequence reads, which can be hard or impossible to obtain from traditional low throughput methods or even well designed microarrays. Since the introduction of high throughput sequencing, huge amounts of efforts have been devoted to developing new tools for data analysis. Most of these analysis tools deal with two questions: 1) Is there difference in RNA expression among the samples? 2) Are there alternative splicing events among the samples? Other types of analysis methods have also been developed and an increasing number of tools are now available to get more knowledge out of the large amounts of RNA-seq data available from public databases. These new methods arise due to the realization that properly performed RNA-seq experiments can be reused for different purposes and RNA-seq reads can store interesting information about RNA structure and sequence in unexpected ways (Figure 1.12).

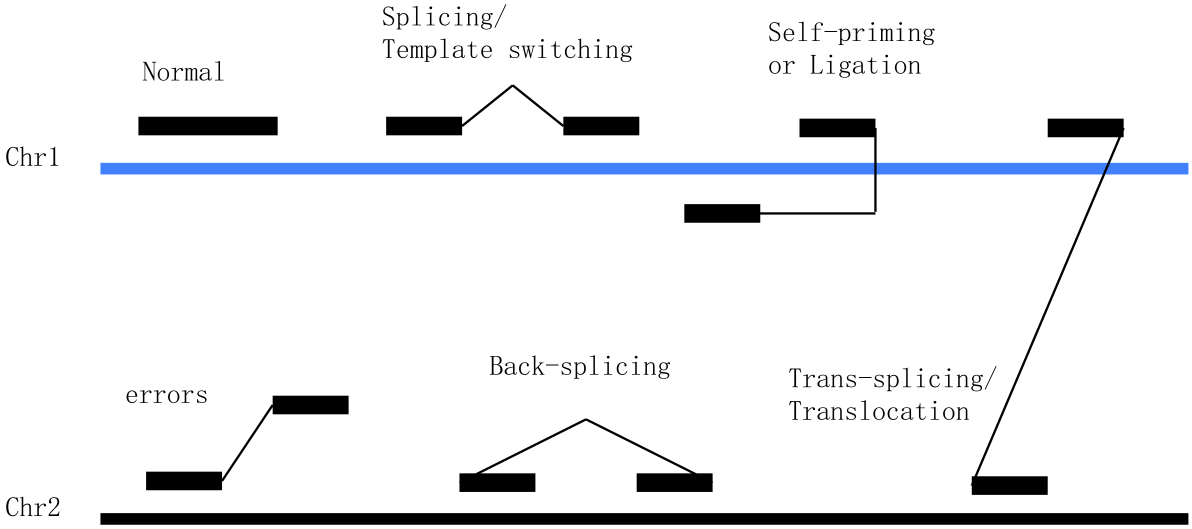


Figure 1.12: **All kinds of RNA-seq Reads and their mapping patterns.** The most commonly seen types of RNA-seq reads are displayed on two chromosomes. Please see text for detailed descriptions.

In order to help understand the methods developed for RNA-seq data, and also provide a broader background for the Vicinal mapping method I developed in Chapter 3, here I summarize the types of RNA-seq reads (not the types of experiments, which were discussed in part in the previous section) we usually see and the biological meaning of each. Normal RNA-seq reads map to the genome nicely, without mismatches, insertions or deletions (most of them to single locations) (Figure 1.12). Some of the reads that are mapped nearly perfectly, with single mismatches, could come from RNA editing (Ramaswami *et al.*, 2013; Park *et al.*, 2012; Bahn *et al.*, 2012). Errors or polymorphisms in genome sequences could also contribute to consistent single mismatches, and care should be taken in interpreting them. An appropriate percentage of the mismatches at the same position, the specific mismatch (for example A to I editing) and sequence context, followed by experimental validation can be used to help distinguish true RNA editing from artifacts.

A significant portion of the reads map to exon-exon junctions, in a discontinuous fashion. These reads are useful, not only for normal quantification of RNA-seq data, but also for the *de novo* identification of splice sites. All the varieties of alternative splicing events can be easily captured in one experiment, e.g. alternative exon use, skipping of exons and inclusion of introns, etc. This is very valuable since prediction of splice sites can be hard due to the degeneracy of the sequence elements that define splice sites and the fact that real splice sites may not always be used in a specific cell type. The availability of the huge amounts of RNA-seq data also leads to the discovery of recursive splicing, which is one of the mechanisms used to remove long introns (Sarah Olson and Brenton Graveley, presentation in the 2013 RNA Society Meeting). More recently, the use of long RNA-seq reads (e.g. from PacBio sequencers) has enabled the direct quantification of full length isoforms produced through alternative use of transcription start sites, alternative splicing, and alternative cleavage and polyadenylation (Tilgner *et al.*, 2013).

Split reads do not always indicate splicing events. Sometimes split reads could be caused by slippery reverse transcription in the presence of repetitive sequences, in other words, template switching. Template switching is physiologically relevant as an important step in RNA virus replication. During the normal *in vitro* reverse transcription, template switching can go forward or backward, creating reads that are similar the ones in splicing and back-splicing (back-splicing will be discussed next), therefore, care should be taken in analyzing splicing and back-splicing events.

In addition to continuous and discontinuous reads that come from normal intron containing genes, other types of reads also exist, but in very low levels (Figure 1.12). One specific situation that is analyzed in Chapter 3 is the mapping of two parts of the same read to opposite strands in close distance. The presence of these reads indicates the occurrence of self-priming or ligation events in the context of terminal stem loops of ncRNAs (See Chapter 3 for details) (Lu and Matera, 2014).

In some instances, two parts of the same reads could be mapped to different genes, distant chromosomal locations or different chromosomes (Figure 1.12A). These reads could come from trans-splicing, gene fusions or chromosome translocations (McManus *et al.*, 2010; Edgren *et al.*, 2011; Sakarya *et al.*, 2012).

Yet another class of chimeric reads is the reordering of the two parts in a chiastic manner (Figure 1.12). These chimeric reads could come from template switching as discussed above, or from back-splicing and other types of RNA processing reactions. Back-splicing is a proposed model for explaining the large number of exonic circular RNAs recently discovered (Jeck *et al.*, 2013; Salzman *et al.*, 2012, 2013; Hansen *et al.*, 2013; Memczak *et al.*, 2013; Hentze and Preiss, 2013). It has also been shown that archaeal tRNA splicing, rRNA, snoRNA and RNaseP processing that generate circular RNAs could also produce chimeric reads that map in a chiastic manner (Danan *et al.*, 2012).

Of course there are also chimeric reads that could simply come from sequencing errors. Even if they are not sequencing errors, low numbers of chimeric reads could arise by chance. There are several *in silico* ways of distinguishing meaningful chimeric reads from artifacts, including using deeper read coverage, in addition, piles of meaningful chimeric reads should not be identical in sequence or shifted by one to two bases, instead, they should be characterized by a stacked/ladder-like pattern (Edgren *et al.*, 2011) (Figure 1.12B).

With the increasing amount of RNA-seq experiments and increasing variety of methods for performing the experiments and generating sequencing libraries, more and more data will be available for the analysis of each of the different types of 'abnormal reads' discussed above. This is both a challenge and an opportunity for biologists and bioinformaticians in both hypothesis-driven and data-driven research.

CHAPTER 2: RIP-seq analysis of eukaryotic Sm containing ribonucleoproteins

Authors: Zhipeng Lu[1], Xiaojun Guan[2], Casey A Schmidt[3] and A Gregory Matera[4]

2.1 Abstract

Background: Sm proteins are multimeric RNA-binding factors, found in all three domains of life. Eukaryotic Sm proteins, together with their associated RNAs, form small ribonucleoprotein (RNP) complexes important in multiple aspects of gene regulation. Comprehensive knowledge of the RNA components of Sm RNPs is critical for understanding their functions.

Results: We developed a multi-targeting RNA-immunoprecipitation sequencing (RIP-seq) strategy to reliably identify Sm-associated RNAs from *Drosophila* ovaries and cultured human cells. Using this method, we discovered three major categories of Sm-associated transcripts: small nuclear (sn)RNAs, small Cajal body (sca)RNAs and mRNAs. Additional RIP-PCR analysis showed both ubiquitous and tissue-specific interactions. We provide evidence that the mRNA-Sm interactions are mediated by snRNPs, and that one of the mechanisms of interaction is via base pairing. Moreover, the Sm-associated mRNAs are mature, indicating a splicing-independent function for Sm RNPs.

Conclusion: This study represents the first comprehensive analysis of eukaryotic Sm-containing RNPs, and provides a basis for additional functional analyses of Sm proteins and their associated snRNPs outside of the context of pre-mRNA splicing. Our findings expand the repertoire of eukaryotic Sm-containing RNPs and suggest new functions for snRNPs in mRNA metabolism.

[1] This chapter was previously published in the journal *Genome Biology* 2014, 15:R7.
Department of Biology, Integrative Program for Biological and Genome Sciences (IBGS), UNC Chapel Hill

[2] Center for Bioinformatics, UNC Chapel Hill. Current address: Sequenom, San Diego, CA

[3] Curriculum in Genetics and Molecular Biology, IBGS, UNC Chapel Hill

[4] Departments of Biology and Genetics, IBGS, UNC Chapel Hill, Correspondence: matera@unc.edu

2.2 Background

Sm proteins are a family of highly conserved RNA-binding proteins present in all three domains of life (Valentin-Hansen *et al.*, 2004; Salgado-Garrido *et al.*, 1999). In bacteria and archea, Sm homologs form either homohexameric (for example, Sm2 and Hfq) or homoheptameric (Sm1) ring-shaped complexes (Toro *et al.*, 2002; Sauter *et al.*, 2003). These complexes regulate the stability and translation of mRNAs by facilitating base pairing interactions between small RNAs (sRNAs) and mRNAs (Sledjeski *et al.*, 2001; Zhang *et al.*, 1998, 2002). In eukaryotes, more than 20 Sm protein homologs assemble into several distinct heteroheptameric rings (Wilusz and Wilusz, 2005). There are two major eukaryotic Sm classes: the canonical Sm proteins and the Sm-like (Lsm) proteins (Matera *et al.*, 2007). Canonical Sm proteins also form heptamers that bind the major and minor uridine-rich small nuclear ribonucleoprotein (snRNP) particles (U1, U2, U4, U4atac, U5, U7, U11 and U12). These small RNPs carry out important metabolic reactions such as pre-mRNA splicing and 3' end processing (Matera *et al.*, 2007; Gunderson *et al.*, 1998; Kaida *et al.*, 2010; Berg *et al.*, 2012; Pillai *et al.*, 2003). Lsm proteins form two distinct heteroheptameric complexes. The Lsm1-7 ring directly binds the 3' end of oligoadenylated mRNAs and is involved in regulating mRNA decay (Tharun and Parker, 2001), while the Lsm2-8 ring binds to the 3' oligouridine tail of U6 and U6atac small nuclear (sn)RNAs to form RNP particles that participate in pre-mRNA splicing (Tharun, 2009; Achsel *et al.*, 1999; Mayes *et al.*, 1999; Vidal *et al.*, 1999). Thus, the Lsm proteins, which regulate mRNA stability, are thought to be more akin to their archaeal and bacterial brethren.

A growing body of evidence points to potential new roles for canonical Sm proteins and Sm class snRNPs outside of the spliceosome in the processing, localization and translational control of messenger RNPs (mRNPs). In *Caenorhabditis elegans*, Sm proteins, but not other splicing factors, localize to germline P granules and are required for their integrity (Barbee *et al.*, 2002; Barbee and Evans, 2006). In *Drosophila melanogaster*, SmB and SmD3 are enriched at the posterior pole of developing oocytes (Gonsalvez *et al.*, 2010), and a hypomorphic mutation in SmD3 causes mislocalization of oskar mRNPs and pronounced defects in germ cell specification that are independent from splicing (Gonsalvez *et al.*, 2010). Moreover, loss of the Sm protein methyltransferase PRMT5 results in failure to specify the germline (Gonsalvez *et al.*, 2010, 2007; Anne *et al.*, 2007). Furthermore, a genetic screen for modifiers of *FMR1* (Fragile X mental retardation 1) in *Drosophila* identified SmD3 as a suppressor of dFMR1's translational repression function, and SmD3 and dFMR1 were found to colocalize within neuronal mRNP granules (Cziko *et al.*, 2009). In vertebrates, Sm proteins are enriched in the nuage and mitochondrial cement (Bilinski *et al.*, 2004), structures that share many components with the invertebrate germ plasm. The U1 snRNP, in addition to its splicing role, protects pre-mRNA from premature polyadenylation at cryptic poly(A) signals in introns (Gunderson *et al.*, 1994; Kaida *et al.*, 2010; Berg *et al.*, 2012), and inhibits HIV RNA polyadenylation (Ashe *et al.*, 1997, 2000). In addition, RNA sequence elements complementary to the U1 5' end play important roles in the stabilization of promoterdownstream transcripts and thus contribute to promoter directionality (Almada *et al.*, 2013; Ntini *et al.*, 2013). The U1 snRNP not only regulates gene expression via RNA processing; a modified form of U1 can also target HIV RNA to reduce viral protein expression (Sajic *et al.*, 2007). Moreover, the U2 and U12 snRNPs play an unexpected role in promoting U7-snRNP-dependent processing of intronless histone mRNAs in human cells, and both protein-RNA interaction and RNA-RNA base-pairing suffice for the activity (Friend *et al.*, 2007). Collectively, these studies suggest additional functions for Sm proteins and snRNPs in RNA metabolism; however, little is known about the in vivo RNA targets that might be regulated by Sm proteins/snRNPs, in these processes.

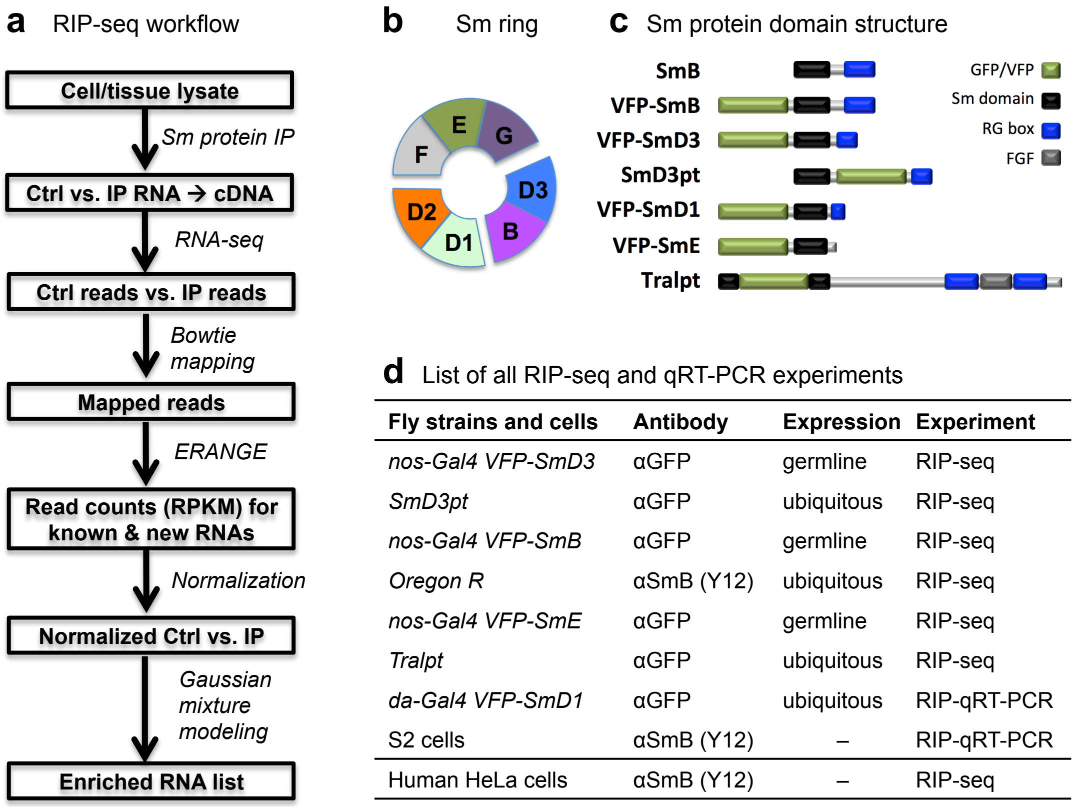
To systematically identify Sm protein-containing RNPs, we carried out RNA-immunoprecipitation (RIP) against multiple Sm proteins from *Drosophila* ovaries and HeLa cells, followed by high-throughput sequencing (RIP-seq) of the immunopurified RNAs. Using this robust and reproducible multi-targeting RIP-seq approach, we recovered most of the spliceosomal snRNAs. In addition, we discovered a new *Drosophila*-specific snRNA, many Smassociated small Cajal body-specific RNAs (scaRNAs), and numerous Sm-associated mRNAs from both Drosophila and human cells. The new snRNA is highly conserved in the melanogaster group of Drosophilids, although it is not essential for organismal viability. Two major categories of the Sm-associated mRNAs encode mitochondrial and translation-related proteins. Using quantitative reverse transcriptase PCR (qRT-PCR), we found that some of the RNA-Sm interactions are tissue-specific, whereas others are more widespread. The Sm-associated mRNAs are properly spliced and polyadenylated, indicating that the mRNA-Sm interactions reported here are distinct from those involved in pre-mRNA splicing and Lsm1-7 dependent degradation. We also provide evidence that the mRNA-Sm association is mediated by snRNPs, and we show that a predicted U1 snRNP base pairing region on an mRNA is required for interaction with this snRNP. These mature mRNA-snRNP interactions are very stable and distinct from other previously studied interactions (pre-mRNA splicing, 'telescripting' and regulation of promoter directionality). Taken together, the data identify additional direct targets of canonical Sm proteins, and suggest that Sm class snRNPs may have novel, evolutionarily conserved functions in mRNA localization, stability and translation.

2.3 Results

2.3.1 Identification of RNAs that co-purify with eukaryotic Sm proteins

As mentioned above, the Sm and Sm-like proteins comprise a family of ancient evolutionary origin that functions to modulate the stability and translation of several classes of RNA, including mRNAs (Valentin-Hansen *et al.*, 2004; Sobrero and Valverde, 2012). Based on these ancestral roles, the involvement of eukaryotic Sm proteins in splicing is generally thought to be a derived function, and additional RNA targets of Sm proteins remain to be discovered.

To characterize the repertoire of RNA targets that are associated with Sm proteins in *Drosophila* ovarian lysates, we performed RIP-seq analysis of individual subunits of the canonical Sm ring. We also performed RIP-seq on Trailer Hitch (Tral), a protein that contains an Sm domain (Figure 2.1c). Tral is not incorporated into the canonical Sm ring; therefore, we expected it to associate with a distinct subset of transcripts (Wilhelm *et al.*, 2005). An outline of the experimental strategy and data analysis pipeline is shown in Figure 2.1. Immunoprecipitations (IPs) were carried out using either anti-SmB (monoclonal antibody Y12) or anti-green fluorescent protein (anti-GFP) antibodies (for the GFP- and Venus fluorescent protein (VFP)-tagged proteins). Normal goat serum was used as control for the IP. Immunoprecipitated RNA



Fly strains and cells	Antibody	Expression	Experiment
nos-Gal4 VFP-SmD3	αGFP	germline	RIP-seq
SmD3pt	αGFP	ubiquitous	RIP-seq
nos-Gal4 VFP-SmB	αGFP	germline	RIP-seq
Oregon R	αSmB (Y12)	ubiquitous	RIP-seq
nos-Gal4 VFP-SmE	αGFP	germline	RIP-seq
Tralpt	αGFP	ubiquitous	RIP-seq
da-Gal4 VFP-SmD1	αGFP	ubiquitous	RIP-qRT-PCR
S2 cells	αSmB (Y12)	–	RIP-qRT-PCR
Human HeLa cells	αSmB (Y12)	–	RIP-seq

Figure 2.1: **RIP-seq experimental analysis strategies.** (a) Outline of RIP-seq analysis pipeline. See Materials and methods for details. (b) Schematic diagram of the canonical Sm ring. The three sub-complexes are shown separately. (c) Schematic diagram of the Sm-domain-containing proteins used in this study. (d) Summary of the RIP-seq and RIP-qRT-PCR experiments performed, targeting all three sub-complexes of the canonical Sm ring and Tral. See Table 2.1 for details. Ctrl, control; GFP, green fluorescent protein; IP, immunoprecipitation; RPKM (reads per kilobase per million reads); VFP, Venus fluorescent protein.

was reverse transcribed to cDNA, fragmented, ligated with adapters, PCR-amplified and sequenced on an Illumina Genome Analyzer II.

To reduce potential non-specific interactions and artifacts, we carried out RIP-seq on several Sm proteins expressed from three different genomic contexts: (i) native endogenous genes, (ii) VFP-tagged transgenes, or (iii) a gene-trapped (GFP-tagged) endogenous gene (Figure 2.1c). Comparisons among this wide variety of experimental conditions helps to minimize problems associated with genetic background, transgene overexpression, and antibody specificity. Four different transgenic lines were employed, including VFP-tagged SmD3, SmB, SmD1 and SmE (Gonsalvez *et al.*, 2010). Transgenes were expressed using the UAS/Gal4 system, crossed to a

Fly strain	Antibody	Expression	Priming		Sample No.
Nos-Gal4 VFP-SmD3	αGFP	germline	Exp1	Oligo dT	Lu001(Ctrl), Lu002(IP)
				6mer	Lu003(Ctrl), Lu004(IP)
			Exp2	Oligo dT	Lu005(Ctrl), Lu006(IP)
				6mer	Lu007(Ctrl), Lu008(IP)
SmD3pt	αGFP	ubiquitous	6mer		Lu025(Ctrl), Lu026(IP)
Nos-Gal4 VFP-SmB	αGFP	germline	Oligo dT		Lu013(Ctrl), Lu014(IP)
			6mer		Lu015(Ctrl), Lu016(IP)
Oregon R	Y12	ubiquitous	6mer		Lu023(Ctrl), Lu024(IP)
Nos-Gal4 VFP-SmE	αGFP	germline	Oligo dT		Lu009(Ctrl), Lu010(IP)
			6mer		Lu011(Ctrl), Lu012(IP)
Da-Gal4 VFP-SmD1	αGFP	ubiquitous	6mer		qRT-PCR
S2 cells	Y12	–	6mer		qRT-PCR
HeLa cells	Y12	–	6mer		Lu045 and Lu046 (Ctrl) Lu047 and Lu048 (Ctrl)

Table 2.1: **Details about the RIP-seq and RIP-qRT-PCR experiments.** 6mer: random hexamer. The RIP-seq experiments on nos-Gal4 VFP-SmD3 fly ovaries were performed as biological replicates. All of these RIP-seq experiments were performed on different days.

nanos-Gal4 driver for germline-specific expression or, in the case of VFP-SmD1, to a *daughterless*-Gal4 driver for ubiquitous expression (Brand and Perrimon, 1993). SmB and SmD3 form an obligate dimer (Figure 2.1b), whereas SmD1 and SmE are present in distinct subcomplexes within the heteroheptameric ring structure (Matera *et al.*, 2007). Thus, IPs targeting different components of the Sm ring further reduced potential artifacts resulting from epitope tagging, as these proteins form a complex that is expected to bind a similar set of RNAs. RIP-seq experiments were performed on SmB, SmD3 and SmE, whereas RIP-qRT-PCR was performed on VFP-SmD1 for identified targets. To broaden the scope of our study, we also performed RIP-seq analysis in cultured human HeLa cells, using the Y12 antibody mentioned above (Figure 2.1d; see details in Table 2.1).

2.3.2 Enrichment analysis of Sm RIP-seq experiments

We obtained between 8 and 28 million 35-nucleotide single-end reads per *Drosophila* ovary RIP-seq library, and roughly 20 million 48-nucleotide paired-end reads per human HeLa cell RIP-seq library. All of the fly and human sequencing data are of high quality (Figure 2.2).

samples	sample No.	unique	spliced	RNAFAR	multi	mapped	total reads	mappability	adjust
nos Gal4 VFP-SmD3 single-end	Lu001	39148	1118	144	107024	147434	10426289	0.014140602	
	Lu002	106073	2199	869	5070549	5179690	8063387	0.6423715	1.57
	Lu003	316701	14664	1434	100558	433357	11666984	0.037143875	
	Lu004	1314704	61607	12153	11152871	12541335	18950426	0.661796996	3.46
nos Gal4 VFP-SmD3 single-end	Lu005	12986	384	131	202801	216302	19219893	0.011254069	
	Lu006	59348	1479	531	3145331	3206689	20415816	0.157068863	3.32
	Lu007	137627	6658	603	171802	316690	19641080	0.016123859	
	Lu008	469701	20478	3860	3755700	4249739	23267823	0.182644461	2.28
nos Gal4 VFP-SmE single-end	Lu009	339674	16882	1420	33122	391098	13775633	0.028390565	
	Lu010	1675331	84927	11474	111670	1883402	21298219	0.088430023	5.62
	Lu011	555161	28406	1296	39507	624370	20733108	0.030114636	
	Lu012	1343009	70417	8133	76377	1497936	22561102	0.066394629	2.46
nos Gal4 VFP-SmB single-end	Lu013	195195	5139	1620	55055	257009	23034642	0.011157499	
	Lu014	2415750	87425	33291	130114	2666580	23043259	0.115720611	10.84
	Lu015	418808	18375	903	40119	478205	19020871	0.025141067	
	Lu016	3866621	186827	29650	155750	4238848	21121289	0.200690782	8.63
Tralpt single-end	Lu019	4519439	241175	26155	355607	5142376	29032697	0.177123607	
	Lu020	8798978	444498	89240	392601	9725317	28358846	0.342937685	
Oregon R single-end	Lu023	799995	39024	6860	203022	1048901	23128732	0.045350562	
	Lu024	3905235	219000	37621	1131728	5293584	27762576	0.190673373	3.95
SmD3pt single-end	Lu025	3273832	167489	89240	205543	3736104	28031693	0.133281425	
	Lu026	8265251	430878	105031	1226538	10027698	24649805	0.406806382	2.13
HeLa cells paired-end	Lu045	–	–	–	28212679	30598572	32324208	0.946614748	
	Lu046	–	–	–	25909287	27795460	28647941	0.970242853	0.8402
	Lu047	–	–	–	37043623	38550693	39243007	0.982358284	0.5779
	Lu048	–	–	–	38302969	39946114	41314245	0.966884763	0.6545

Table 2.2: **RIP-seq library statistics** RNAFAR: Reads that cluster to putative new genes or new exons of known genes. Adjust: the adjustment (normalization) factors between each pair of Ctrl and IP (i.e. raw read numbers from the IP in each pair should be divided by the adjustment factor to give normalized read numbers).

Despite differences in total read numbers, the IPs consistently yielded many more mappable reads than did the controls (Table 2.2, 'mapped' and '%mappable' columns). This was to be expected; due to the low amount of input cDNA, most of the reads in the control IPs are not mappable (for example, rRNAs, primer/adapter dimers or even random sequences; Table 2.3) and those that do map to the genome typically correspond to abundant RNAs that stick to the beads non-specifically. Library statistics show that random hexamer priming yielded more mappable reads than did oligo(dT)20 priming (Table 2.4). Thus, we used the random hexamer-primed libraries for the subsequent enrichment analyses. We built a data analysis pipeline (Figure 2.1a) by integrating previously published programs (see Materials and methods for details). Sequence reads for the *Drosophila* RIP-seq experiments were mapped to the *Drosophila* expanded genome and quantified using ERANGE (Mortazavi *et al.*, 2008). Then, for each experiment, we filtered out transcripts with read coverage less than 10. Assuming that

library	Note	Mappable (%)	Primers (%)	rRNA (%)	Other (%)	Total (%)
VFP-SmD3_Lu003	Non-size-selected	3.7	78.1	6.4	11.8	100
VFP-SmD3_Lu004		66.2	11.2	11.5	11.1	100
VFP-SmD3_Lu007		1.6	81.9	1.5	15.0	100
VFP-SmD3_Lu008		18.3	56.2	1.1	24.4	100
VFP-SmE_Lu011	Size-selected	3.0	0.0	48.3	48.7	100
VFP-SmE_Lu012		6.6	0.0	53.1	40.3	100
VFP-SmB_Lu015		2.5	0.0	37.1	60.4	100
VFP-SmB_Lu016		20.1	0.0	36.6	43.3	100
Tralpt_Lu019		17.7	1.4	24.1	56.8	100
Tralpt_Lu020		34.3	1.6	7.2	56.9	100
SmB_Lu023		4.5	2.5	22.0	71.0	100
SmB_Lu024		19.1	1.7	14.5	64.7	100
SmD3pt_Lu025		13.3	1.4	30.7	54.6	100
SmD3pt_Lu026		40.7	2.3	18.6	38.5	100

Table 2.3: **Mappable and unmappable read statistics in random hexamer primed libraries.** The random hexamer-primed libraries were used for the enrichment analysis presented in the main text. The first four libraries were prepared without size-selection, whereas the latter ones were size-selected to remove primer-dimers. The rRNA reads were quantified by mapping total reads to a single copy of the 45S rRNA repeat unit and 5S rRNA repeat unit, allowing no-mismatches. The reads derived from primes/adaptors were quantified using the fast-qc program. Taken together, these three categories encompass 30% - 90% of each library, whereas the balance (i.e. the 'Other' column) are unmappable reads. The vast majority of reads in the 'Other' category are random sequences that do not match to any known genome, along with a few trace contaminants (e.g. bacterial Propionibacterium acnes, or rainbow trout, Oncorhynchus mykiss sequences). Among each pair of experiments shown in the table, the rRNA and primer-dimer reads are more abundant in the control libraries (odd numbers) than they are in the IP libraries (even numbers). This is to be expected due to the limited amount of RNA brought down in the IPs (or control/mock IPs).

the majority of RNA species are not associated with Sm proteins, we normalized the remaining transcripts against the median of all enrichment ratios: %(raw_IP + 2)/(raw_Ctrl + 2). After normalization, we defined the enrichment ratio as (norm_IP + 2)/ (norm_Ctrl + 2). The use of median-normalized raw read numbers is similar to the upper-quartile normalization method used by others (Bullard *et al.*, 2010). In this way, we made a conservative estimate of the enrichment of RNAs in IPs versus controls.

To visualize the enrichment data, scatter plots were constructed using the log-transformed and normalized read numbers. Data for the native SmB-associated RNAs (Oregon R, Y12 IPs) are shown in Figure 2.3a; data for the other Sm protein constructs are presented in Figure 2.2. In any co-IP experiment, there are two populations of molecules: those that interact specifically

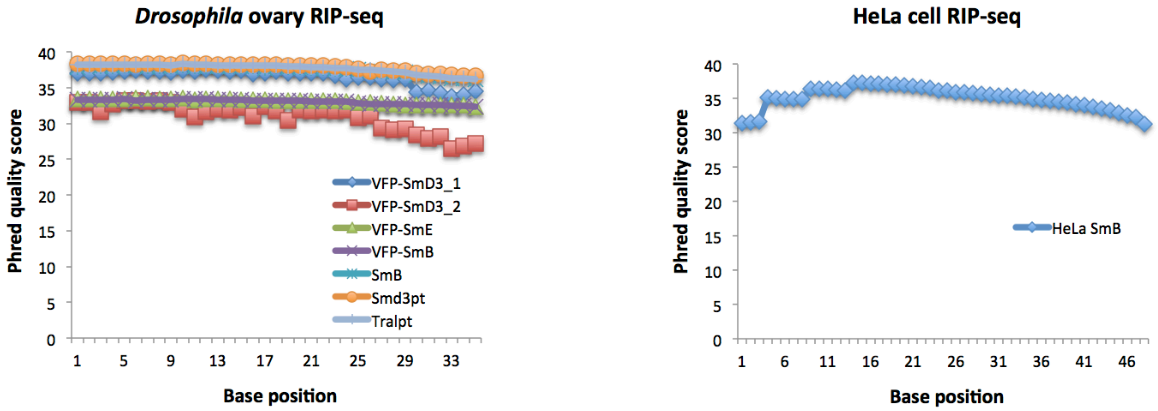


Figure 2.2: **Per base quality of the *Drosophila* (35nt) and human (48nt) RIP-seq data calculated using FastQC** (reads are pooled for each experiment).

with the antibody and those that stick non-specifically to the beads. Non-specific interaction was observed for many transcripts, as depicted by the main cluster along the diagonal line (Figure 2.3a). The dots located above the main cluster represent the enriched RNAs. In order to objectively identify Sm-associated RNAs, we employed Gaussian mixture modeling (Pearson, 1894), which has been used to analyze RIP-chip experiments (Morris *et al.*, 2008). Distributions of the enrichment ratios were first plotted as histograms. Next, we used mixtools to fit a combination of two Gaussian functions to the enrichment ratio distribution (Benaglia *et al.*, 2009).

As shown in Figure 2.3b, the distribution of the logtransformed enrichment ratios (red line) can best be explained by two different Gaussian functions, one that corresponds to the background RNAs (black dotted line) and one that represents the Sm-associated RNAs (blue dotted line). The cutoff between Sm-associated and background mRNAs was defined by the log of the odds (LOD) ratio between the two Gaussian functions. The transcripts with a LOD >1 (that is, those that had a greater likelihood of being in the Sm distribution) were considered to be Sm-associated RNAs. Using this threshold, we then mapped these assignments back onto the scatter plots. As shown in Figure 2.3a (blue dots), the enriched RNAs are clearly seen to be above the diagonal (black dots represent the background distribution). This same analysis was performed on the other Sm protein datasets, with strikingly similar results (Figure 2.4). Thus, the Gaussian mixture modeling procedure provides an unbiased and less arbitrary method for

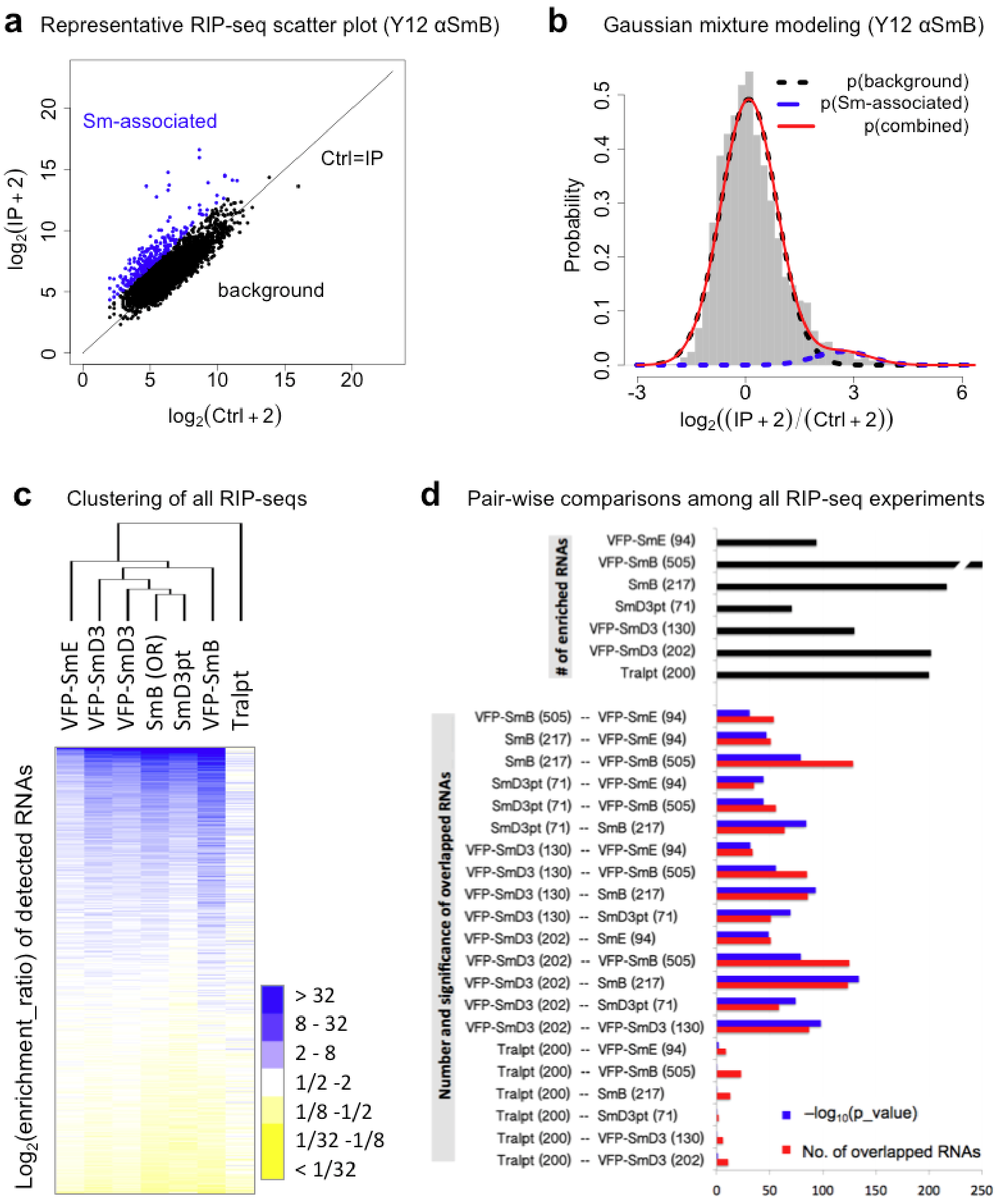


Figure 2.3: **RIP-seq data analysis.** (a) Scatterplot of a control (Ctrl)-IP pair of RIP-seq data (SmB IP Lu023-Lu024), where normalized and log-transformed read numbers for each known transcript in an IP are plotted against that of Ctrl (Ctrl + 2 and IP + 2 to avoid division by zero). Black dots represent background RNAs, while the blue dots represent enriched RNAs, as determined by Gaussian mixture modeling. Only RNAs with read coverage >10 are plotted. See Figure 2.2 for the rest of the scatterplots. (figure legend continued on the next page)

Figure 2.3: (b) Gaussian mixture modeling of the RIP-seq data (SmB IP), where the enrichment ratios for all the transcripts were plotted as a histogram (in gray) and fitted with a combination of two Gaussian curves. (c) Log-transformed enrichment ratios of the 5,296 RNAs (with coverage d>10) in all 7 experiments were clustered (average linkage clustering using correlation (uncentered) as similarity metric) and visualized as a heat map. (d) Pair-wise comparisons among all seven experiments. Numbers of enriched RNAs are listed next to the experiment labels. Black bars, number of enriched RNAs in each experiment; red bars, number of overlapped RNAs in each pair; blue bars, negative log10 transformed Fisher's exact test P-values (within a superset of 5,296 RNAs). See Figure 2.4 for pairwise comparisons excluding non-coding RNAs.

identifying enriched RNAs (Morris *et al.*, 2008). Using the aforementioned analysis pipeline, we identified roughly 200 Sm-associated RNAs in any given RIP-seq experiment, representing 0.7% of the Drosophila transcriptome, or 4% of the significantly expressed transcripts.

2.3.3 A multi-targeting RIP strategy identifies Sm-associated RNAs

To assess the robustness and reproducibility of the Drosophila RIP-seq experiments and analysis pipeline, we visualized the log-transformed enrichment ratios for the transcripts with a read coverage greater than 10. Out of the >15,000 annotated genes in the fruitfly genome, 5,296 of them showed sufficient read depth (d >10). To determine the relationship between the profiles of the seven RIP-seq experiments without prior assumptions, we performed an unsupervised hierarchichal clustering analysis. The top of the map represents RNAs that are significantly enriched (Figure 2.3c). As shown by the dendrogram (Figure 2.3c) and consistent with expectation, the six canonical Sm protein RIP-seq experiments clustered together, whereas the data from the Tral IP formed an outgroup. The most-highly enriched transcripts among the random hexamer-primed libraries from six Sm IP experiments (including one VFP-SmD3 biological replicate) revealed extensive overlap. Detailed analysis showed that 25 RNAs (9 snRNAs, 16 mRNAs) were common among all 6 Sm protein IPs, and 52 transcripts (12 snRNAs, 40 mRNAs) were shared in 5 of the 6 (see Table 2.5 for detailed enrichment ratios). The top 86 transcripts (13 snRNAs, 1 small nucleolar RNA (snoRNA), and 72 mRNAs) were shared by at least 4 of the experiments. Since four Drosophila snRNAs (U1, U2, U4, and U5) have multiple variant paralogs, we reassigned uniquely mappable reads to them and we found that all of the snRNAs with significant coverage are enriched in all Sm IPs (Table 2.6). In addition,

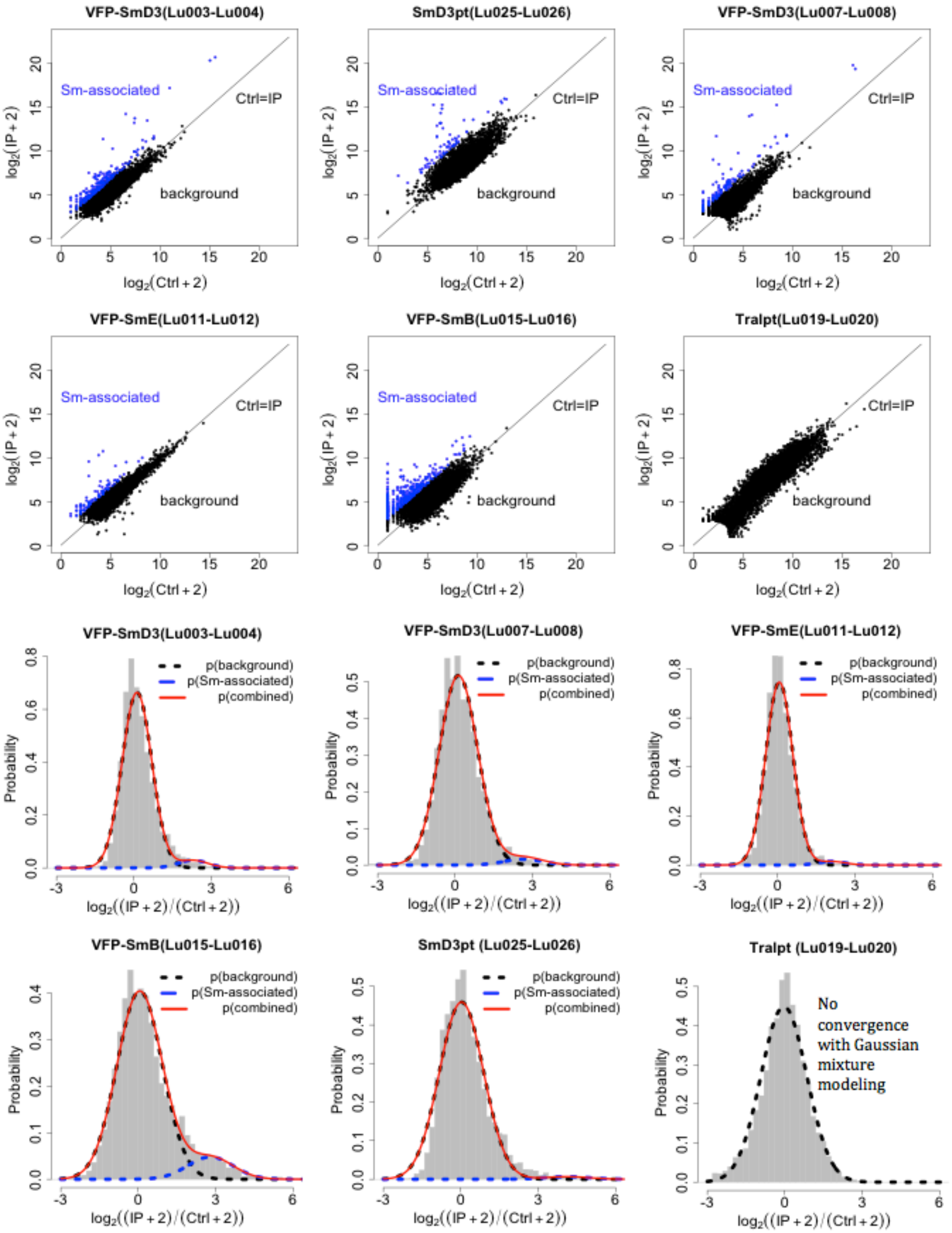


Figure 2.4: **Additional Scatterplots and Gaussian mixture modeling plots .**

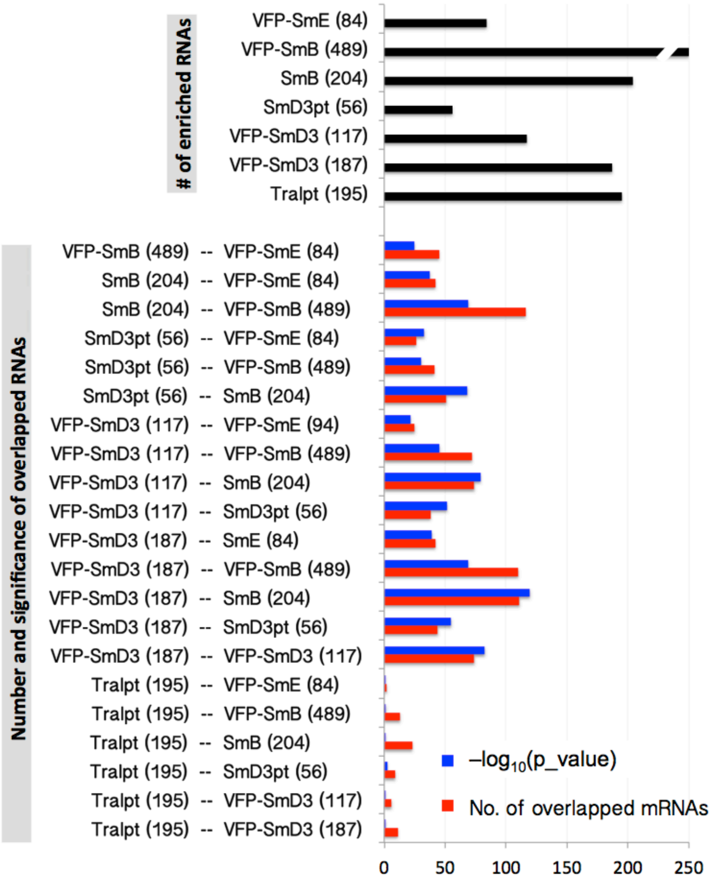


Figure 2.5: **Comparisons among all RIP-seq experiments, excluding ncRNAs.** Numbers of enriched RNAs are listed next to the experiment labels. Black bars: number of enriched RNAs in each experiments; red bars: number of overlapped RNAs in each pair; blue bars: negative log10 transformed Fisher's exact test p-values (within a superset of 5270 RNAs).

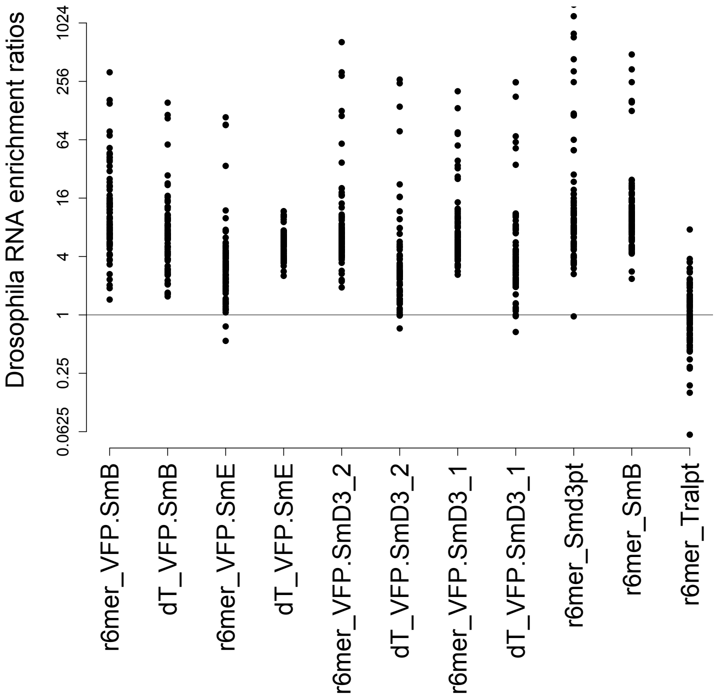


Figure 2.6: **Enrichment ratios of the consensus set of Sm-associated RNAs, plotted by experiment and priming methods.** Tralpt is used as control.

we analyzed the consensus set of 86 Sm-associated RNAs in the oligo(dT)20 primed libraries, and we found that they are also highly enriched, despite the lower number of mappable reads (Figure 2.6). Thus, our multi-targeting RIP-seq approach is robust despite the differences in library statistics (Table 2.2). We operationally defined the Sm-associated RNAs as being those that were enriched in at least four of the six experiments.

Next, we carried out pair-wise comparisons among the seven RIP-seq experiments and performed Fisher's exact test to assess the significance of any overlapping subsets (Figure 2.3d).

Experiments	No. RNAs (d>10)		mRNA reads		snRNA reads	
	OligodT	6mer	OligodT	6mer	OligodT	6mer
Lu001-Lu004	1478	6545	57603	706143	3250538	3352434
Lu005-Lu008	693	5384	13111	301516	1151580	1840716
Lu009-Lu012	7065	6780	653812	1175978	11002	24683
Lu013-Lu016	6980	7087	394807	919921	13805	13499

Table 2.4: **Comparison of oligo(dT) and random hexamer primed libraries.** Oligo(dT) primed libraries produced fewer mRNAs with significant coverage, and fewer reads for mRNAs.

Interestingly, among the top 200 RNAs in the Tral IP experiment, very few of them overlapped with any of the RNAs that associated with canonical Sm proteins. As seen in the heat map (Figure 2.3c), the enrichment ratios for the VFP-SmE IP were typically lower than those of the other Sm proteins. However, the pairwise comparisons show that SmE associates with a similar group of RNAs (see also Figure 2.6). The overlaps between the different Sm protein IPs were highly significant, as shown by their extremely small Pvalues (10E-32 to 10E-135, plotted as negative logarithms; Figure 2.3d). Even when all of the snRNAs were taken out of the pairwise comparisons, the P-values remained extremely small (Figure 2.3d; Figure 2.5). Despite the different experimental parameters (tagged versus untagged, native versus ectopic, and so on), the lists of enriched RNAs are essentially the same. This high degree of reproducibility suggests that the multi-subunit targeting approach is superior to the conventional biological replication of experiments for RNP analysis. Indeed, the variability between biological replicates was greater in the case of VFP-SmD3 than it was between some of the other RIPs (Figure 2.3c). Collectively, these data demonstrate a high degree of specificity in the Sm protein IPs, showing that canonical Sm proteins co-precipitate with essentially the same set of mRNAs.

2.3.4 Sm proteins associate with three major classes of RNAs

The RIP-seq experiments in both *Drosophila* and human cells confirmed the well-studied snRNAs as major targets of Sm proteins, and in addition indicate novel classes of Sm targets. A detailed analysis of the known and newly discovered RNAs from our study suggests that Sm proteins associate with three major classes of RNAs (Figures 2.7 and 2.8; Figures 2.6 and 2.12).

2.3.5 RIP-seq identifies Sm class snRNAs

The Sm-associated transcripts and their enrichment ratios are listed in Figure 2.7. As expected, all spliceosomal snRNAs were among the top-scoring transcripts in terms of their enrichment ratios. The only missing Sm class snRNA from the list of Sm-associated RNAs is U7 snRNA, because it is too short (71 nucleotides in *Drosophila*, and 63 nucleotides in human) to be included in the size-selected cDNA libraries (Figure 2.7a; Table 2.5) (Dominski *et al.*, 2003;

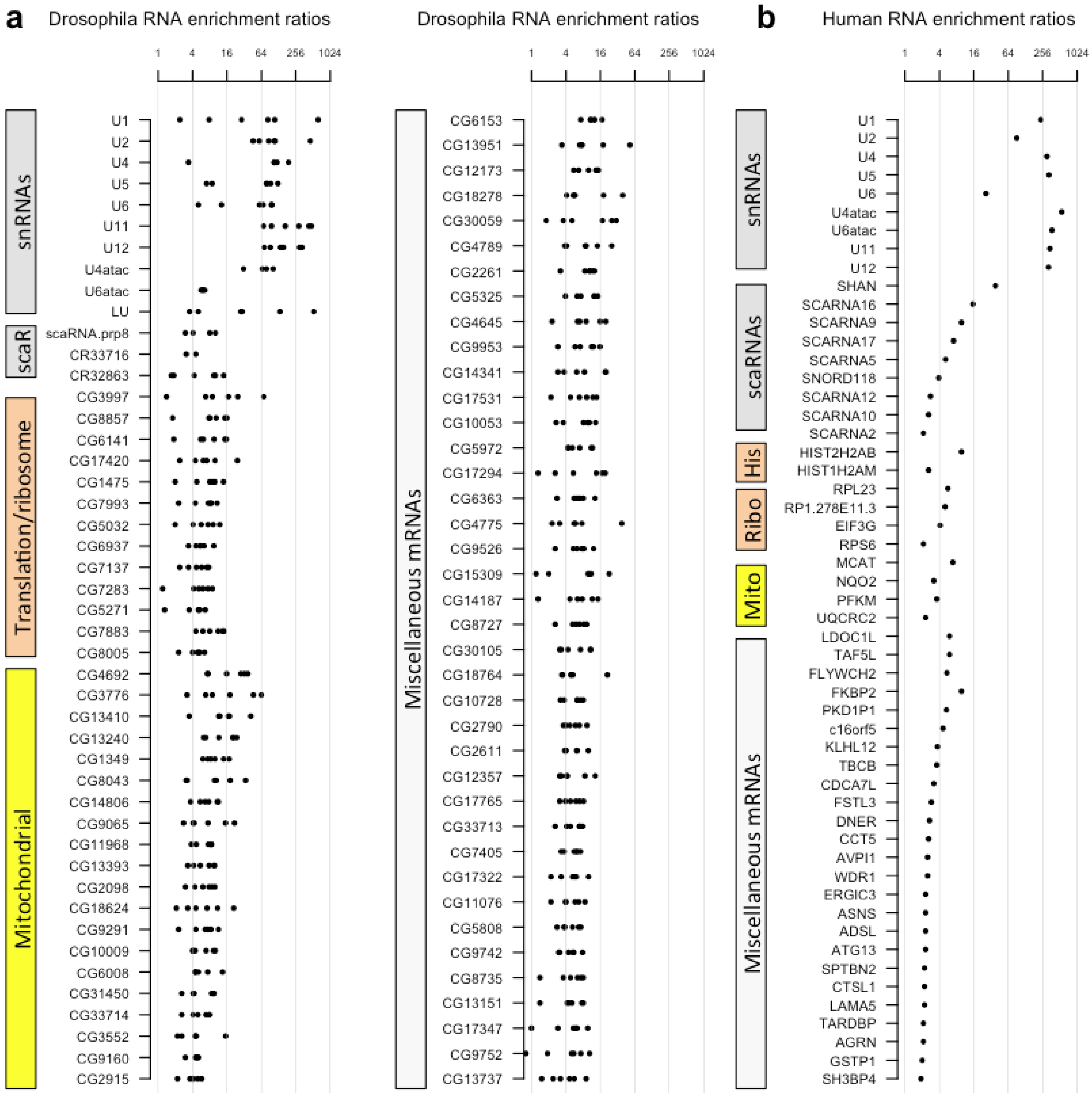


Figure 2.7: **Three categories of Sm-associated RNAs in Drosophila and human.** Different categories of Sm-associated RNAs are color-coded. (a) Drosophila Sm-associated RNAs, with enrichment ratios from all six Sm RIP-seq experiments. For snRNAs with multiple distinct paralogs (U1, U2, U4 and U5), all the reads were pooled for calculation of enrichment ratios. The three U6 paralogs are identical in sequence. See Table 2.6 for assignment of reads to distinct paralogs. U7 was not plotted due to low read coverage. See Table 2.5 for detailed enrichment ratios. (b) Human Sm-associated RNAs. Medians of enrichment ratios were plotted for snRNAs with multiple paralogs. See Table 2.7 for detailed enrichment ratios.

Mowry and Steitz, 1987). Other highly abundant non-coding RNAs (ncRNAs; for example, 7SK snRNA, SRP RNA, 5.8S ribosomal RNA and so on, data not shown) were not enriched in the IPs, demonstrating the specificity of the approach. Multiple distinct paralogs exist for four of the *Drosophila* snRNAs, U1, U2, U4 and U5, and they share long stretches of identical regions (Figure 2.11). In order to accurately analyze each paralog without the confounding repetitive reads, we reassigned uniquely mappable reads to U1, U4 and U5 paralogs (Table 2.6). We used the variant nucleotides in U2 to calculate the fractions of each isoform and redistribute the total number of U2 reads among the gene paralogs. Not surprisingly, all snRNAs with significant read coverage are enriched in the IPs (Table 2.6). With regard to the HeLa cell analysis, there are hundreds of snRNA genes in the human genome, and only a small fraction of them are properly annotated. Not surprisingly, most of the annotated human spliceosomal snRNAs were identified in our IPs, all of which have very high enrichment ratios (Figure 2.7b).

ERANGE analysis and manual inspection of the *Drosophila* RIP-seq data revealed several clusters of reads that could not be mapped to gene models. Four of them are new genes that had not been previously annotated. During preparation of this manuscript, two transcriptomic studies have since identified these putative new transcripts (Graveley *et al.*, 2011; Jung *et al.*, 2010): CR43708, CR43600, snoRNA:2R:9445410 (CR43574) and snoRNA:2R:9445205 (CR43587). Two of the four novel transcripts, CR43708 and CR43600, showed significant enrichment in the IPs.

We characterized the two Sm-associated ncRNAs and found that one, CR43708, has features typical of an snRNA. CR43708 is located in the second intron of *fas2* (CG3524, fatty acid synthase 2), a homolog of the human fatty acid synthase gene (Figure 2.9a). We defined the accurate 5' and 3' ends of CR43708, and found that this transcript is 116 nucleotides long (ZL and AGM, unpublished). Detailed analysis of sequences upstream of CR43708 revealed conserved proximal sequence elements PSEA and PSEB, highly similar to Sm-class snRNA promoters (Figure 2.9a; Figure 2.10a) (Hernandez, 2001; Jensen *et al.*, 1998). To examine the subcellular localization of CR43708, we carried out in situ hybridization in *Drosophila* S2 cells and found that this RNA accumulates in the nucleus (Figure 2.9c). Using the transcribed region and the promoter sequences, we searched genome and transcriptome databases for homologs. We

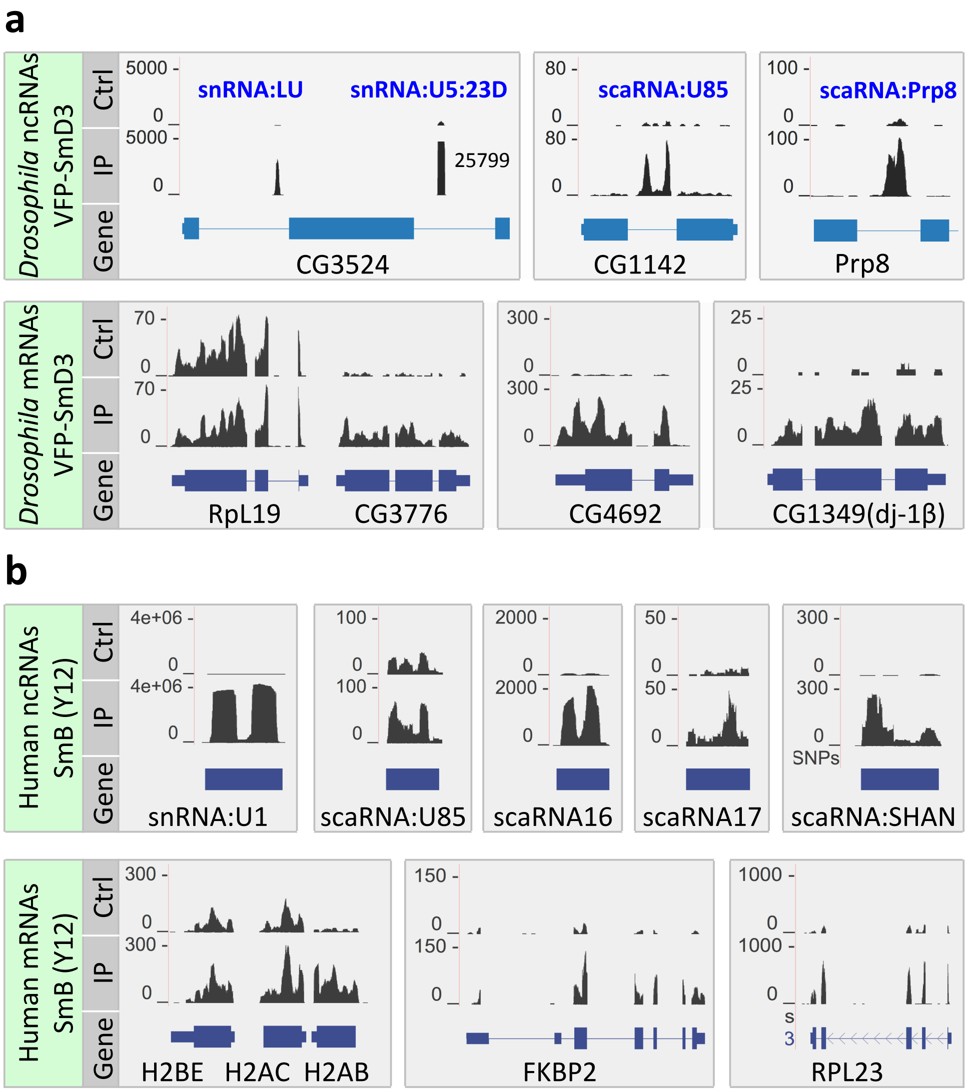


Figure 2.8: **Examples of the three categories of Sm-associated RNAs in *Drosophila* and human.** For genes with multiple transcripts, the gene model that is most similar to the read coverage pattern is shown. The y-axis corresponds to the normalized number of reads per nucleotide. (a) Examples of *Drosophila* Sm-associated RNAs from VFP-SmD3, control (Ctrl; Lu003) and IP (Lu004). For the non-coding RNAs that are associated with Sm proteins, their host genes are also shown. The read coverage for U5:23D is off scale, and thus truncated. (b) Examples of human Sm-associated RNAs from Y12 (SmB), Ctrl (Lu045) and IP (Lu047). The histone mRNAs H2BE, H2AC and H2AB are short for HIST2H2BE, HIST2H2AC and HISTH2AB, respectively.

Gene ID	Annotation	Enrichment ratios					
		VFP-SmD3	VFP-SmD3	VFP-SmE	VFP-SmB	SmB	SmD3pt
snRNAs							
Five U1	U1	28.9	7.8	2.4	110.2	84.2	633.5
Six U2	U2	59.0	111.3	87.1	46.0	108.7	459.1
Three U4	U4	109.3	120.3		3.4	107.8	191.6
Seven U5	U5	78.7	93.6	8.8	7.1	124.6	81.4
Three U6	U6	97.8	59.6	13.0	5.1	96.9	69.1
CR34151	U11	167.0	290.7	96.9	70.1	484.1	433.9
CR32162	U12	72.4	299.3	92.4	137.8	155.1	329.3
CR32860	U4atac	31.4	66.4			79.0	103.6
CR32989	U6atac	6.3	6.0			6.6	5.7
CR43708	LU	135.7	530.6	3.6	5.1	29.8	27.9
scaRNAs							
CR32863	snoRNA:MeU5-C46 (U85)	10.1	1.9	1.7	14.1	4.3	9.4
CR33716	snoRNA:MeU5-U42	4.6	3.1				
CR43600 (new)	scaRNA:Prp8	10.1	4.1				8
Mitochondrial							
CG4692	ATP synthase	32.5	37.2	7.5	7.3	15.9	28
CG3776	Jhebp29	8.9	6.8	3.2	46.5	18.1	63.9
CG13410	mRpL35	12.1	17	3.5	41.9	11.5	17.7
CG13240	NADH dehydrogenase subunit	11.6	20.3	6.3	6.8	20.8	23.6
CG1349	dj-1beta/PARK7	9.9	14.1	7.4	6.1	8.4	17.5
CG8043	IBA57, Fe/S assembly	3.2	18.3	3.1	34.2	10.5	9.7
CG14806	Apopt1	6.7	3.7	5.5	11.4	10.8	7.7
CG9065	cox17	2.8	4.1	4.3	21.6	15.2	7.5
CG11968	Ras-related GTP binding A	8.4	4.7	3.8	8.4	8.6	7.6
CG13393	DAD1, phospholipase	4.2	5.5	3.3	9.8	7.8	9.4
CG2098	ferrochelatase	8.8	4.4	3	6.1	9.9	7.8
CG18624	NADH dehydrogenase subunit	3.3	4.6	2.1	21	11	7.2
CG9291	elongin-C	6.7	8.4	4.6	2.3	11.4	7.4
CG10009	NOA36/ZNF330	4.1	4.5	4	6.9	10.1	9.3
CG6008	NADH dehydrogenase subunit	4.5	4.5	4.5	7.4	5.1	13.5
CG31450	mRpS18A	4.3	9.5	4.1	2.6	8.6	9.3
CG33714	RNA binding protein	5	7.9	2.6	4.1	7.8	6.9
CG3552	GDP-D-glucose phosphorylase	4.5	4.5	2.2	15.3	4.7	2.6
CG9160	mtacp1	4.8	4.7	3	5.1	4.5	5
CG2915	carboxypeptidase M14 like	3.6	5	2.2	4.2	5.8	4
Translation							
CG3997	RpL39	6.8	17	1.4	70.7	24.7	8.9
CG8857	RpS11	8	10.4	1.8	16	14.8	7.8
CG6141	RpL9	5.6	6.2	1.9	14.8	15.7	9.5
CG17420	RpL15	4.5	7.1	2.4	6.2	24.3	9.9
CG1475	RpL13A	4.8	9.9	2	14	8	9
CG7993	rpf2, ribosome production	4.5	7.6	2.3	8.5	8.2	10.8
CG5032	rRNA methyltransferase	5.7	4.1	2	12.1	9.2	7.5
CG6937	MKI67IP, ribosome biogenesis	4.6	6.4	3.4	5.5	9.4	5.5
CG7137	rRNA-processing protein 8	5.8	7.1	2.4	7.7	4.7	3.4
CG7283	RpL10Ab	4.2	5.2	1.2	8.9	6.1	7.5
CG5271	RpS27A	5.1	6.7	1.3	5.4	5.2	3.5
CG7883	eIF2Balpha	8	6.1	4.6	14.2	13.5	11.2
CG8005	Dhps, eIF5A modification	4	2.3	5.2	6.5	4.9	5.4

Table 2.5: **Enrichment ratios of *Drosophila* Sm-associated RNAs**

recovered matches in nine species, all of which are in the melanogaster group of the *Drosophila* genus, and all are located within the same intron of the *fas2* gene (Figure 2.9e,f). Among the sequenced *Drosophila* species in the melanogaster group, the *Drosophila erecta* genome does not appear to contain CR43708, suggesting that it may have been lost. Interestingly, we found a truncated version of this gene within an intron of the Ac3 gene in D. melanogaster (Figure

Gene ID	Annotation	Enrichment ratios					
		VFP-SmD3	VFP-SmD3	VFP-SmE	VFP-SmB	SmB	SmD3pt
Miscellaneous							
CG6153	pithd1	12.4	10.8	7.3	17.1	12.7	10.7
CG13951	Zfp511	7	7.6	3.4	52.6	17.9	7.6
CG12173	enoph1	5.5	6.7	9.9	14.2	13.3	14.9
CG18278	N-acetylglucosamine-6-sulfatase	5.8	5.4	5.6	39.7	18.2	4.1
CG30059	N-acetylglucosamine-6-sulfatase	25.3	17.4	1.8	30.4	5.1	3.6
CG4789	Rabl3	3.8	8.7	4.1	25.2	9	14.3
CG2261	CstF-50, WD40	10.5	8.6	3.2	10.2	11.7	12.7
CG5325	Pex19	14.5	7.3	3.9	6.3	12.5	12.4
CG4645	yipf1	9	6.4	2.3	7	19.9	15.9
CG9953	peptidase s28	7.1	5.8	2.9	15.8	12.1	11.2
CG14341	only in arthropods	3.7	20.1	2.9	19.5	6.2	8.4
CG17531	GstE	9.2	4.8	2.2	13.8	6.9	11.8
CG10053	ccdc75/CENP-Y	8.1	2.7	3.6	10.2	13.2	9.5
CG5972	arp-p20	11	7	4.4	4.3	5.2	11.7
CG17294	HDHD2	2.6	5.4	1.3	13.6	17.3	19.5
CG6363	MRG15	5.6	6.3	2.8	8	12.9	7
CG4775	Tango14, alkyl/aryl transferase	3.1	5.6	2.3	38.2	5.9	7.6
CG9526	frj, mboat family protein	5.4	6.3	2.6	12.2	7.6	8.2
CG15309	Yippee-like	9.7	10.2	1.2	2	22.8	11
CG14187	only in Drosophila species	4.7	6.4	1.3	7.7	14.6	11.5
CG8727	Cycle/dBMAL	5.2	6	2.6	9.2	8.2	6.9
CG30105	rnaseH2	7.2	4.3	3.1	3.3	10.6	10.8
CG18764	only in Drosophila species	5.3	3.4	4.9	21.2	4.9	3.5
CG10728	vls	6.4	6.3	3.2	3.7	7.2	8
CG2790	Hsp70 binding	5.9	4	3.6	6.9	9.3	4.7
CG2611	only in arthropods, DUF872	3.8	3.9	4	9.8	6.3	6.2
CG12357	CBP20	4.2	4.1	3.1	8.6	13	3.3
CG17765	Ca-binding, EF hand	6	3.9	3.1	4.8	8.1	6.9
CG33713	Acbd6, acyl-CoA binding	4.8	7	2.6	4.1	7.8	6.9
CG7405	cyclin H	5.6	7.2	3.3	3.7	6.1	6.4
CG17322	UDP glucosyl transferase	5.2	2.2	3.3	9.9	6.1	5.7
CG11076	only in Drosophila species	4	5.8	2.2	6.7	8.6	3.9
CG5808	PPIL4, cyclophilin	3.7	2.8	3.8	6.7	7.3	5.2
CG9742	SmG	5.5	7.8	3.1	5.3	4.4	3
CG8735	LNP1, zinc finger, transmembrane	3.6	6.4	1.4	8	7.3	4.8
CG13151	AT-hook, DNA binding	4.5	4.2	5.1	1.4	8	7.6
CG17347	dynactin6	6.3	5.9	2.9	5.4	9.6	1
CG9752	c9orf64	5	5.5	0.8	1.9	10.3	7.3
CG13737	only in Drosophila species	3.2	4.6	1.5	5.6	2.4	9

Table 2.5: **Enrichment ratios of *Drosophila* Sm-associated RNAs** U1, U2, U4, U5 and U6 are multi-copy snRNAs, and for each snRNA species, all reads mapped to all the paralogs are pooled for calculation of enrichment ratios. Please refer to the methods section for details of the analysis. Please refer to Table 2.5 for assignment of reads to distinguishable snRNA paralogs. Empty spaces means that there are not enough reads for that particular snRNA or scaRNA. Enrichment of the scaRNAs and mRNAs as determined by Gaussian mixture modeling is indicated by the highlighting. Note: even though these RNAs are not deemed significantly enriched in some of the samples (unhighlighted cells), their enrichment ratios are mostly larger than 1.

2.10c). The homology extends through the first 70 bp of CR43708, and lacks the promoter and the 3' end, suggesting that this paralog is a pseudogene. The predicted secondary structure of CR43708 closely resembles that of a canonical snRNA, including the presence of 5' and 3' end stem loops that flank a putative Sm binding site (Figure 2.9c). Structured sequence alignments clearly show that the putative Sm binding site (except in *Drosophila kikkawai*) and the terminal stem loops are well conserved. In addition, we identified many covariant base pairs within

	VFP-SmD3_1 Oligo(dT)		VFP-SmD3_1 6mer		VFP-SmD3_2 Oligo(dT)		VFP-SmD3_2 6mer		VFP-SmE Oligo(dT)		VFP-SmE 6mer		VFP-SmB Oligo(dT)		VFP-SmB 6mer		SmB 6mer		SmD3pt 6mer	
snRNA paralogs	**Lu001**	**Lu002**	**Lu003**	**Lu004**	**Lu005**	**Lu006**	**Lu007**	**Lu008**	**Lu009**	**Lu010**	**Lu011**	**Lu012**	**Lu013**	**Lu014**	**Lu015**	**Lu016**	**Lu023**	**Lu024**	**Lu025**	**Lu026**
U1:21D,U1:95Ca,U1:95Cb	2676	71161	1785	193643	4824	27630	4467	39845	15	49	9	26	0	42	0	26	1692	85180	330	93882
U1:82Eb	278	16868	330	31305	1069	7143	880	9741	1	6	5	1	0	6	0	3	58	3068	21	4932
U1:95Cc	44	1852	37	3923	117	868	92	1275	0	2	0	2	0	1	0	0	22	1329	8	1869
U2:14B	531	26255	654	43704	84	7087	69	10009	1	380	6	292	9	188	0	214	110	11935	94	43015
U2:34ABa	1766	87235	2172	145213	281	23548	229	33255	4	1262	21	970	29	623	0	713	364	39656	311	142924
U2:34ABb,U2:34ABc	5017	247854	6171	412580	797	66904	651	94485	11	3586	60	2755	81	1771	0	2025	1035	112672	883	406078
U2:38ABa	333	16440	409	27367	53	4438	43	6267	1	238	4	183	5	117	0	134	69	7474	59	26936
U2:38ABb	67	3309	82	5508	11	893	9	1261	0	48	1	37	1	24	0	27	14	1504	12	5421
U4:25F	0	13	0	12	0	5	0	9	0	0	0	0	0	0	0	0	0	7	0	19
U4:38AB	1	156	0	295	0	63	0	134	0	0	0	0	0	0	0	1	0	104	0	241
U4:39B	5	268	3	386	1	70	1	146	0	1	0	1	0	1	0	2	0	174	1	250
U5:14B	0	43	0	56	0	3	0	11	0	0	0	0	0	0	0	0	0	4	0	0
U5:23D	0	73	1	69	0	5	0	18	0	0	0	0	0	0	0	0	0	5	0	3
U5:34A	0	70	0	13	0	2	0	4	0	0	0	0	0	0	0	0	0	5	0	0
U5:35D	0	6	0	3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
U5:38ABa	0	49	2	47	0	4	0	7	0	0	0	0	0	0	0	0	0	1	0	1
U5:38ABb	0	55	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
U5:63BC	1	195	6	165	2	12	1	37	0	0	0	0	0	0	0	0	0	2	0	2

Table 2.6: **Assignment of unique reads to *Drosophila* snRNA paralogs** (except U2, which is reassignment of all reads, please see the methods section for details). snRNA paralogs with identical sequences are listed in one row. The highlighted columns are the IPs, whereas the unhighlighted ones are the controls.

the two stem loops, supporting the predicted secondary structure (Figure 2.9f). Uridine-rich, Sm-class snRNAs such as U1 and U2 are known to contain a trimethyl-guanosine (TMG) 5' cap structure that is generated upon formation of the Sm core RNP (Matera *et al.*, 2007). As expected, CR43708 was efficiently immunoprecipitated by anti-TMG antibodies (Figure 2.16a). Taken together, these features led us to conclude that this transcript is a novel Sm-class snRNA, which we termed snRNA:LU (Like U).

Interestingly, the U5:23D snRNA gene is located near LU, within a neighboring intron of the *fas2* protein coding gene (Figure 2.9a). We were unable to deduce the precise origin of LU; however, its juxtaposition with U5:23D suggests that it could have evolved from a U5 gene duplication, followed by rapid divergence. Supporting this notion, the 3' end stem-loops of the LU snRNA homologs are quite similar to those of U5 snRNAs (Figure 2.10), although there is a lack of overall sequence similarity between the two genes.

To study the function of LU snRNA, we first considered the possibility that it might base pair with other snRNAs, as we found a nearly invariant single-stranded region located in the middle of LU snRNA (Figure 2.9d,f). Notably, we identified extensive base complementarity

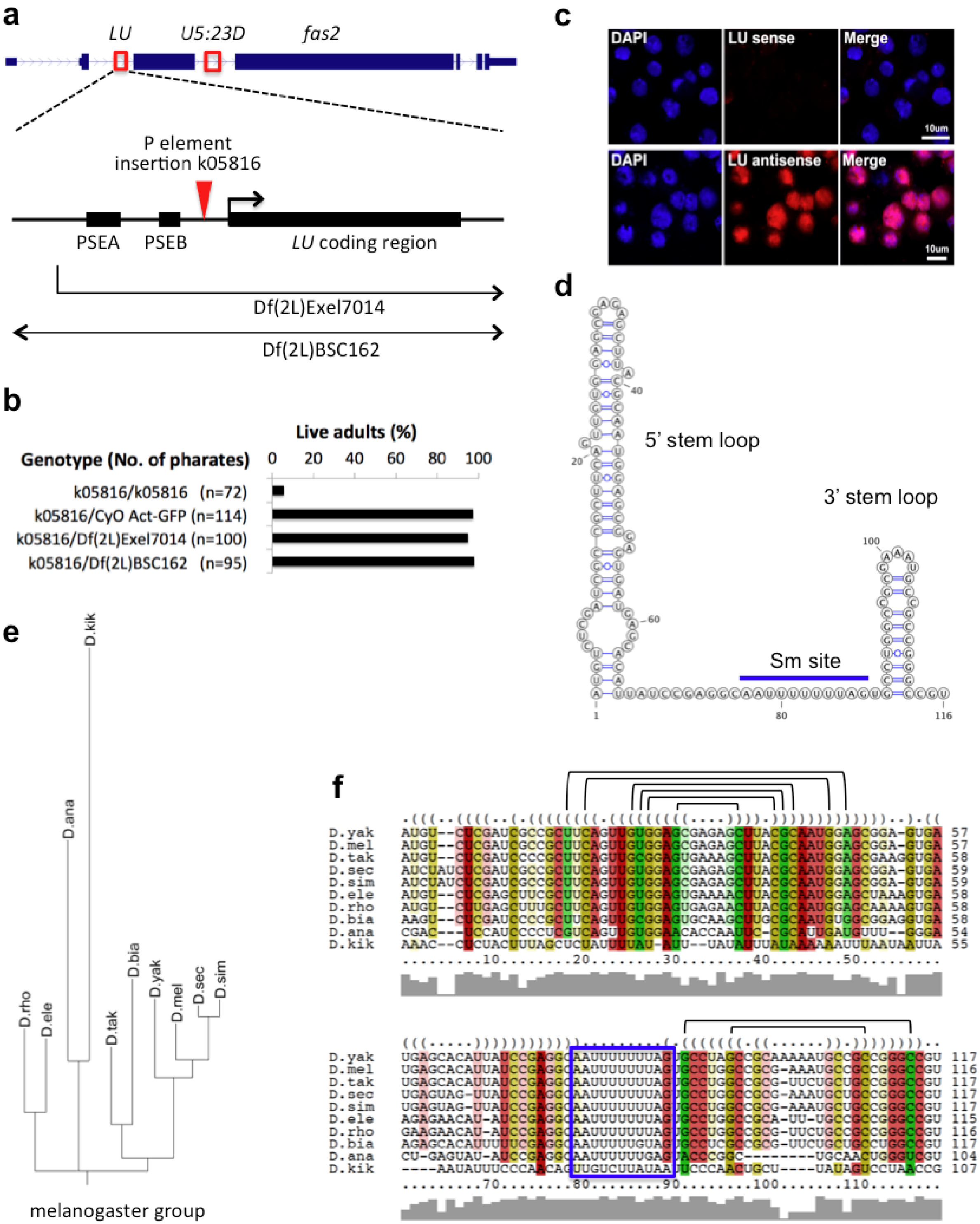


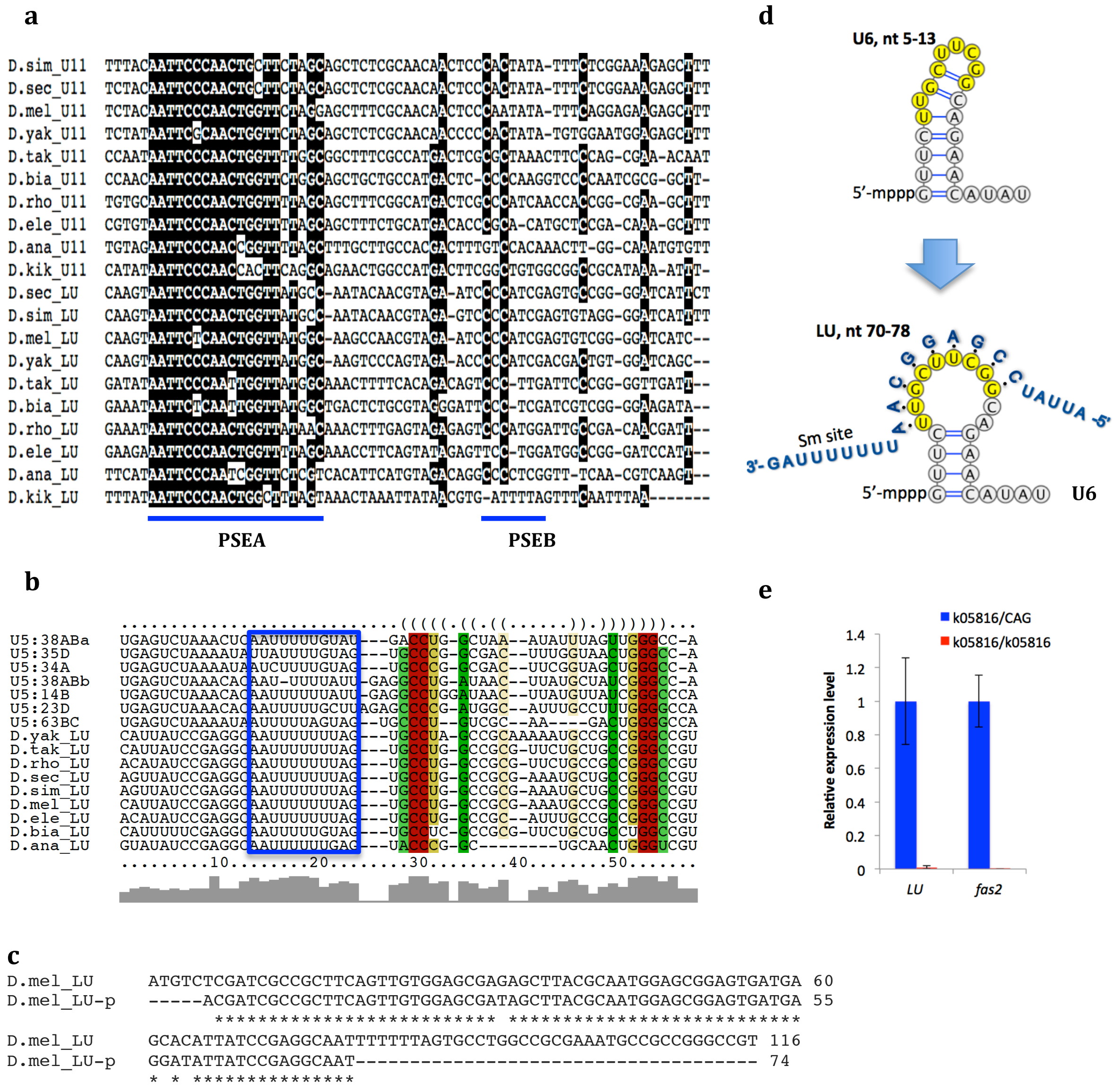
Figure 2.9: **Characterization of the *Like-U* (*LU*) snRNA gene.** (Figure legend on the next page.)

Figure 2.9: **Characterization of the *Like-U* (*LU*) snRNA gene.** (a) Genomic and genetic contexts of the LU snRNA locus. LU snRNA is encoded within the second intron of fas2; U5:23D is located in the third intron. PSEA/PSEB, proximal sequence element A/B (see Figure 2.10 for alignment of the U11 and LU promoters in Drosophilids). Locations of a P-element insertion and two deficiencies are indicated. The arrows on the deficiencies indicate that the regions extend beyond the displayed area. (b) Complementation analysis of LU snRNA mutations and deficiencies. Numbers of third instar larvae are indicated in parentheses. (c) Localization of LU snRNA in S2 cells determined by in situ hybridization using LU sense and antisense probes. (d) Predicted secondary structure of D. melanogaster LU snRNA. (e) Phylogeny of LU snRNA. (f) Alignment of Drosophilid LU snRNA orthologs using LocARNA. The blue box indicates the Sm site. Half-brackets indicate covariant base pairs

between this region of LU and the 5' end of U6 (Figure 2.10d). This putative base-pairing suggests that LU may be involved in splicing regulation. We identified four independent transposon insertions in and around the LU gene locus (see Materials and methods), and we confirmed that one of these insertion lines, *fas2[k05816]*, disrupts expression of both the fas2 host gene and the LU snRNA gene (Figure 2.9a; Figure 2.10e). Although homozygotes die around eclosion; complementation analysis between fas2k05816 and two other deletion lines uncovering this region suggests that neither the fas2 host gene nor the LU snRNA gene are required for organismal viability (Figure 2.9b). We conclude that, although it may well contribute to organismal fitness, LU is not an essential gene. This conclusion is supported by the independent loss of LU snRNA in D. erecta. Taken together, our RIP-seq analysis of Sm proteins reveals that a total of 11 distinct species of Sm-class snRNAs are present in *Drosophila*: U1, U2, U4, U5, U6, U7, U4atac, U6atac, U11, U12 and LU.

2.3.6 Sm proteins associate with scaRNAs

scaRNAs are ncRNAs that guide methylation and pseudouridylation of snRNAs, the specificity of which is determined by base-pairing with targets (Darzacq *et al.*, 2002). A previous study showed that in human cells, several scaRNAs specifically associate with SmB and SmD3, including U85, U87, U89 and human telomerase RNA (hTR) (Fu and Collins, 2006). Coprecipitation of SmB/D3 with these scaRNAs was shown to require the conserved CAB box (Fu and Collins, 2006), which is essential for scaRNA localization to Cajal bodies (Richard



```
D.sim_U11  TTTACAATTCCCAACTGCTTCTAGCAGCTCTCGCAACAACTCCCACTATA-TTTCTCGGAAAGAGCTTT
D.sec_U11  TCTACAATTCCCAACTGCTTCTAGCAGCTCTCGCAACAACTCCCACTATA-TTTCTCGGAAAGAGCTTT
D.mel_U11  TCTACAATTCCCAACTGGTTCTAGGAGCTTTCGCAACAACTCCCAATATA-TTTCAGGAGAAGAGCTTT
D.yak_U11  TCTATAATTCGCAACTGGTTCTAGCAGCTCTCGCAACAACCCCCACTATA-TGTGGAATGGAGAGCTTT
D.tak_U11  CCAATAATTCCCAACTGGTTTTGGCGGCTTTCGCCATGACTCGCGCTAAACTTCCCAG-CGAA-ACAAT
D.bia_U11  CCAACAATTCCCAACTGGTTCTGGCAGCTGCTGCCATGACTC-CCCCAAGGTCCCCAATCGCG-GCTT-
D.rho_U11  TGTGCAATTCCCAACTGGTTTTAGCAGCTTTCGGCATGACTCGCCCATCAACCACCGG-CGAA-GCTTT
D.ele_U11  CGTGTAATTCCCAACTGGTTTTAGCAGCTTTCTGCATGACACCCGCA-CATGCTCCGA-CAAA-GCTTT
D.ana_U11  TGTAGAATTCCCAACCGGTTTTAGCTTTGCTTGCCACGACTTTGTCCACAAACTT-GG-CAAATGTGTT
D.kik_U11  CATATAATTCCCAACCACTTCAGGCAGAACTGGCCATGACTTCGGCTGTGGCGGCCGCATAAA-ATTT-
D.sec_LU   CAAGTAATTCCCAACTGGTTATGCC-AATACAACGTAGA-ATCCCCATCGAGTGCCGG-GGATCATTCT
D.sim_LU   CAAGTAATTCCCAACTGGTTATGCC-AATACAACGTAGA-GTCCCCATCGAGTGTAGG-GGATCATTTT
D.mel_LU   CAAGTAATTCTCAACTGGTTATGGC-AAGCCAACGTAGA-ATCCCCATCGAGTGTCGG-GGATCATC--
D.yak_LU   CAAGTAATTCCCAACTGGTTATGGC-AAGTCCCAGTAGA-ACCCCCATCGACGACTGT-GGATCAGC--
D.tak_LU   GATATAATTCCCAATTGGTTATGGCAAACTTTTCACAGACAGTCCC-TTGATTCCCGG-GGTTGATT--
D.bia_LU   GAAATAATTCTCAATTGGTTATGGCTGACTCTGCGTAGGGATTCCC-TCGATCGTCGG-GGAAGATA--
D.rho_LU   GAAATAATTCCCAACTGGTTATAACAAACTTTGAGTAGAGAGTCCCATGGATTGCCGA-CAACGATT--
D.ele_LU   GAAGAAATTCCCAACTGGTTTTAGCAAACCTTCAGTATAGAGTTCC-TGGATGGCCGG-GATCCATT--
D.ana_LU   TTCATAATTCCCAATCGGTTCTCGTCACATTCATGTAGACAGGCCCCTCGGTT-TCAA-CGTCAAGT--
D.kik_LU   TTTATAATTCCCAACTGGCTTTAGTAAACTAAATTATAACGTG-ATTTTAGTTTCAATTTAA-------
```

```
           ..............................(((((.((.((......)).)))))))...
U5:38ABa   UGAGUCUAAACUCAAUUUUUGUAU---GACCUG-GCUAA--AUAUUUAGUUGGGCC-A
U5:35D     UGAGUCUAAAAUAUUAUUUUGUAG---UGCCCG-GCGAC--UUUGGUAACUGGGCC-A
U5:34A     UGAGUCUAAAAUAAUCUUUUGUAG---UGCCCG-GCGAC--UUCGGUAGCUGGGCC-A
U5:38ABb   UGAGUCUAAACACAAU-UUUUAUU-GAGGCCUG-AUAAC--UUAUGCUAUCGGGCC-A
U5:14B     UGAGUCUAAACACAAUUUUUUAUU-GAGGCCUGGAUAAC--UUAUGUUAUCGGGCCCA
U5:23D     UGAGUCUAAACACAAUUUUUGCUUAGAGCCCCG-AUGGC--AUUUGCCUUUGGGGCCA
U5:63BC    UGAGUCUAAAAUAAUUUUUAGUAG--UGCCCU--GUCGC--AA----GACUGGGGCCA
D.yak_LU   CAUUAUCCGAGGCAAUUUUUUUAG---UGCCUA-GCCGCAAAAAUGCCGCCGGGCCGU
D.tak_LU   CAUUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGCG-UUCUGCUGCCGGGCCGU
D.rho_LU   ACAUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGCG-UUCUGCCGCCGGGCCGU
D.sec_LU   AGUUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGCG-AAAUGCUGCCGGGCCGU
D.sim_LU   AGUUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGCG-AAAUGCUGCCGGGCCGU
D.mel_LU   CAUUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGCG-AAAUGCCGCCGGGCCGU
D.ele_LU   ACAUAUCCGAGGCAAUUUUUUUAG---UGCCUG-GCCGC--AUUUGCCGCCGGGCCGU
D.bia_LU   CAUUUUUCGAGGCAAUUUUUGUAG---UGCCUC-GCCGCG-UUCUGCUGCCUGGCCGU
D.ana_LU   GUAUAUCCGAGGCAAUUUUUUGAG---UACCCG-GC--------UGCAACUGGGUCGU
           .........10........20........30........40........50.......
```

```
D.mel_LU     ATGTCTCGATCGCCGCTTCAGTTGTGGAGCGAGAGCTTACGCAATGGAGCGGAGTGATGA 60
D.mel_LU-p   -----ACGATCGCCGCTTCAGTTGTGGAGCGATAGCTTACGCAATGGAGCGGAGTGATGA 55
                  ************************** ***************************

D.mel_LU     GCACATTATCCGAGGCAATTTTTTTAGTGCCTGGCCGCGAAATGCCGCCGGGCCGT 116
D.mel_LU-p   GGATATTATCCGAGGCAAT------------------------------------- 74
             * * ***************
```

Figure 2.10: **Additional characterization of LU snRNA.** Alignment of U11 and LU snRNA promoters in 10 *Drosophila* species that have the LU gene. Alignment by ClustalW2, coloring by the Color Align Conservation app in Sequence Manipulation Suite. Highlighted nucleotides are >70% identical in all sequences. PSEA element is highly conserved, but not the PSEB element. Species name abbreviation: D.mel: *Drosophila melanogaster*; D.yak: *D. yakuba*; D.sec: *D. sechellia*; D.sim: *D. simulans*; D.tak: *D. takahashii*; D.rho: *D. rhopaloa*; D.ele: *D. elegans*; D.bia: *D. biarmipes*; D.ana: *D. ananassae*; D.kik: *D. kikkawai*. b Sequence structure alignment of the 3' end of *D. melanogaster* U5 paralogs and LU snRNA orthologs using LocARNA (global standard alignment). Note the conservation of the 3' end stem loop. Blue box: Sm site. c. Sequence alignment of the LU with its pseudogene paralog residing in an intron of *Ac3* (chr2L:21644292-21644365) in *D. melanogaster*. d Putative base pairing between LU and U6. Only the 5' end of U6 and the interal single stranded region in LU are shown. e Expression of LU and fas2 RNA is reduced to less than 1% in the P element insertion line 10580, as determined by quantitative RT-PCR. CAG: *CyO actin::GFP* balancer.

et al., 2003). To determine whether other ncRNAs co-purify with Sm proteins in Drosophila and human cells, we systematically analyzed the enrichment values of snoRNAs and scaRNAs in our RIP-seq datasets. Consistent with the findings of Fu and Collins (Fu and Collins, 2006), we found that two previously identified Drosophila scaRNAs, U85 (CR32863 or snoRNA:MeU5-C46) and CR33716 (snoRNA:MeU5:U42), were enriched in the Sm protein IPs (Figure 2.8a; Table 2.5). Interestingly, the new Sm-associated ncRNA identified in this study (CR43600 or snoRNA:Prp8) also appears to have features of box H/ACA scaRNAs. Indeed, evolutionary comparisons identify conserved H/ACA and CAB box elements present within the detected orthologs (Figure 2.12b,c). snoRNA:Prp8 folds into a predicted secondary structure similar to that of other box H/ACA scaRNAs, which is further supported by the presence of multiple covariant base pairs. In support of the notion that snoRNA:Prp8 is an H/ACA box scaRNA, we searched snRNAs for sequence complementarity to the pseudouridylation pocket sequences, and found potential target sites in U1, U5, U7 and U11 (Figure 2.12d). Therefore, we have renamed this transcript scaRNA:Prp8. We detected homologs of scaRNA: Prp8 in both Diptera (Drosophilids, Anopheles gambiae) and Hymenoptera (Apis mellifera), but not in Coleoptera (Tribolium castaneum) (Figure 2.12b). The orthologous scaRNA:Prp8 RNAs are highly conserved, suggesting their functional importance. Many scaRNA and snoRNA genes reside within introns of splicing and translation-related genes, respectively (Lestrade and Weber, 2006). The nested gene structures are thought to facilitate transcriptional co-regulation. Thus, it is not surprising that the Prp8 host gene encodes a splicing factor (Figure 2.12a) (Lossky *et al.*, 1987; Pinto and Steitz, 1989). Although Fu and Collins (Fu and Collins, 2006) reported that only SmB and SmD3 co-purified with scaRNAs such as hTR, we found that IP targeting VFP-SmD1 also pulled down snoRNA:Prp8 (Figure 2.17a). It has been shown that many H/ACA box scaRNAs are TMG-capped (Jady *et al.*, 2004; Seto *et al.*, 1999; Tang *et al.*, 2012; Simoes-Barbosa *et al.*, 2012); consistent with these studies, we also found that scaRNA: Prp8 co-immunoprecipitates with anti-TMG antibodies (Figure 2.16a).

To identify additional Sm-associated ncRNAs in HeLa cells, we examined known human sno/scaRNA loci. Several of the previously reported scaRNAs, including U85, U87 and U89, showed moderate but significant enrichment in Y12 IPs (Figure 2.8b; Table 2.7). In addition,

```
U1_21D     ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_95Ca    ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_95Cb    ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_95Cc    ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_82Eb    ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGTACCTCGGCTGAGTTGACCTCTGCGATTATT 100
           ********************************************************************* ******************************
U1_21D     CCTAATGTGAATAACTCGTGCGTGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_95Ca    CCTAATGTGAATAACTCGTGCGTGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_95Cb    CCTAATGTGAATAACTCGTGCGTGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_95Cc    CCTAATGTGAATAACTCGTGCGCGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_82Eb    CCTAATGTGAATAACTCGTGCGTGTAATTTTTGTTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
           ********************** ********** ******************************


U2_38Aba   ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_38ABb   ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCT-AACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 99
U2_14B     ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_34ABc   ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_34ABb   ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_34ABa   ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
           ******************************************************* ********************************************
U2_38ABa   GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_38ABb   GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 191
U2_14B     GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGACTTCGGCCCAAC 192
U2_34ABc   GATTTTTGGAATCAGACGGAGTGCTAGGAGCTTGCTCCACCTCTGTCGCGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_34ABb   GATTTTTGGAATCAGACGGAGTGCTAGGAGCTTGCTCCACCTCTGTCGCGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_34ABa   GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCGCGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
           ****************************.******************.******************************** ***********


U4_38AB    ATCTTTGCGCAGAGGCGATATCGTAACCAATGAAG-TTCTACTGAGGTGCGATTATTGCTAGTTGAAAACTTTAACCAATACCCCGCCATGGGGACGTGA 99
U4_39B     ATCTTTGCGCAGTGGCAATACCGTAACCAATGAAG-TCCTCCTGAGGTGCGGTTATTGCTAGTTGAAAACTTTAACCAATACCCCGCCATGGGGACGTGA 99
U4_25F     AACCTTGTGCAGTGGCAACATCGCAAGCAATGAAGTTCCAACTGAGCTGCGATTATTGCTAGTTGAAAACTAAAACCAATATCTCGCCCAGCGTAAG-GA 99
           *:* *** ****:***.* * ** ** ******** * *:.***** ****.******************::******** * ****.:* * *.* **
U4_38AB    AATACCGTC----CACTACGGCAATTTTTGGAAG-CCCGAGAGGGCCA- 142
U4_39B     AATACCGTC----CACTACGGCAATTTTTGGAAG-CCCGAGAGGGCTAA 143
U4_25F     TCTACGATCTTTAAGCTAAGGCAATTTTTTTAGGCCCCAAGTGGGCTGA 148
           :.*** .**    ..***.**********  *.* ***.**:**** .


U5_23D     ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTTGCTTA 100
U5_38ABb   ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTT--ATT 98
U5_38ABa   ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACTCAATTTTTG---T 97
U5_34A     ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAAATAATCTTTTG---T 97
U5_35D     ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAAATATTATTTTG---T 97
U5_14B     ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTTT-ATT 99
U5_63BC    ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAA-ATAATTTTTA-GTA 98
           ************************************************************************************* : :: ****    :
U5_34A     AGTG-CCCGGCGACTTCGGTAGC-----TGGG-CCA- 129
U5_35D     AGTG-CCCGGCGACTTTGGTAAC-----TGGG-CCA- 127
U5_63BC    -GTG-CCCTGTCGC----AAGAC-----TGGGGCCA- 122
U5_38ABa   -ATGACCTGGCTAAATATTTAGT-----TGGG-CCA- 126
U5_38ABb   -GAGGCCTGATAACTT--ATG-CT---ATCGGGCCA- 126
U5_14B     -GAGGCCTGATAACTT--ATG-TT---ATCGGGCCCA 129
U5_23D     -GAGCCCCGATGGCAT--TTGCCT---TTGGGGCCA- 128
              * **                     * ** **
```

Figure 2.11: **Sequence alignment of *D. melanogaster* U1, U2, U4 and U5 paralogs.** The paralogs of U1 and U2 have very few nucleotide variations and they are highlighted. U4 and U5 paralogs have significant differences among them.

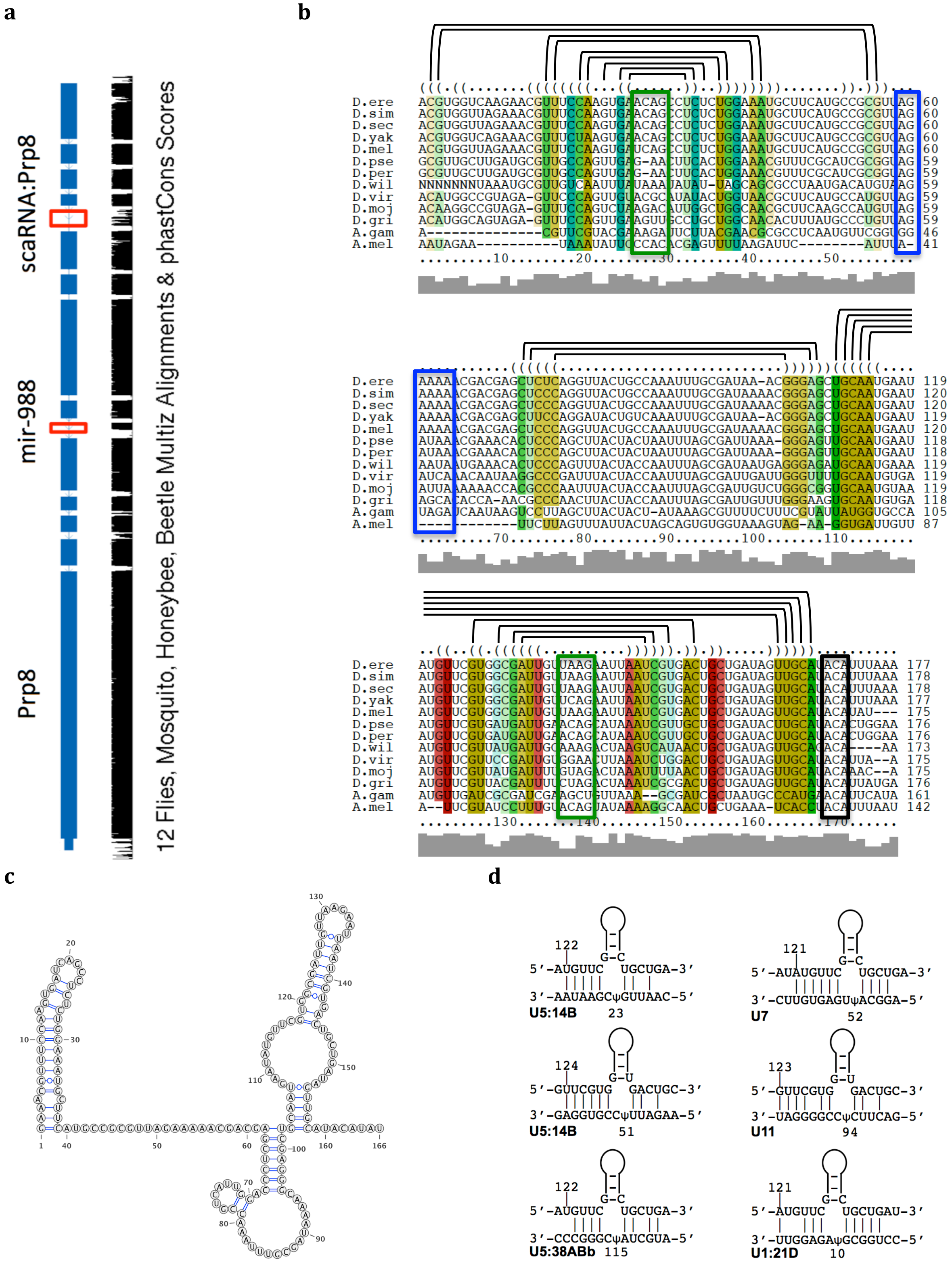
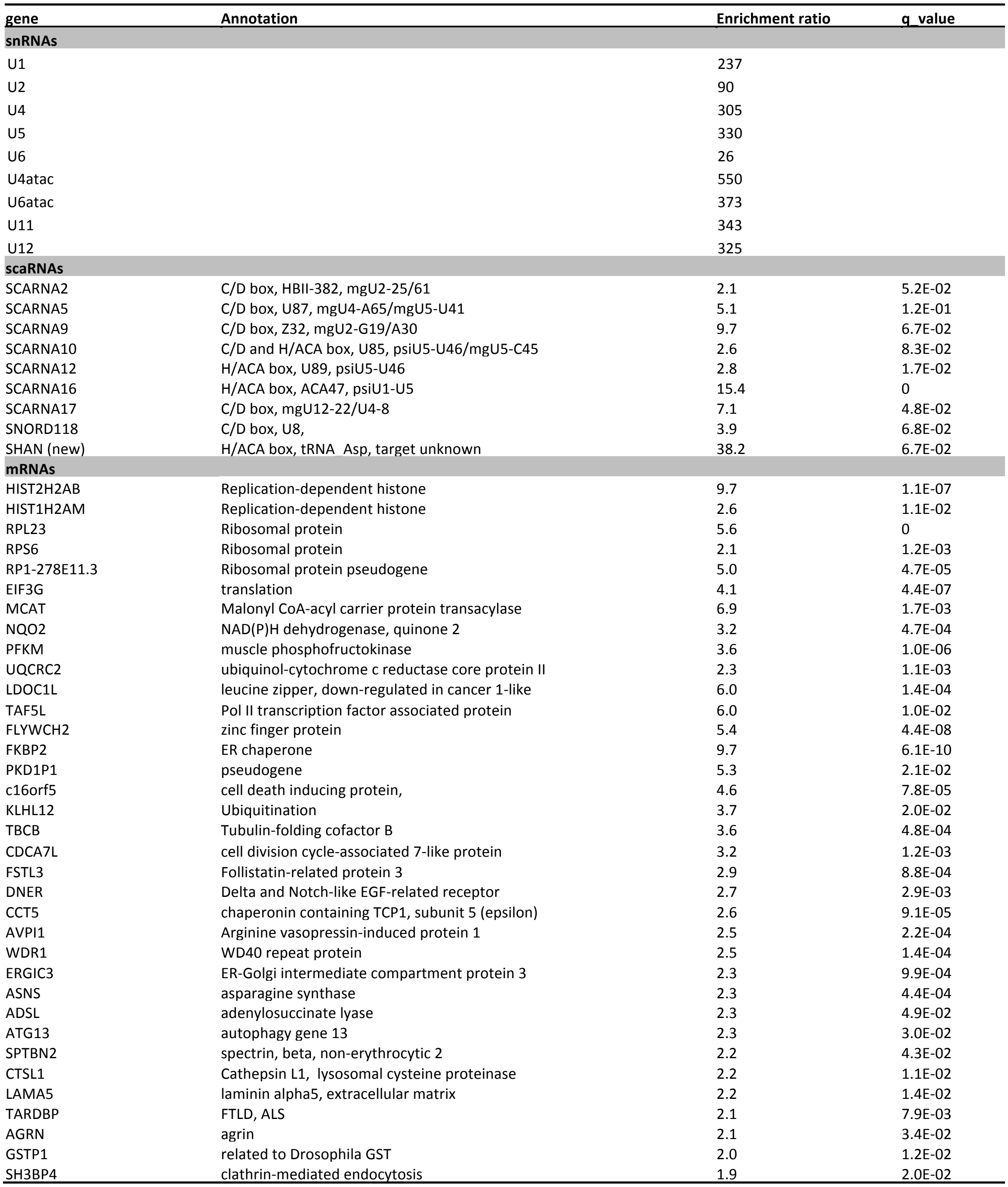


Figure 2.12: **Characterization of scaRNA:Prp8.** (See figure legend on the next page.)



gene	Annotation	Enrichment ratio	q_value
snRNAs			
U1		237	
U2		90	
U4		305	
U5		330	
U6		26	
U4atac		550	
U6atac		373	
U11		343	
U12		325	
scaRNAs			
SCARNA2	C/D box, HBII-382, mgU2-25/61	2.1	5.2E-02
SCARNA5	C/D box, U87, mgU4-A65/mgU5-U41	5.1	1.2E-01
SCARNA9	C/D box, Z32, mgU2-G19/A30	9.7	6.7E-02
SCARNA10	C/D and H/ACA box, U85, psiU5-U46/mgU5-C45	2.6	8.3E-02
SCARNA12	H/ACA box, U89, psiU5-U46	2.8	1.7E-02
SCARNA16	H/ACA box, ACA47, psiU1-U5	15.4	0
SCARNA17	C/D box, mgU12-22/U4-8	7.1	4.8E-02
SNORD118	C/D box, U8,	3.9	6.8E-02
SHAN (new)	H/ACA box, tRNA Asp, target unknown	38.2	6.7E-02
mRNAs			
HIST2H2AB	Replication-dependent histone	9.7	1.1E-07
HIST1H2AM	Replication-dependent histone	2.6	1.1E-02
RPL23	Ribosomal protein	5.6	0
RPS6	Ribosomal protein	2.1	1.2E-03
RP1-278E11.3	Ribosomal protein pseudogene	5.0	4.7E-05
EIF3G	translation	4.1	4.4E-07
MCAT	Malonyl CoA-acyl carrier protein transacylase	6.9	1.7E-03
NQO2	NAD(P)H dehydrogenase, quinone 2	3.2	4.7E-04
PFKM	muscle phosphofructokinase	3.6	1.0E-06
UQCRC2	ubiquinol-cytochrome c reductase core protein II	2.3	1.1E-03
LDOC1L	leucine zipper, down-regulated in cancer 1-like	6.0	1.4E-04
TAF5L	Pol II transcription factor associated protein	6.0	1.0E-02
FLYWCH2	zinc finger protein	5.4	4.4E-08
FKBP2	ER chaperone	9.7	6.1E-10
PKD1P1	pseudogene	5.3	2.1E-02
c16orf5	cell death inducing protein,	4.6	7.8E-05
KLHL12	Ubiquitination	3.7	2.0E-02
TBCB	Tubulin-folding cofactor B	3.6	4.8E-04
CDCA7L	cell division cycle-associated 7-like protein	3.2	1.2E-03
FSTL3	Follistatin-related protein 3	2.9	8.8E-04
DNER	Delta and Notch-like EGF-related receptor	2.7	2.9E-03
CCT5	chaperonin containing TCP1, subunit 5 (epsilon)	2.6	9.1E-05
AVPI1	Arginine vasopressin-induced protein 1	2.5	2.2E-04
WDR1	WD40 repeat protein	2.5	1.4E-04
ERGIC3	ER-Golgi intermediate compartment protein 3	2.3	9.9E-04
ASNS	asparagine synthase	2.3	4.4E-04
ADSL	adenylosuccinate lyase	2.3	4.9E-02
ATG13	autophagy gene 13	2.3	3.0E-02
SPTBN2	spectrin, beta, non-erythrocytic 2	2.2	4.3E-02
CTSL1	Cathepsin L1, lysosomal cysteine proteinase	2.2	1.1E-02
LAMA5	laminin alpha5, extracellular matrix	2.2	1.4E-02
TARDBP	FTLD, ALS	2.1	7.9E-03
AGRN	agrin	2.1	3.4E-02
GSTP1	related to Drosophila GST	2.0	1.2E-02
SH3BP4	clathrin-mediated endocytosis	1.9	2.0E-02

Table 2.7: **Enrichment ratios of human Sm-associated RNAs.**Note the enrichment ratios for the snRNAs are from the medians of mappable human snRNAs.

we found several other scaRNAs that are highly enriched (Figure 2.8b; Table 2.7). However, we did not detect any significant enrichment of hTR as previously reported (Fu and Collins,

Figure 2.12: **Characterization of scaRNA:Prp8.** a. The scaRNA:prp8 is located in the conserved 4th intron of Prp8 gene, and a previously identified microRNA mir-988 is located in the conserved 8th intron. b. Sequence structure alignment of scaRNA:Prp8 in 13 selected insect species by LocARNA (global standard alignment). Blue box: H box; black box: ACA box; green boxes: CAB boxes; black half brackets: covariant basepairs. Species name abbreviation: D.ere: *D. erecta*; D.pse: *D. pseudoobscura*; D.per: *D.persimilis*; D.wil: *D. willistoni*; D.vir: *D. virilis*; D.moj: *D. mojavensis*; D.gri: *D. grimshawi*; A.gam: *Anopheles gambiae*; A.mel: *Apis mellifera.* c. Predicted secondary structure of the scaRNA:Prp8. d. Potential basepairing between the putative peudouridylation guide sequence and target snRNAs.

2006) (data not shown). We identified a novel, unannotated Sm-associated ncRNA, which we named SHAN (Sm-associated Hybrid tRNAAspcontaining NcRNA); its predicted secondary structure is shown in Figure 2.13c. This new transcript appears to be a chimera between a tRNA gene and an H/ACA type scaRNA gene. Supporting this hypothesis, we detected H box, ACA box and CAB box motifs in the orthologous sequences from other primates (Figure 2.13b,c). In summary, our RIP-seq analysis revealed both evolutionarily conserved and newly evolved interactions between Sm proteins and scaRNAs, suggesting that Sm proteins play roles in the biogenesis/function of a subset of scaRNAs. However, we did not identify sequence/ structural features that distinguish Sm-associated scaRNAs from other scaRNAs.

2.3.7 Sm-associated mRNAs encode mitochondria/translation-related proteins

Due to a relative lack of comprehensive annotation of Drosophila gene ontology, we manually annotated the Sm-associated mRNAs by homolog searching, protein domain analysis, and literature mining. This analysis surprisingly revealed two major categories of mRNAs: those encoding ribosome/translation-related proteins (13/86), and mitochondrial proteins (including mitochondrial ribosomal proteins, 19/86). As discussed above, the enrichment of ribosomal protein mRNAs is not simply due to high levels of expression. Only a subset of ribosomal protein mRNAs is enriched in the Sm protein IPs. For example, mRNAs encoding RpS11 (CG8857) and RpL39 (CG3997) are highly enriched in Sm protein IPs (Figure 2.7a; Table 2.5), whereas RpL19 (CG2746) and RpL4 (CG5502) are not enriched at all (Figure 2.8a and data not shown). Anecdotally, the mRNA encoded by CG3776, which is highly enriched, is located immediately adjacent to RpL19 in the Drosophila genome, demonstrating the high degree of specificity of

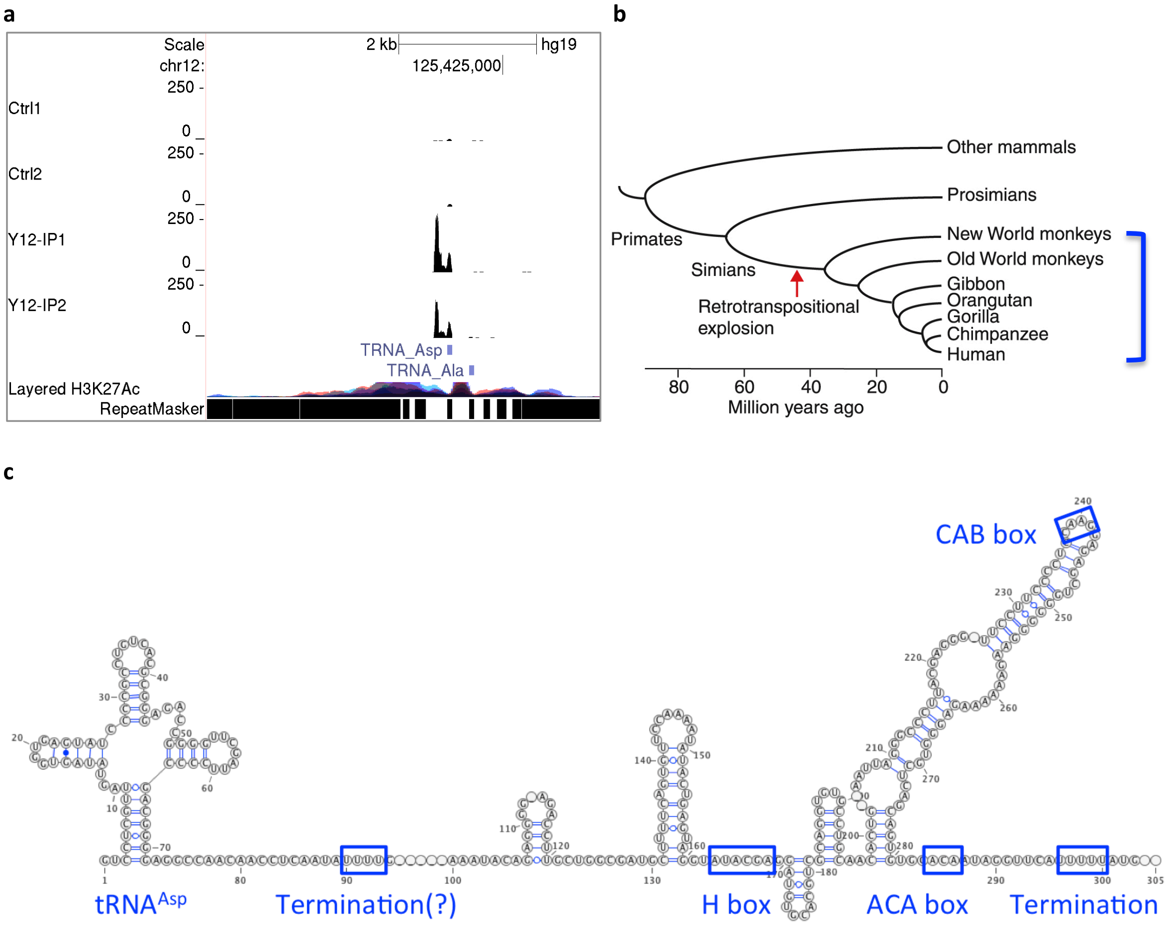


Figure 2.13: **Enrichment, structure, and phylogeny of SHAN scaRNA.** (see legend on the next page)

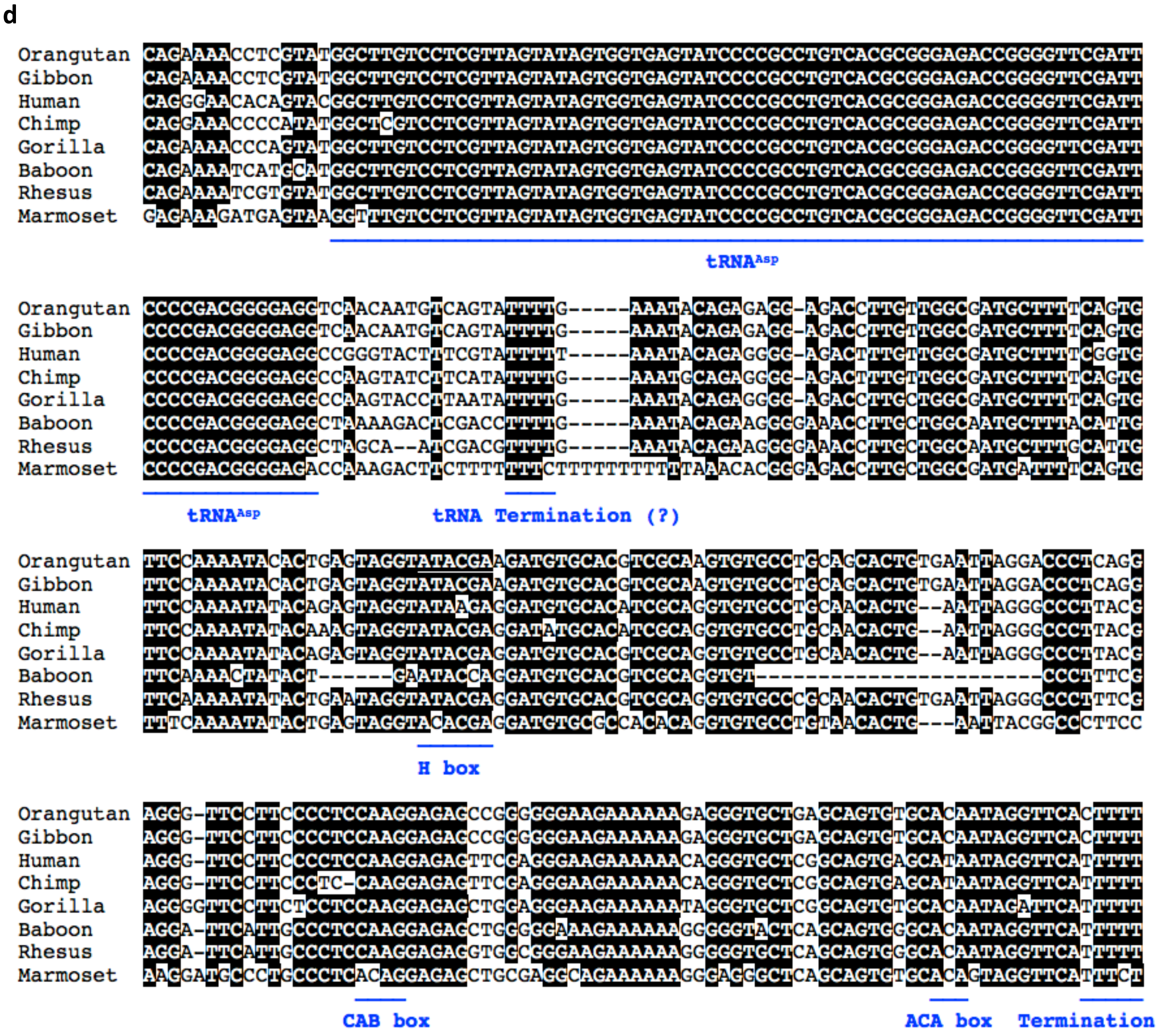


Figure 2.13: **Enrichment, structure, and phylogeny of SHAN scaRNA.** a: Genome browser view of the SHAN locus, showing two Ctrls and two IPs. TRNA_Asp is part of the new gene SHAN, while TRNA_Ala is a separate tRNA gene. b: consensus secondary structure of SHAN orthologs. tRNAAsp: the tRNA part of this new gene SHAN; tRNA Termination (?): the presumed pol III transcription termination signal for tRNA; H/ACA/CAB boxes: putative scaRNAs sequence elements. Termination: putative pol III transcription termination signal for this new gene SHAN. c: SHAN scaRNA evolved from the root of simians (blue bracket), after the retrotranspositional explosion (reproduced from Ohshima et al., 2003 Genome Biology, published by Biomed Central). d: Alignment of SHAN scaRNAs orthologs from eight simian species. Human: *Homo sapiens*; Chimp: *Pan troglodytes*; Gorilla: *Gorilla gorilla*; Orangutan: *Pongo abelii*; Gibbon: *Nomascus leucogenys*; Rhesus: *Macaca mulatta*: macaque (Old World monkey); Baboon: *Papio anubis* (Old World monkey); Marmoset: *Callithrix jacchus* (New World monkey).

our approach.

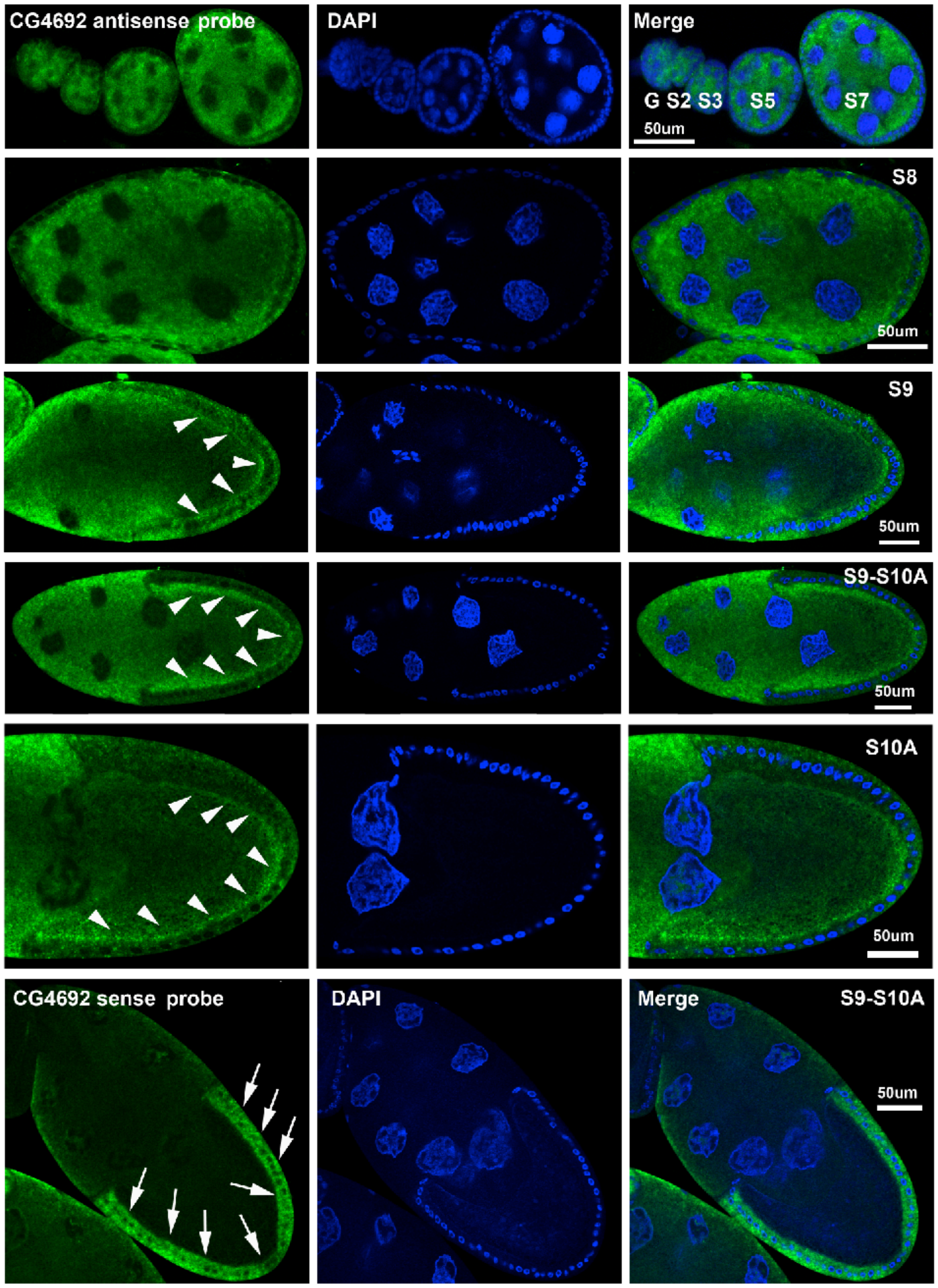


Figure 2.14: **CG4692 mRNA localizes along oocyte cortex.** CG4692 mRNA (detected with CG4692 antisense probe) was enriched in the oocyte cortex from stages 9 to 10A (arrow heads, S9-S10A), but not in earlier stages (from germarium to stage 8). The CG4692 sense probe shows a localization pattern opposite to that of CG4692 mRNA (in follicle cells, arrows) and does not label the cortex. This latter pattern is likely due to the existence of a putative antisense transcript from the CG4692 locus (see EST data in Flybase.org).

Two other *Drosophila* Sm-associated mRNAs merit special interest. CG4692 encodes a predicted mitochondrial F1-FO ATP synthase subunit that was consistently enriched in our IPs. We found that this mRNA localizes to the actin-rich oocyte cortex of late-stage Drosophila egg chambers (Figure 2.6), in a pattern that is very similar to that of VFP-tagged Sm proteins, as described previously (Gonsalvez *et al.*, 2010). Analysis of several other high-scoring mRNAs from Figure 2.7a and Figure 2.6 did not display this pattern (data not shown), so it is not a general feature of Sm-associated mRNAs, but was nonetheless interesting. CG1349 (dj-1beta) encodes a Drosophila homolog of the human DJ-1/PARK7 (Parkinson autosomal recessive, early onset 7) gene. DJ-1/ PARK7 is one of 10 genes identified to date that cause familial Parkinson disease (Houlden and Singleton, 2012). A subpopulation of DJ-1 protein is localized to mitochondria in a regulated manner, and is required for proper mitochondrial function (Zhang *et al.*, 2005). Thus, it is possible that Sm proteins play a role in regulating the localization and/or translation of associated mRNAs.

In contrast to the more than 70 Sm-associated mRNAs in the fruitfly (Figure 2.7a), we identified roughly 30 high-scoring mRNAs in human cells (Figure 2.7b). The lower number in the human dataset is potentially due to a reduced coverage of the transcriptome. Nevertheless, we found that one of the replication-dependent histone mRNAs, HIST2H2AB, is highly enriched in the IPs (Figures 2.7b and 2.8b). In contrast, two adjacent histone genes, HIST2H2BE and HIST2H2AC, were not enriched (Figure 2.8b). Another histone mRNA (HIST1H2AM), was also significantly enriched (Figure 2.7b). Interestingly, Steitz and colleagues (Friend *et al.*, 2007) previously showed that the U2 snRNP binds to (intronless) histone pre-mRNAs and stimulates 3' end processing. Our identification of histone mRNAs in Sm protein co-IPs may reflect a snRNP-mediated interaction between Sm proteins and mRNAs. However, none of the Drosophila replication-dependent histone mRNAs were enriched in the Sm protein IPs (Figure 2.15). Taken together, our data suggest that the mode of interaction between Sm proteins, snRNPs and mRNAs is conserved between vertebrates and invertebrates.

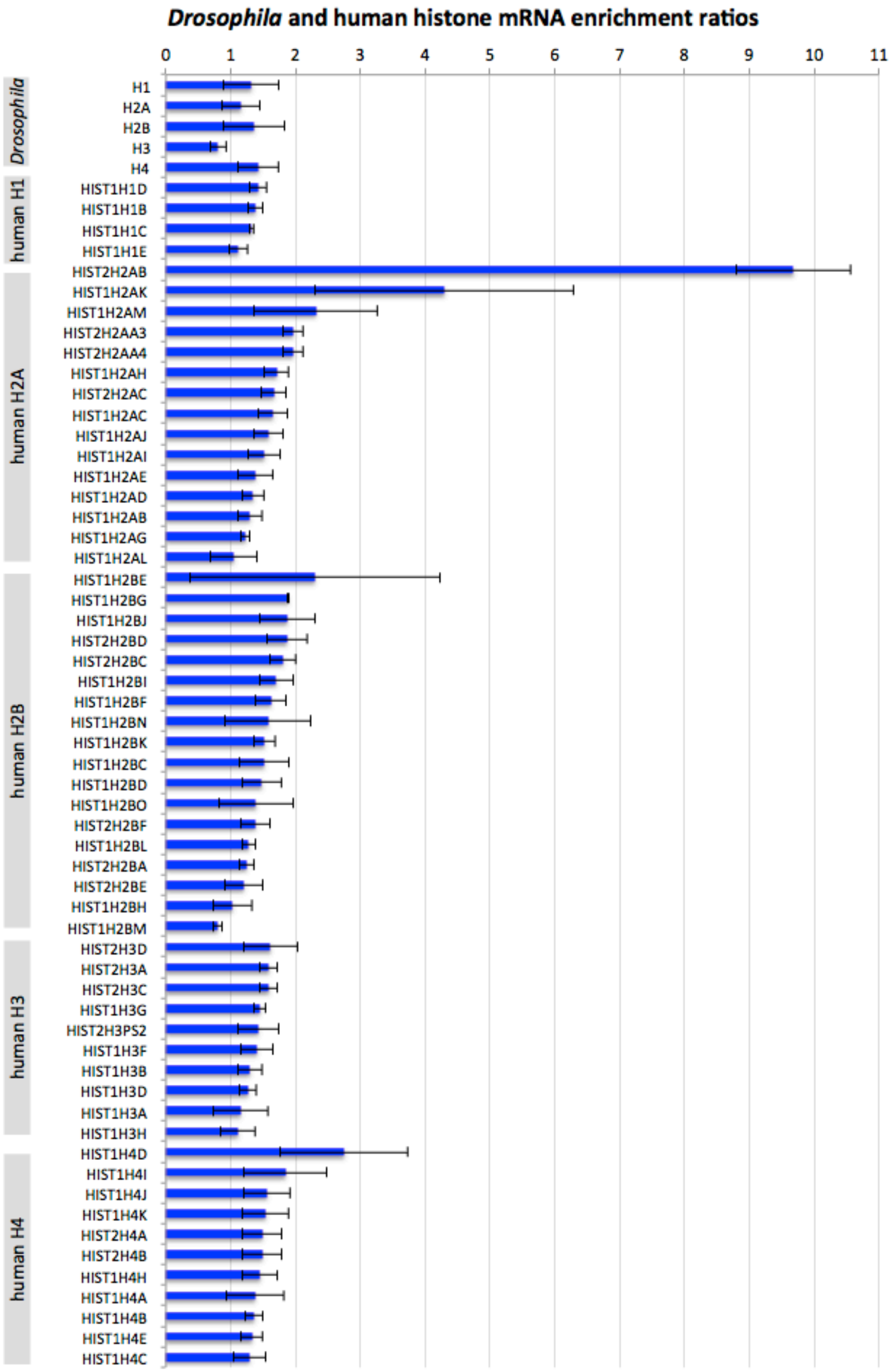


Figure 2.15: **Enrichment ratios for *Drosophila* and human replication-dependent histone mRNAs.** (See figure legend on the next page.)

Figure 2.15: **Enrichment ratios for *Drosophila* and human replication-dependent histone mRNAs.** Five *Drosophila* histone genes were displayed, summarizing all six *Drosophila* RIP-seq experiments. For the 86 human histone genes in the four clusters, 58 of them are detectable and shown in this figure. Error bars for *Drosophila* histone mRNAs represent standard deviation of enrichment ratios of six RIP-seq experiments, while human histone mRNA genes represent standard deviation of IP raw read numbers divided by the average of Ctrl raw read numbers. Note that most histone mRNAs are moderately, even though not significantly, enriched. Exact binomial test of all human histone mRNAs (58) gave a p-value of 4.1E-16, suggesting that human histone mRNAs are associated with Sm proteins (57 mRNAs with enrichment ratios >1), even though most of the enrichment ratios are less than 2-fold and not significant.

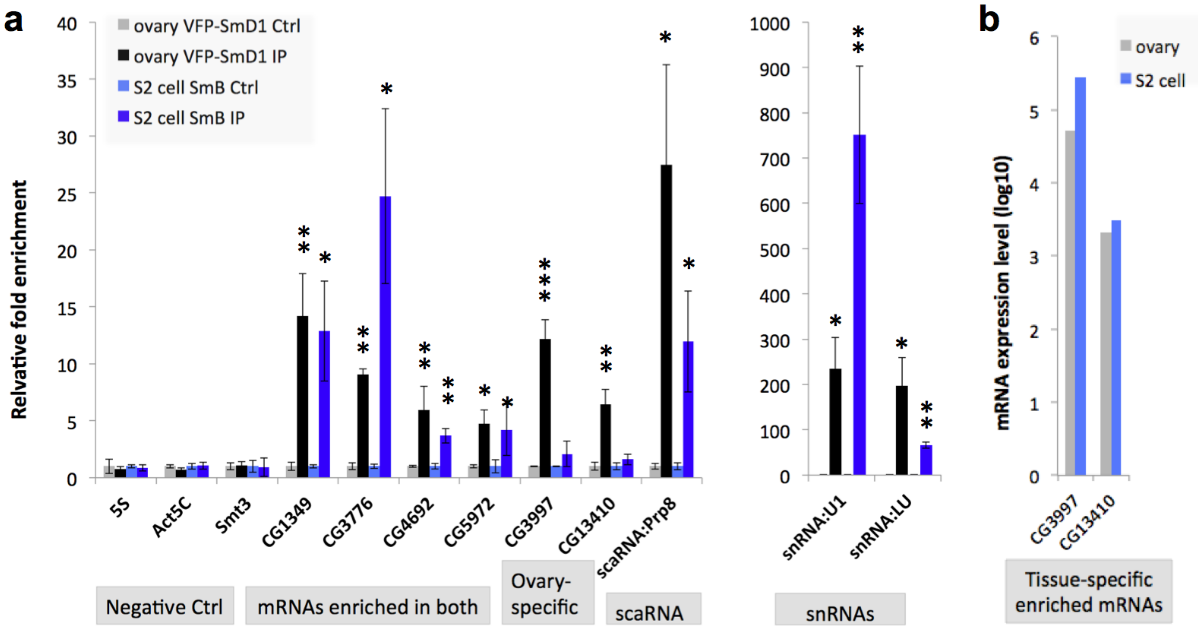


Figure 2.16: **snRNPs associate with mature mRNAs in S2 cells.** (a) Sm-associated mRNAs, as well as scaRNAs and snRNAs, can be pulled down by a TMG antibody in S2 cells. CG9042 (Gapdh) is used for normalization. (b) Enrichment analysis of the U1-70 K RIP-seq data in a volcano plot. The most highly enriched transcripts were labeled. The inset rectangular boxes highlight CG3776 and CG8108 mRNAs in the plot. Note: CG1349 and CG4692 could be associated with other snRNPs, and therefore not pulled down by U1-70 K. (c) CG8108 mRNA can be pulled down by TMG and Y12 antibodies in S2 cells. (d) CG8108 is expressed in similar levels in Drosophila ovary and S2 cells (data from FlyBase). (e) CG8108 mRNA is not enriched in ovary Sm RIP-seq. t-Test for significance between IP and control (Ctrl): *P <0.05, **P <0.01, ***P <0.001). Error bars reflect the standard deviation.

2.3.8 Validation and tissue-specificity of RNA-Sm protein interactions

We have shown that the B/D3 and E/F/G subcomplexes bind essentially the same set of target RNAs. To determine whether SmD1 (which forms heterodimers with SmD2; Figure

2.1b) also associates with the RNAs listed in Figure 2.7a, we immunopurified ovarian RNA from daGal4,VFP-SmD1 flies (using anti-GFP) and carried out qRT-PCR. Furthermore, to assay the observed interactions in another cell type, we also performed qRT-PCR on RNAs immunopurified from S2 cells using anti-Sm antibody Y12. We chose six of the top-ranking mRNAs that were identified in the RIP-seq experiments (targeting SmB, SmD3 and SmE), and found that they were all highly enriched in the VFP-SmD1 IPs (Figure 2.17a). Two snRNAs (U1 and LU) were used as positive controls, whereas three RNAs not expected to interact with Sm proteins (Act5C and Smt3 mRNAs and 5S rRNA) were used as negative controls (Figure 2.17a). In contrast to the results in ovaries, only four out of the six mRNAs we tested were significantly enriched in the S2 cell IPs (Figure 2.17a). Given that the Sm proteins and the six mRNAs we tested all have comparable expression levels in both ovaries and S2 cells (Figure 2.17b and data not shown), these findings suggest that the interactions between mRNAs and Sm proteins can be tissue-specific. A potential concern in all RIP experiments is that the co-purification of the components might be due to reassortment of complexes following cell lysis (Mili and Steitz, 2004; Riley and Steitz, 2013). However, the fact that CG3997 and CG13410 fail to associate with Sm proteins despite the fact that they are well expressed in S2 cells argues strongly against this artifact.

2.3.9 Sm proteins associate with fully spliced and polyadenylated mRNAs

The identification of significantly enriched mRNAs in the co-IP fractions led us to ask whether the association between Sm proteins and mRNAs was due to the splicing reaction itself. In other words, do Sm proteins interact with partially spliced or fully mature mRNAs? A quick glance at Figure 2.7 shows that the read depth over intronic sequences is very low. Meta-gene analysis of both Drosophila and human Sm-associated intron-containing mRNAs showed that the vast majority of reads map to exons, and the IPs did not pull down more pre-mRNAs than the controls did (Figure 2.18a). Among the few transcripts that showed significant numbers of intronic reads, most of those were actually candidates for either new exons or new genes (for example, scaRNA:Prp8 and snRNA:LU; Figure 2.8a). Thus, this analysis demonstrates that the mRNAs that associate with canonical Sm proteins are fully spliced. Importantly, 6 of the

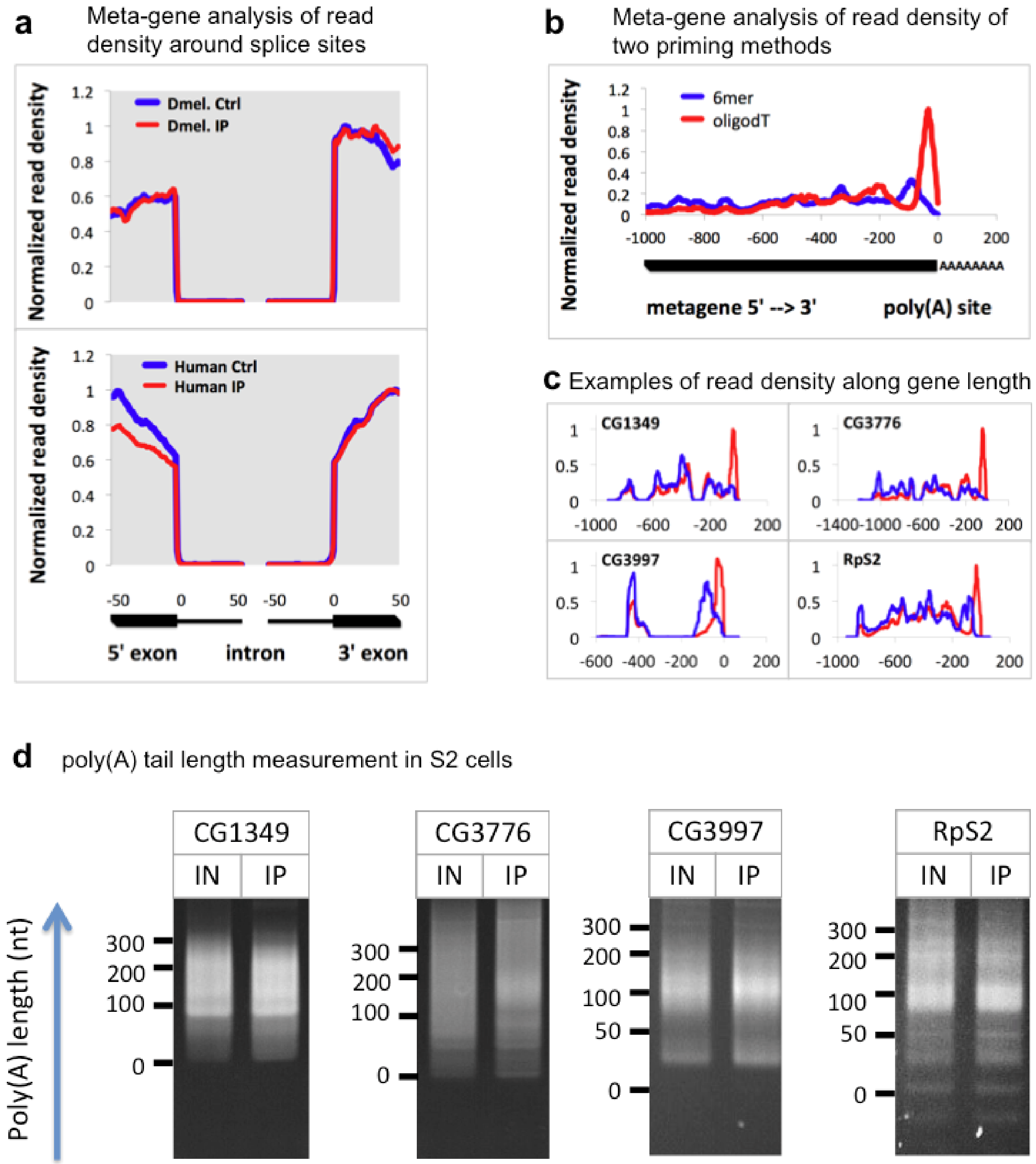


Figure 2.17: **RNA-Sm association is cell type-specific and not due to re-assortment.** (a) RIP-qRT-PCR in da-Gal4 VFP-SmD1 fly ovary (anti-GFP) and S2 cells (Y12). Negative controls (Ctrl) used are 5S rRNA, Act5C and Smt3. CG9042 (Gapdh) is used as the normalization standard. snRNAs are shown separately due to the difference in scale. (b) mRNAs associated with Sm proteins in ovaries but not in S2 cells are expressed in S2 cells. t-Test for significance between IP and Ctrl: *P <0.05, **P <0.01, ***P <0.001. Error bars show standard deviation.

72 Drosophila Sm-associated mRNAs (CG6008, CG13151, CG13951, CG17531, CG11076 and CG7137), and 2 of the 30 human Sm-associated mRNAs (HIST2H2AB and HIST2H2AM) are intronless, suggesting that splicing is not a prerequisite for Sm protein interaction.

The highly conserved eukaryotic Lsm1-7 complex is known to bind to mRNA degradation intermediates, preferentially those with oligoadenylated tails (Tharun and Parker, 2001; Chowdhury *et al.*, 2007). We therefore asked whether the canonical Sm ring shares this same recognition specificity. Taking advantage of the oligo(dT)20 and random hexamer primed RIP-seq cDNA libraries, we compared the read coverage patterns for the various mRNAs. As shown in Figure 2.18b,c, there is a dramatic 3' end bias in the oligo(dT)20 primed libraries compared to the randomly primed ones. We also confirmed the presence of adenylated tails of Sm-associated and non-associated mRNAs by examining the unmappable reads in the oligo(dT)20 primed RIP-seq files (Figure 2.19). In order to measure polyA tail lengths, we performed RACE-PAT (rapid amplification of cDNA ends-poly(A) tail assay) on immunopurified RNAs from S2 cells (Salles and Strickland, 1995). This analysis demonstrates that the poly(A) tails of the Sm-associated mRNAs are roughly the same length as the input mRNAs (Figure 2.18d). Taken together, these data show that Sm and Lsm proteins have distinct specificities and modes of mRNA interaction.

2.3.10 Sm protein interaction with mRNAs is mediated by snRNPs

The association of snRNAs and scaRNAs with Sm proteins is thought to be mediated by direct binding to Sm sites and CAB boxes, respectively (Fu and Collins, 2006; Leung *et al.*, 2011; Urlaub *et al.*, 2000). We therefore wanted to determine whether Sm proteins associate with mRNAs directly or indirectly. Toward that end, we carried out PAR-CLIP (photoactivatable ribonucleosideenhanced crosslinking and immunoprecipitation) on native and VFP-tagged Sm complexes (Hafner *et al.*, 2010); however, we were unable to detect any significant crosslinking events in the precipitated RNA (data not shown). We note that canonical Sm proteins are notoriously poor at crosslinking. Even on extremely abundant targets such as U1 snRNA, the UV crosslinking efficiency was rather low, with SmG being the predominant crosslinked member of the heptameric ring (Urlaub *et al.*, 2001). More recently, Castello et al. Castello *et al.* (2012)

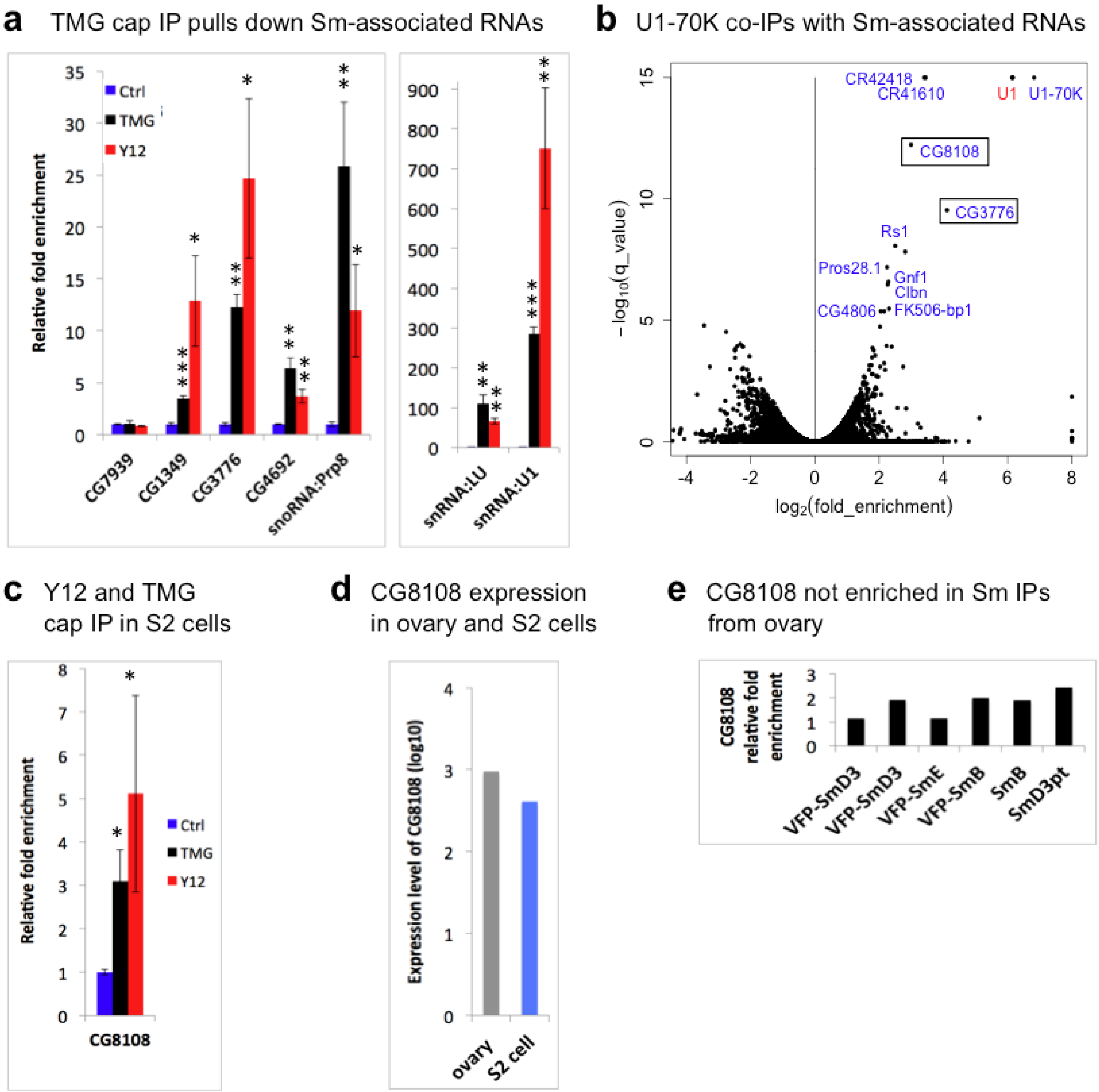


Figure 2.18: **Sm proteins associate with mature mRNAs.** (a) Meta-gene analysis of read density around splice sites for all Drosophila and human Sm-associated intron-containing mRNAs in all RIP-seq experiments. (b) Meta-gene analysis of read density along the gene length for all Drosophila Sm-associated mRNAs quantified from oligodT and random hexamer primed libraries. (c) Example tracks for read density along the gene length for oligodT and random hexamer primed libraries. (d) Poly(A) tail length Sm-associated mRNAs (CG3997, CG1349 and CG3776) and non-associated mRNA (RpS2) from Y12 IP in S2 cells. IN, input total RNA; IP, immunoprecipitated RNA. The labels denote the length of poly(A) tails. Oligo(dT)20 was used as the reverse primer for the reverse transcription and subsequent PCR, therefore producing the 'smear' of poly(A) tail. See Figure 2.19 for analysis of poly(A) containing reads for selected Sm-associated mRNAs.

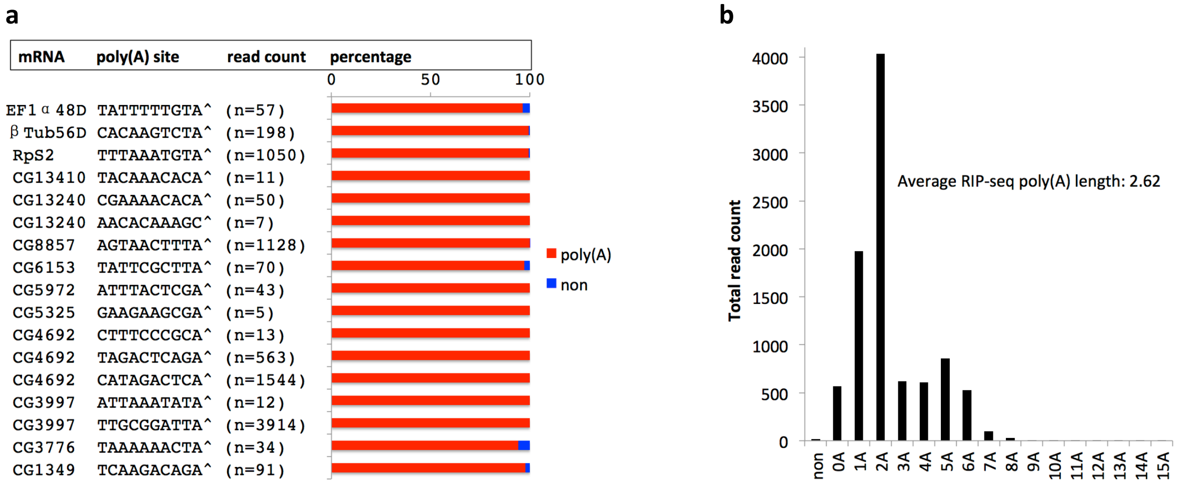


Figure 2.19: **Analysis of the polyadenylation of Sm-associated mRNAs.** a. Percentage of polyadenylated and non-polyadenylated mRNAs that are associated with Sm proteins. Ten Sm-associated mRNAs were analyzed from the RIP-seq data. Some of them have multiple cleavage and polyadenylation sites. EF1alpha48D, betatub56D and RpS2 are used as control, non-Sm-associated mRNAs. b. Distribution of sequenced polyA lengths for the selected Sm-associated mRNAs in the RIP-seq data. Note: the lengths of sequenced polyA tails do not represent the real lengths of polyA tails, because the reads are short (35nt), and the ability of the sequencer to cover long homopolymer stretches is limited.

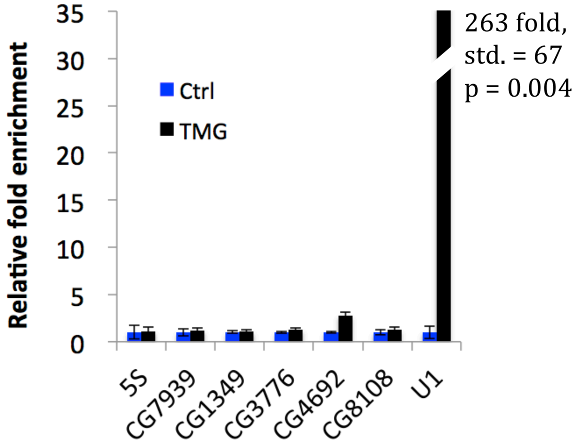


Figure 2.20: **Sm-associated mRNAs are not TMG-capped.** Purified S2 cell total RNA was immunoprecipitated using TMG antibody (K121, unconjugated), and the immunoprecipitated RNA was measured using quantitative RT-PCR. Experiments were performed as quadruplicates and the error bars represent standard deviations. U1 snRNA was used as positive control. Std: standard deviation from four biological replicates.

carried out UV- and PAR-CLIP in parallel to generate a comprehensive mRNA interactome in HeLa cells. As part of their studies, they identified the Lsm1-7 proteins as mRNA binding proteins, but the canonical Sm proteins were not detected, again supporting the idea that Sm proteins are not efficiently crosslinked to mRNAs.

However, the fact that we found all three Sm subcomplexes in association with the same set of mRNAs (Figures 2.3 and 2.7) suggested interaction with a complex that contains an intact Sm ring. Furthermore, the previously reported binding between histone mRNAs and U2 snRNPs (Friend *et al.*, 2007), coupled with our identification of H2A mRNAs in our RIP-seq data (Figure 2.8) led us to ask whether the mRNA-Sm interaction might be indirect, mediated by snRNPs. Sm-class spliceosomal snRNAs are transcribed by a specialized form of RNA polymerase II and contain a 5' TMG cap structure (Matera *et al.*, 2007). Using anti-TMG antibodies, we immunopurified RNPs from S2 cell lysate and used qRT-PCR to assess the enrichment of mRNAs. As expected, the U1 and LU snRNAs (positive controls) were highly enriched in the anti-TMG IPs, whereas CG7939 (RpL32) mRNA was not (Figure 2.16a). Notably, the scaRNA:Prp8 transcript and all three of the Sm-associated mRNAs we tested (CG1349, CG3776 and CG4692) were significantly enriched in the anti-TMG pulldowns (Figure 2.16a). In parallel, we performed anti-TMG IPs using purified S2 cell RNA (that is, the IP was not performed in lysates). We detected significant enrichment of U1 snRNA but not the mRNAs (Figure 2.20). Therefore, the Sm-associated mRNP complex contains a TMG cap component that is structurally distinct from the mRNAs themselves, suggesting the presence of snRNPs.

In order to test whether the interactions with mRNAs are indirectly mediated by snRNPs, we took advantage of a database from a large-scale Drosophila S2 cell RIP-seq analysis of 29 RNA binding proteins, including U1-70 K (modENCODE). The U1-70 K protein binds to U1 snRNA directly and specifically, thus allowing it to be used as an additional, independent epitope for pulldown experiments (Urlaub *et al.*, 2001). We mined the database for RNAs that associate with U1-70 K by analyzing RNAs that were enriched in IPs from U1-70 K transfected versus non-transfected cells. The RIP-seq data were displayed on a volcano plot to identify transcripts that are highly enriched in the IPs. As shown in Figure 2.16b, U1 snRNA, but

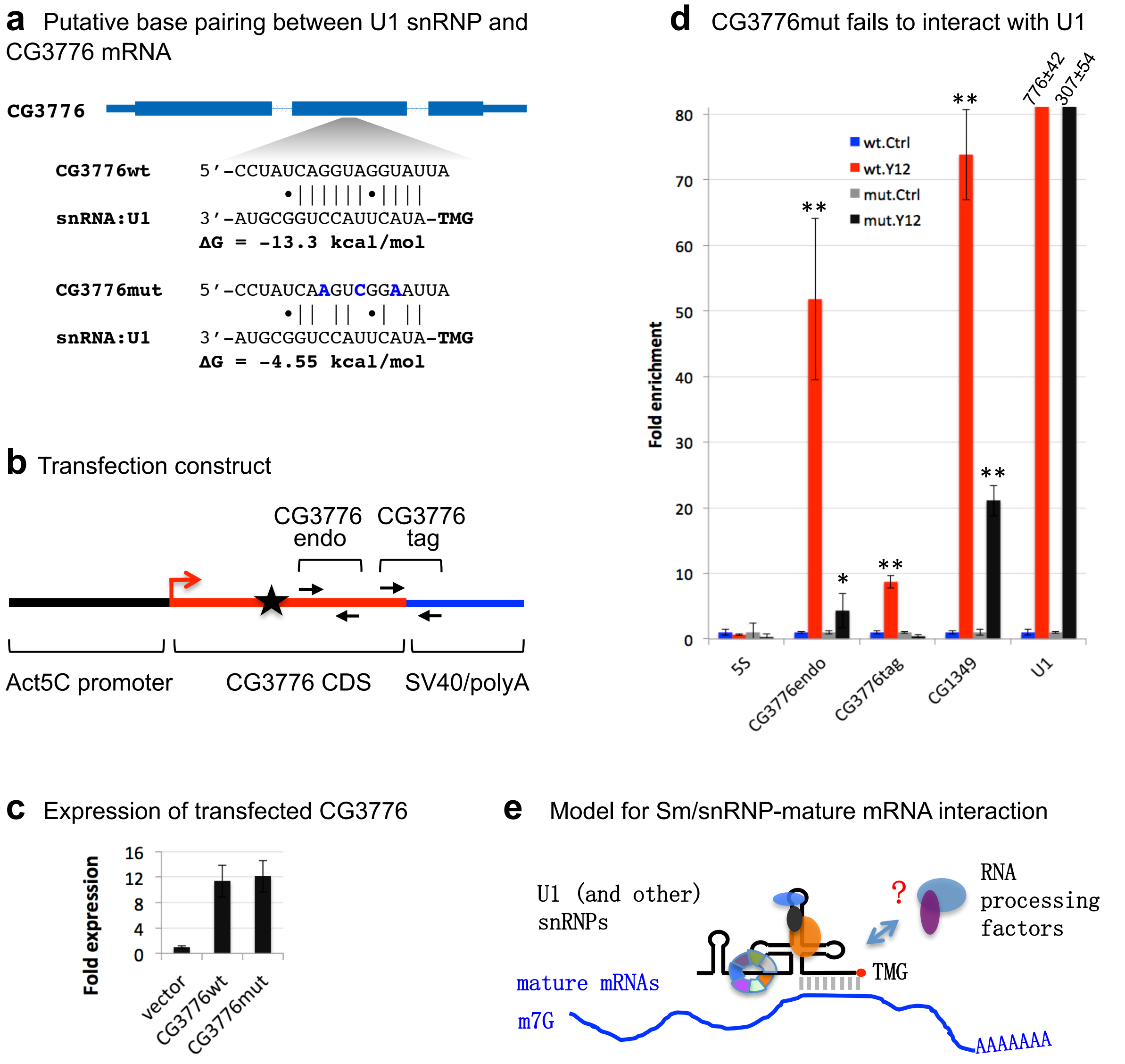
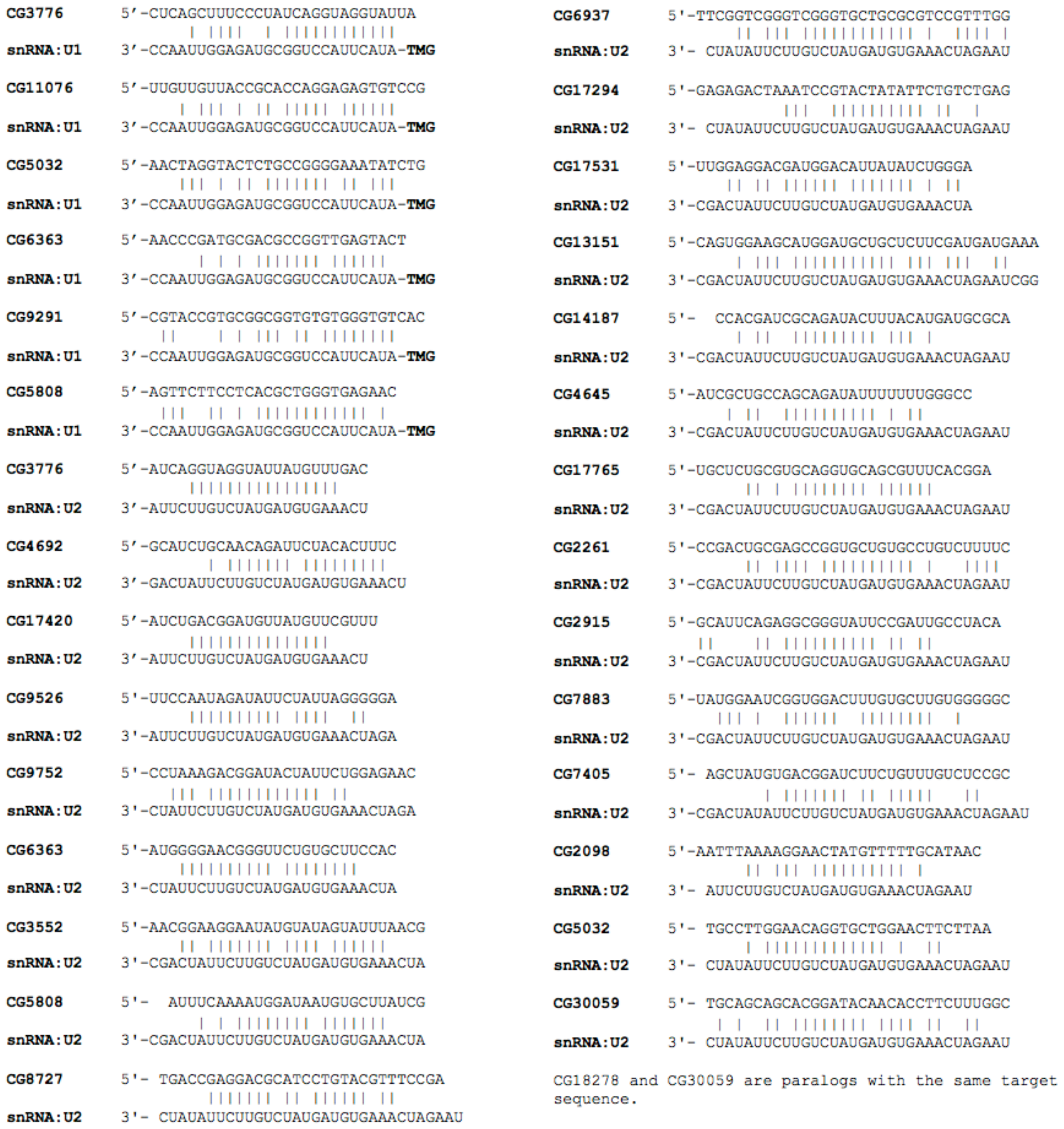


Figure 2.21: **U1 snRNP binds mature mRNAs.** (a) Putative base pairs between the 5' end of U1 snRNA and the CG3776 mRNA coding region (upper panel). Within the putative region of base pairing, three translationally silent point mutations were introduced (bold blue letters) to disrupt the helix (lower panel). (b) Cartoon of the S2 cell transfection construct, showing the CG3776 expression unit. CG3776endo and CG3776tag indicate locations of primers for qRT-PCR. CG3776endo amplifies both endogenous and transfected CG3776 mRNAs, whereas CG3776tag amplifies transfected CG3776 mRNA only. The black star indicates the location of the putative U1 binding site. (c) pAW vector, pAW-CG3776wt and pAW-CG3776mut were transfected into S2 cells, and CG3776wt and CG3776mut expression was measured using qRT-PCR with the CG3776endo primer pair. GAPDH was used as normalization standard. (d) After pAW-CG3776wt and pAW-CG3776mut were transfected, anti-Sm (Y12) IPs were performed using S2 cell lysate. GAPDH was used as normalization standard. (figure legend continued on the next page)

Figure 2.21: (e) Proposed model of snRNP-mRNA interactions. Distinct snRNPs (U1 and potentially others) associate with mature mRNAs via base pairing and/or protein-mediated interaction. Such interactions could serve as a platform to recruit RNA processing factors that act on multiple levels of RNA metabolism. t-Test for significance between IP and control (Ctrl): *P <0.05, **P <0.01, ***P <0.001. Mut, mutant; wt, wild-type.

not the other spliceosomal snRNAs, was dramatically enriched in the IP fractions, along with a number of other ncRNAs and mRNAs. Among this latter category, three mRNAs were particularly noteworthy: CG3776, CG8108 and U1-70 K (CG8749) itself. Although U1-70 K protein may well bind to its own mRNA for some type of autologous feedback, one must view this result with caution because the cells were transiently transfected with U1-70 K cDNAs, artificially inflating expression of this transcript. However, CG3776 and CG8108 remain good candidates. Interestingly, CG3776 was one of the top-ranking candidates in our ovarian RIP-seq experiments (Figures 2.7 and 2.8), but CG8108 was not identified as being enriched, even though it is expressed at similar levels in S2 cells (Figure 2.16d,e). Because the U1-70 K data were generated from S2 cells, we performed anti-TMG and anti-SmB (Y12) IPs in S2 cells, followed by qRT-PCR. As shown in Figure 2.16c, we detected significant enrichment of CG8108 in both the TMG and Sm protein IPs. These data provide additional support for the idea that the Sm-mRNA interactions are cell-type specific and not due to reassortment, as CG8108 is expressed in Drosophila ovaries (Figure 2.16d) but not significantly enriched in Sm protein IPs (Figure 2.16e).

In addition to CG3776, we also found other U1-70 K associated RNAs that overlapped with our Sm protein dataset, including CG5972 and CR32863. Although it is likely that U1-70 K binds to certain RNAs in a manner that is independent of the U1 snRNP, the overlap between our anti-Sm and anti-TMG data suggests that a cadre of mature mRNAs interacts with intact snRNPs outside of the spliceosome. Thus, we checked for sequence complementarity in CG3776 mRNA and found a 12 bp perfect duplex with the 5' end of U1 snRNA (Figure 2.21a). The complementary region is in the middle of the second exon of CG3776, far from any intron-exon boundaries and the base-pairing potential is much greater than is typical for a 5' splice site. Similarly, we found stretches of complementarity between U1 snRNA and exonic regions of



```
CG3776        5’-CUCAGCUUUCCCUAUCAGGUAGGUAUUA
                     | ||||  | ||||||||||||
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG11076       5’-UUGUUGUUACCGCACCAGGAGAGTGTCCG
                    | ||| | || ||||| ||||||
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG5032        5’-AACTAGGTACTCTGCCGGGGAAATATCTG
                    ||| | || ||||||| || |||
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG6363        5’-AACCCGATGCGACGCCGGTTGAGTACT
                      | | | ||||||| ||||||
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG9291        5’-CGTACCGTGCGGCGGTGTGTGGGTGTCAC
                  ||    | | ||| || ||||||||
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG5808        5’-AGTTCTTCCTCACGCTGGGTGAGAAC
                  |||  || | |||||||||||| |
snRNA:U1      3’-CCAAUUGGAGAUGCGGUCCAUUCAUA-TMG

CG3776        5’-AUCAGGUAGGUAUUAUGUUUGAC
                     ||||||||||||||||
snRNA:U2      3’-AUUCUUGUCUAUGAUGUGAAACU

CG4692        5’-GCAUCUGCAACAGAUUCUACACUUUC
                       | ||||||| |||||||||
snRNA:U2      3’-GACUAUUCUUGUCUAUGAUGUGAAACU

CG17420       5’-AUCUGACGGAUGUUAUGUUCGUUU
                     |||||||||||||||
snRNA:U2      3’-AUUCUUGUCUAUGAUGUGAAACU

CG9526        5'-UUCCAAUAGAUAUUCUAUUAGGGGA
                     |||||||||| ||||  ||
snRNA:U2      3'-AUUCUUGUCUAUGAUGUGAAACUAGA

CG9752        5'-CCUAAAGACGGAUACUAUUCUGGAGAAC
                   ||| |||||||||||| ||
snRNA:U2      3'-CUAUUCUUGUCUAUGAUGUGAAACUAGA

CG6363        5'-AUGGGGAACGGGUUCUGUGCUUCCAC
                    |||||||||| ||||||||
snRNA:U2      3'-CUAUUCUUGUCUAUGAUGUGAAACUA

CG3552        5'-AACGGAAGGAAUAUGUAUAGUAUUUAACG
                    || ||||||| |||| ||||||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUA

CG5808        5'-  AUUUCAAAAUGGAUAAUGUGCUUAUCG
                      | | |||||||| |||||||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUA

CG8727        5'- TGACCGAGGACGCATCCTGTACGTTTCCGA
                       ||||||| || |||||| ||
snRNA:U2      3'- CUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU
```

```
CG6937        5'-TTCGGTCGGGTCGGGTGCTGCGCGTCCGTTTGG
                     || ||| |||||||||||| |  |||| |
snRNA:U2      3'- CUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG17294       5'-GAGAGACTAAATCCGTACTATATTCTGTCTGAG
                          |||  |||||||||| ||  |
snRNA:U2      3'- CUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG17531       5'-UUGGAGGACGAUGGACAUUAUAUCUGGGA
                    || || |||||| ||||||| | ||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUA

CG13151       5'-CAGUGGAAGCAUGGAUGCUGCUCUUCGAUGAUGAAA
                     | ||| ||||||||||| ||| |||  ||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAUCGG

CG14187       5'-  CCACGAUCGCAGAUACUUUACAUGAUGCGCA
                     | ||  ||||||||| ||| |
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG4645        5'-AUCGCUGCCAGCAGAUAUUUUUUUGGGCC
                    | ||  |||||||||| | ||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG17765       5'-UGCUCUGCGUGCAGGUGCAGCGUUUCACGGA
                      || | |||||||| ||||||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG2261        5'-CCGACUGCGAGCCGGUGCUGUGCCUGUCUUUUC
                      || |||| |||||||||| |   ||||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG2915        5'-GCAUUCAGAGGCGGGUAUUCCGAUUGCCUACA
                 ||    || |||||||||| || ||
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG7883        5'-UAUGGAAUCGGUGGACUUUGUGCUUGUGGGGGC
                   ||| |  ||||||  ||||||||  |
snRNA:U2      3'-CGACUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG7405        5'- AGCUAUGUGACGGAUCUUCUGUUUGUCUCCGC
                        | ||||||| || |||||   ||
snRNA:U2      3'-CGACUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG2098        5'-AATTTAAAAGGAACTATGTTTTTGCATAAC
                      || ||| |||||||||| |
snRNA:U2      3'- AUUCUUGUCUAUGAUGUGAAACUAGAAU

CG5032        5'- TGCCTTGGAACAGGTGCTGGAACTTCTTAA
                      | ||||||||||||| |  ||
snRNA:U2      3'- CUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU

CG30059       5'- TGCAGCAGCACGGATACAACACCTTCUUUGGC
                   | |  || |||||||| |||| ||  ||
snRNA:U2      3'- CUAUAUUCUUGUCUAUGAUGUGAAACUAGAAU
```

CG18278 and CG30059 are paralogs with the same target sequence.

Figure 2.22: **Additional predicted snRNP-mRNA base pairings.** These are only a subset of the most stable duplexes.

CG8108, CG5972 and many other transcripts (Figure 2.22). Those mRNAs within our dataset that are missing from the U1-70 K pulldowns (for example, CG1349 and CG4692) are plausibly bound by other Sm snRNPs such as U2, U4/U6, U5, U11 and U12. A list of such potential base pairing interactions was compiled by taking known single-stranded regions from snRNAs, and using them to find putative binding sites on the list of Smand U1-70 K-associated mature mRNAs (Figure 2.22). We found many potential sites with a duplex length and minimum free energy profile similar to the ones shown in Figure 2.16f. Taken together with the Sm and TMG IPs, these data suggest that snRNPs associate with subsets of mature Drosophila mRNAs, in a mode that is distinct from their interactions within the spliceosome.

To test whether base pairing between U1 snRNP and CG3776 mRNA is responsible for their interaction, we introduced three synonymous point mutations within the twelve-nucleotide complementary region in CG3776 mRNA that should completely block putative pairing with U1 snRNA (Figure 2.21a). We then transfected both wild-type and mutant CG3776 mRNA expression constructs into S2 cells (Figure 2.21b). The constructs are transcribed by an Act5C promoter and are terminated using the SV40 polyA signal and a heterologous 3' UTR. We confirmed that both transfections produced similar levels of chimeric CG3776 mRNAs (Figure 2.21c) and then performed Y12 IPs on S2 cell lysates, using normal goat serum as a control. As expected, 5S rRNA was not enriched in the IP fractions, whereas CG1349 mRNA and U1 snRNA were both significantly enriched in the transfections. Both endogenous and transfected CG3776wt mRNAs were pulled down by the Y12 antibody, whereas transfected CG3776mut mRNA was not (Figure 2.21d). These results support two conclusions. First, splicing is not required for U1 snRNP binding, and the binding site for U1 snRNP is located within the CG3776 mRNA coding sequence, since it can be efficiently pulled down by Y12 antibody. Second, the predicted U1 binding site is indeed necessary for U1 snRNP binding. Taken together, our results suggest that snRNPs bind mature mRNAs, and that at least one mechanism requires U1 snRNP base pairing with target mRNAs.

2.4 Discussion

We have developed an experimental and analytical pipeline to identify RNAs that stably associate with Sm proteins, an evolutionarily ancient group of RNA binding factors. The targeting of multiple subunits of an RNA-binding complex in this RIP-seq approach, along with the use of different genetic backgrounds, ensures that the identified RNPs are bona fide. Notably, this pipeline can be easily adapted to study other RNA-binding complexes.

2.4.1 Sm proteins in scaRNP complexes

We found that subsets of scaRNAs associate with Sm proteins, in both Drosophila and human cells. These include the highly conserved U85 scaRNA and newly evolved and non-canonical scaRNAs, such as scaRNA: Prp8 and SHAN, identified in this study. The involvement of Sm proteins in scaRNP biogenesis and function has been shown in several previous studies. Notably, both budding and fission yeast telomerase RNA precursors contain canonical Sm sites and are directly bound by Sm proteins (Leonardi *et al.*, 2008; Seto *et al.*, 1999). In fission yeast, Sm binding to telomerase RNA stimulates spliceosome-mediated cleavage that mimics the first step of splicing (Box *et al.*, 2008; Tang *et al.*, 2012). However, none of the scaRNAs we found in our IPs contain readily identifiable Sm sites. Fu and Collins Fu and Collins (2006) reported that SmB and SmD3, but not other Sm proteins, specifically associate with several human scaRNAs, and that this association requires a conserved CAB box sequence. Tycowski et al. Tycowski *et al.* (2009) showed that this CAB box is bound by a protein called WDR79. In our comprehensive analysis of fruit fly and human Sm-associated scaRNAs, we did not find additional sequence or structural features that distinguish them. Thus, these studies suggest an evolutionarily conserved role for Sm proteins in scaRNA biogenesis and function; however, the mechanism through which scaRNAs that lack identifiable Sm sites associate with Sm proteins is not well understood.

2.4.2 Splicing-independent, evolutionarily ancient functions for Sm-class snRNPs

The available single-stranded regions of snRNPs, which are used to identify intron-exon boundaries and intronic splicing elements, also serve as prime candidates for base pairing with mature mRNAs. We propose a model whereby Sm-class snRNPs interact with their targets via a combination of base pairing and protein-RNA interactions, as shown in Figure 2.21e. Indeed, this model has precedence, as the efficacy of this combination of interactions has already been demonstrated. Steitz and colleagues (Friend *et al.*, 2007) showed that both RNA-RNA and protein-RNA interactions are individually sufficient for function of the SF3b-hPrp43 subcomplex within the U2 snRNP in stimulating histone mRNA 3'-end maturation. In the current study, we showed that a sequence within CG3776 mRNA that potentially base pairs with the 5' end of U1 snRNP is required for binding. Mutation of this sequence abrogates U1 binding. By such a mechanism, snRNAs and/or specific proteins that bind to snRNPs could recruit other factors that, together, serve to regulate the processing, localization, translation or degradation of target mRNAs (Figure 2.21e).

Recently, Berg et al. Berg *et al.* (2012) proposed a function for U1 snRNPs, termed 'telescripting', whereby binding of U1 to nascent transcripts acts to suppress premature cleavage and polyadenylation at cryptic sites. Reduction of U1 snRNP levels elicited shortening of 3' UTR length and proximal 3' exon switching of numerous transcripts in a dose-dependent fashion (Kaida *et al.*, 2010; Berg *et al.*, 2012). This process is distinct from the interactions described here, as our data clearly showed snRNPs associating with mature mRNAs. Moreover, we did not observe significant enrichment of intronic regions in our RIP-seq datasets, as might have been expected if the telescripting interactions between U1 and post-splicing lariats were stable. Thus, the interactions described here with mature mRNAs are stable, likely taking place either in the cytoplasm or just prior to mRNA export.

Furthermore, the data indicate that U1 snRNP is not the only Sm RNP that associates with mature mRNAs. The U2 snRNP-histone mRNA interaction (Friend *et al.*, 2007) (and this work) is a case in point. We did not detect any downstream flanking sequences in our RIP-seq data, suggesting that the U2 snRNP maintains contact with the histone mRNA long after 3' end maturation, and therefore a potential function downstream of 3' end formation, for example, translational control. We also identified Sm- and TMG-associated mRNAs in S2 cells that are not enriched in U1-70 K IPs, most prominently CG1349 and CG4692. Interestingly, we found that the localization pattern of *Drosophila* CG4692 within stage 10 egg chambers (Figure 2.14) mirrored that of VFP-tagged Sm proteins (Gonsalvez *et al.*, 2010). Taken together, these findings suggest a general role for Sm-class snRNPs in post-splicing mRNA metabolism.

The Sm family of proteins is evolutionarily ancient. The eukaryotic Lsm1-7 complex regulates mRNA decapping and degradation by association with oligoadenylated mRNAs (Tharun, 2009; Parker and Sheth, 2007; Tharun *et al.*, 2000). The bacterial Sm orthologue, Hfq, also functions to regulate the translation and stability of a number of transcripts (for review see (Vogel and Luisi, 2011)). Similar to eukaryotic Sm proteins, prokaryotic Hfq forms a toroidal ring that binds a class of 50- to 200-nucleotide small (s)RNAs. These so-called 'sRNPs' bind to their targets, which include ribosomal protein (RP) mRNAs, via a combination of base pairing and protein-RNA interactions (Zhang *et al.*, 1998, 2002). Although the RP genes are not homologs of the RP mRNAs identified in this study, our findings nevertheless support the hypothesis that regulation of ribosome biogenesis is a deeply conserved function of Sm proteins.

Sequence covariation is generally considered a hallmark of conserved base-pairing interactions, underscoring functional importance. Not surprisingly, we found many covariant base pairs in the stem-loops of snRNA:LU and scaRNA: Prp8, despite their short evolutionary histories (Figure 2.9; Figures 2.12 and 2.10). However, we were unable to analyze this feature in our Drosophila and human Sm/snRNP-associated mRNAs, as no clearly orthologous mRNA transcripts were identified. Instead, we found that most of the targets of Sm proteins and snRNPs are different in the flies and human, with the exception of snRNAs and U85 scaRNA. This is consistent with the idea that protein- RNA and RNA-RNA interaction networks rapidly rewire themselves during evolution, despite the conservation of the individual components. For example, several studies on the RNA targets of Puf family proteins in yeast, fruit fly and human suggest that even though the binding sites of the proteins are conserved, the target mRNAs are not (Morris *et al.*, 2008; Gerber *et al.*, 2004, 2006). Similarly, Graveley and colleagues (Brooks *et al.*, 2011) showed that the binding sites for PS and NOVA1/2 are highly conserved between insects and mammals, but the target gene orthologs associated with PS and NOVA1/2 are almost entirely non-overlapping. This change of regulatory relationships in evolution has also been observed in the processing of minor introns and highly conserved micro- RNAs, such as let-7 and its targets [83,84].

2.4.3 Technical considerations

It is likely that the Sm-associated transcriptome is larger than the one described here. Although RNA-seq is quite sensitive, it may not be sensitive enough to reliably identify all of the low abundance transcripts from the relatively minute amount of immunopurified RNAs. The spliceosomal snRNAs comprise a majority of the immunopurified transcripts, limiting the ability of the sequencer to identify low abundance Sm-associated RNAs, especially scaRNAs and mRNAs. In addition, we employed a very stringent analysis procedure to ensure that the identified targets were not false positives. This procedure could also lead to false negatives. In our normalization, we assumed that the majority of RNAs do not associate with Sm proteins. This may or may not be true. There could be a very large number of transcripts that associate with Sm proteins with lower affinities than the ones identified in this study. The extent to which our assumption holds true will dictate the number of false negatives. Finally, as our qRT-PCR results suggest, certain RNA targets associate with Sm proteins in a tissue-specific fashion. Therefore, a comprehensive RIP-seq analysis of different tissues would be needed in order to identify all the targets of Sm proteins.

Recently, RNA crosslinking has been extensively used in characterizing targets of RNA binding proteins (Urlaub *et al.*, 2000, 2001; Ule *et al.*, 2003; Anko *et al.*, 2012). These methods not only provide evidence for direct interaction between RNAs and proteins, but can also achieve single-nucleotide resolution of the binding sites. However, such methods are not applicable to complexes that are refractory to crosslinking or interactions that are indirect. Canonical Sm proteins are poor substrates for UV crosslinking, even to the highly abundant snRNAs (Urlaub *et al.*, 2000, 2001). A more recent study used two different crosslinking methods to characterize the mRNA-associated proteome; they also failed to detect the canonical Sm proteins (Castello *et al.*, 2012). These investigators also identified the eIF4AIII component of the exon-junction complex (EJC), but not the other three EJC subunits (Castello *et al.*, 2012), which are presumably beyond the effective crosslinking radius. Because only eIF4AIII makes a direct contact with the mRNA, this result further supports the notion that crosslinking is not effective for studying all RNA-protein interactions. Our multiple-targeting strategy is therefore advantageous for the study of multimeric RNP complexes. The use of mock IPs as controls enables direct quantification of enrichment ratios, providing valuable information about the stability and affinity of the protein-RNA complexes. This point is illustrated by our RIP-seq data: the direct snRNA-Sm protein interactions are very stable, and correspondingly have much higher enrichment ratios than the mRNAs, which associate with Sm proteins indirectly.

2.5 Conclusions

The structural and functional similarities between prokaryotic sRNPs and eukaryotic snRNPs suggest that canonical Sm-class snRNPs have the potential to carry out multiple functions inside the eukaryotic cell. This study represents the first comprehensive analysis of eukaryotic Sm-containing RNPs, and provides a basis for additional functional analyses of Sm proteins/snRNPs outside of the context of pre-mRNA splicing. We have developed a flexible experimental procedure and robust statistical analysis methods to identify mRNAs that are associated with canonical Sm proteins in Drosophila and human cells. Using this pipeline, we confirmed and extended previous reports that Sm proteins associate with snRNAs, scaRNAs and histone mRNAs. Importantly, we also identified numerous Sm-associated mRNAs, along with several novel, previously unannotated snRNA and scaRNA transcripts. These newly discovered snRNAs and scaRNAs are highly conserved in the species with detectable homologs, suggesting that they are functionally important. The evidence indicates that the mRNA-Sm protein interaction is neither a consequence of splicing nor a product of Lsm1-7-dependent mRNA degradation. Instead, the interactions are mediated by snRNPs with mature mRNAs. Moreover, the fact that we did not identify intron-retained pre-mRNAs strongly suggests that the association between Sm proteins/snRNPs and mature mRNAs is more stable than the interactions within the spliceosome.

2.6 Materials and methods

2.6.1 Fly strains and cell lines

These previously described fly strains were used: Oregon R (OR, as the wild type), *nos-Gal4 VFP-SmB, nos-Gal4 VFP-SmD3, nos-Gal4 VFP-SmE, da-Gal4 VFP-SmD1, SmD3pt* and *Tralpt* from the fly-trap project (Quinones-Coello *et al.*, 2007a; Gonsalvez *et al.*, 2010). We characterized the insertion sites of P elements around the *LU* gene, and they are listed as follows. Line 10580 (k05816, y^{1} $w^{67}c^{23}$; *P{lacW}v(2)k05816*^{k05816}*, l(2)k05816*^{k05816}*/CyO*, from Bloomington Stock Center) and line 111186 (k05816, y^{d2} w^{1118} *P{ey-FLP.N}2 P{GMR-lacZ.C(38.1)}TPN1; P{lacW}v(2)k05816*^{k05816} *P{neoFRT}40A/CyO* y^{+}, from DGRC, Kyoto): CCCATCGAGT-GTCGGGGATC; line d04154 (*P{XP}v(2)k05816*^{d04154}): TCATAGCAAA-CATCCACCCC; line 203640 (*y1* $w^{67}c^{23}$; *P{GSV7}GS22096/SM1*, from DGRC, Kyoto): CGGCGCAAGT-GGCTGACTCA; line 103535 (*y* w*; P{GawB}v(2)k05816*^{NP0131}*/CyO, P{ UAS-lacZ.UW14}UW14*, from DGRC, Kyoto): CAACTGGTTA-TGGCAAGCCA. The following deficiency lines were obtained from stock collections: *Df(2L)Exel7014/CyO* (Exelixis collection at Harvard, stock no. 7784), and *Df(2L)BSC162/CyO* (BDSC at Bloomington, stock no. 9597). The flies were cultured on standard corn meal food at room temperature (22C) with 12 hour light-12 hour darkness cycles. Drosophila S2 cells were cultured in Express Five (Life Technologies, Carlsbad, CA, USA) plus 10% fetal bovine serum and penicillin/streptomycin, at room temperature (22C). Human HeLa cells were cultured in DMEM (Life Technologies) plus 10% fetal bovine serum and penicillin/streptomycin, in a 37C incubator with 5% CO_2.

2.6.2 RIP-seq experiment

Drosophila ovary RIP-seq These antibodies were used for IPs: Y12 (J Steitz, Yale, New Haven, CT, USA) (Lerner and Steitz, 1979), rabbit anti-GFP antibody (Abcam, ab6556, Cambridge, UK), agarose-conjugated anti-TMG (Calbiochem, La Jolla, CA, USA). For the Drosophila RIP-seq, ovaries were dissected from well-fed 3- to 4-dayold female flies. The IPs, RNA purification and reverse transcription were done essentially as described (Gonsalvez *et al.*, 2010). After first strand synthesis, the second strand was made using RNase H and DNA

Gene	Forward (5'-3')	Reverse (5'-3')	Size (bp)
U1:21D	ATACTTACCTGGCGTAGAGGTTAACC	GGAATGGCGTTCGCGCCGTCCCGA	164
U2:14B	ATCGCTTCTCGGCCTTATGGCTAAGATC	GTTGGGCCGAAGTCCCGGCGGTACTGCA	192
LU	ATGTCTCGATCGCCGCTTCAGTTGT	AATTGCCTCGGATAATGTGCTCATC	80
scaRNA:Prp8	ACGTTTCCAAGTGATCAGCCTCTCTGG	ATATGTATGCAACTATCAGCAGTCACGAT	168
5S rRNA	GCCAACGACCATACCACGCTGAA	AGTTGTGGACGAGGCCAACAACAC	120
Act5C	CGTCTTCCCATCGATTGTGGGACGT	AGTCGGTCAAATCGCGACCAGCCAGA	476
Smt3	CGGCATTCGACGCTCCGCAA	ATGGAGCGCCACCAGTCTGC	350
CG1349	TGTCGAAAAGCGCGCTGGTGAT	CGGCTACGGTGACCTTGATGCC	105
CG3776	CGGGAACGCGGCGAGGAAAT	CCGATTGGTGTCCAGCGGTGA	115
CG3776 mutagenesis	cggaatTATGTTTGACGATGCAAAC	acttgaTAGGGAAAGCTGAGGTATATG	
CG3776 tag (set2F & set2R)	TGTAAAGGAGTTCACCGCTGGACAC	CCTGCTAGCTTACGTCACCACTTTG	133
CG3997	CACAAGTCGTTCAGAATAAAGCAGAAGC	TGACGGCGCTTAGCGTTGTAACGA	119
CG4692	GCCCTTCGGCCAGGTCAAGC	CTTTGGGAACACGTACTTGTGCTGC	130
CG5972	AAATGAAACTGGCGGTCAATGCCAG	ACCGGGTCCATACTGGTTGCCT	91
CG7939 (RpL32)	CATCCGCCCAGCATACAG	CCATTTGTGCGACAGCTTAG	97
CG8108	AGTTCACTCACCACAACTCGAGCA	CGTTTGCGATCATCGCTGCGGTC	103
CG9042 (GAPDH)	CGTCAAGTACCTGAAAGGACACAAGC	CGAAGATCAGGATGTCAGCGTTCTTG	95
CG13410	AGAAGAGCACCTGCGTTTTGTATGGA	CAAACGCTTCGCAGCGCGCTT	99
dT.anchor for PAT:	GCGAGCTCCGCGGCCGCGTTTTTTTTTTTT	NA	NA
RpS2 PAT assay	CCTCGTCTGCACGCCGATGCCTAAGT	NA	116+
CG1349 PAT assay 1	GGTCTTCTTGTGGCCTACAACTAACA	NA	142+
CG1349 PAT assay 2	GCAAGGAGAAAGTCCAGGAGGT	NA	171+
CG3776 PAT assay	CATATAACATCGGCCCATGGCTA	NA	154+
CG3997 PAT assay	CTGTAAGCTGTTGATTCCAGGAG	NA	127+

Table 2.8: **List of primers and oligos used in this study.**The plus sign (+) for the sizes of the poly(A) length assay products indicates the smear, because the annealing of dT.anchor primer is random on the poly(A) tail.

polymerase I (Life Technologies, Carlsbad, CA, USA) according to the manufacturers' instructions. The resultant double-stranded cDNA was fragmented, ligated with Illumina sequencing adapters and sequenced in 36 cycles using the Genome Analyzer II platform at the UNC High Throughput Sequencing Facility. Random hexamer priming was used for reverse transcription for all seven cDNA libraries. In parallel, we also used oligo(dT)20 priming to generate cDNA libraries for four of the seven samples (Table 2.1).

Human HeLa cell RIP-seq HeLa cells were lysed and immunoprecipitated using the Y12 antibody. Four IPs and four normal goat serum controls (mock IP) were performed at the same time. The cDNA from these four controls and four IPs was used for real-time PCR analysis of selected transcripts. The RNA from two controls and two IPs was converted to cDNA libraries according to the Illumina TruSeq RNA SamplePrep Guide (version 2). The HeLa cell RIP-seq libraries were sequenced in 50 cycles. The RIP experiments for qRT-PCR were performed under more stringent conditions: 150 mM NaCl, 0.5% NP-40, 50 mM Tris-HCl, pH7.5 for incubation; 500 mM NaCl, 0.5% NP-40, 50 mM Tris-HCl, pH7.5 for washing. Dithiothreitol

(1 mM), RNase inhibitor (Superase-In, Life Technologies) and protease inhibitors (cOmplete, Roche Diagnostics, Indianapolis, IN, USA) were added to the buffer just prior to use.

2.6.3 RIP-seq read mapping and quantification

For the Drosophila RIP-seq experiments, sequencing reads were filtered using ELAND and those that passed the quality standard (Chastity >0.6) were mapped using Bowtie to the genome plus annotated transcriptome of *D. melanogaster* (Langmead *et al.*, 2009). Next, we used ERANGE software to count the reads that fall into existing gene models and to pile putative new exons (Mortazavi *et al.*, 2008). Clusters of reads that were close to known genes were either assigned as new exons of known genes or identified as novel transcripts on the basis of the read mapping pattern. Furthermore, because a number of Drosophila snRNA genes have multiple (two to seven) paralogs in the genome, we allowed up to ten mapped loci for each read. Subsequently, the repetitive reads were randomly assigned to mapped locations. The ERANGE final RPKM (reads per kilobase per million reads) data were converted to raw read numbers for each gene by using the calculated total number of reads for each sequenced library and the length of each gene. For each pair of control-IP experiments, we defined the read depth of a transcript d as the square root of the sum of the squares of number of reads in control and IP: d = sqrt (Ctrl*Ctrl + IP*IP). Raw read numbers for each gene between control and IP were normalized against the median of enrichment ratios for all expressed genes (with d >10). The HeLa cell RIP-seq experiments were performed in duplicates (two controls and two IPs) with paired-end sequencing technology. We therefore used standard t-tests from the Tophat/Cufflinks pipeline to analyze the human RIP-seq data (Trapnell *et al.*, 2012). The q values and expression difference scores from Tophat/Cufflinks analysis were directly used. The sequencing data are accessible at Gene Expression Omnibus with the accession number GSE35842.

2.6.4 Assignment of reads to Drosophila snRNAs

To calculate the enrichment ratios of snRNAs as shown in Figure 2.7 and Table 2.5, the total numbers of reads mapped to all paralogs of each snRNA species were pooled from both random hexamer primed libraries and oligo(dT) primed libraries (BAM files), and reads with mismatches were discarded. The following strategy is employed to assign reads to distinct snRNA paralogs. For U1, U4 and U5 snRNAs, reads overlapping the variable regions were identified from mapped RIP-seq BAM files, and reads with mismatches were discarded. For U2 snRNA, reads overlapping the four variable regions were used to calculate the fraction each isoform takes, then the total number of U2 reads (without mismatches) was redistributed according to the calculated fractions. (Details available on request; ZL and AGM, manuscript in preparation.)

2.6.5 *Drosophila* histone mRNA read mapping

Since the Drosophila replication-dependent histone genes are highly repetitive, we mapped all the RIP-seq reads to a single unit of the repeat, allowing no mismatches or indels. Then the read numbers were normalized against the median ratios obtained as mentioned above.

2.6.6 In situ hybridizations

Full length LU snRNA and CG4692 mRNA and their antisense transcripts were produced using the T7 in vitro transcription system (MEGAscript T7 Kit, Life Technologies), and labeled with digoxigenin-UTP (DIG). The DIG-labeled probes were hybridized to S2 cells and detected using the tyramide signal amplification kit (Life Technologies) as previously described (Gonsalvez *et al.*, 2010).

2.6.7 Gaussian mixture modeling

Gaussian mixture modeling was performed on logtransformed enrichment ratios for all the RNAs with a read depth >10. The normalmixEM function from the R package mixtools was used for the modeling (Young, 1998). Specifically, we restrained the number of normal distributions to two, and the two distributions were homoscedastic. For example: y <- normalmixEM(x, lambda = 0.5, mu = c (0, 2), sigma = (0.5)). Model fitting for all the six *Drosophila* RIP-seq experiments on canonical Sm proteins converged. However, the Tralpt RIP-seq data did not. Since the canonical Sm RIP-seq yields around 200 enriched RNAs on average, we therefore arbitrarily used the top 200 RNAs from the Tralpt RIP-seq for pairwise comparisons.

2.6.8 Cluster analysis of RIP-seq data

Enrichment ratios for every transcript in each of the seven RIP-seq experiments were log transformed. Then these enrichment ratios were clustered by experiment (but not genes) using Cluster 3.0 [94]. All available similarity metrics and clustering methods from the Cluster package were tried and all gave similar tree topology. After clustering, the data were visualized using Java Treeview (Saldanha, 2004). The aspect ratio of the whole data matrix was scaled to fit the presentation.

2.6.9 Fisher's exact test of the significance of overlap

A total of 5,296 (denoted as N) RNAs with read depth >10 was used as the superset. For each pair of comparison, with a and b enriched RNAs (let a <b), there are n overlapped RNAs. The Fisher's exact test P-value was calculated using the following R function: sum(dhyper(n:a, b, N-b, a, log = FALSE)).

2.6.10 Phylogenetic analysis

To identify the homologs of the newly discovered ncRNAs, we first examined the same syntenic block in other insect species. In addition, the D. melanogaster ncRNA sequences (including the promoter region, for LU snRNA) were used to BLAST against genome and transcriptome databases for homologs (Altschul *et al.*, 1990). Candidates were examined for the presence of signature sequence elements. The recovered sequences were aligned using ClustalW2

(Larkin *et al.*, 2007). The phylogenetic tree of the homologs was constructed using drawtree-0.1.3.

2.6.11 Meta-gene analysis of read density around splice junctions

One transcript from each *Drosophila* or human Smassociated intron-containing mRNA was randomly selected. Only internal exon-intron boundaries were used in this analysis. Reads were mapped using TopHat to increase the coverage around splice junctions. Reads mapped within a fifty nucleotide radius from the splice sites were counted from the following control and IP libraries (only random hexamer primed ones): Lu003-Lu004 (VFPSmD3), Lu007-Lu008 (VFP-SmD3), Lu011-Lu012 (VFPSmE), Lu015-Lu016 (VFP-SmB), Lu023-Lu024 (SmB), Lu025-Lu026 (SmD3pt), Lu045-Lu046-Lu047-Lu048 (human SmB). Scripts used for the analysis are available upon request.

2.6.12 Meta-gene analysis of read density along the entire gene length

One transcript from each *Drosophila* Sm-associated intron-containing mRNA was randomly selected. We manually determined the poly(A) site for each transcript. Read density along the gene length was extracted from wiggle files of the following data. The oligodT primed IP libraries were Lu002, Lu006, Lu010 and Lu014, and the random hexamer primed were Lu004, Lu008, Lu012 and Lu016. For each library preparation method, the reads for all enriched RNAs in four libraries were added and the coordinate adjusted to the poly(A) site. Read density was adjusted so that the maximum equals to 1. Read density as far as 1 kb from the poly(A) site was displayed. Scripts used for the analysis are available upon request.

2.6.13 Quantitative reverse-transcription PCR

Immunoprecipitated RNA was reverse transcribed with SuperScript III (Invitrogen) and digested with RNase H. Quantitative reverse-transcription PCR was performed using the SYBR Green master mix (Fermentas, Pittsburgh, PA, USA) on an ABI PRISM 7700 system (Applied Biosystems, Carlsbad CA, USA) according to the manufacturer's instructions. At least three biological replicates were performed for each experiment. RT-PCR primers are listed in Table 2.8. To test the significance of IP versus control for each RNA, we used one-sided t-test, assuming heteroscedasticity.

2.6.14 CG3776 construct and transfection

The CG3776 mRNA coding sequence (without the stop codon) was first cloned into pDONR221 and then transferred into pAW vectors using the Gateway system (Life Technologies). The three point mutations within the putative U1 binding site were introduced using Q5 Site-Directed Mutagenesis Kit (New England Biolabs, Ipswich, MA, USA). The construct expressed hybrid mRNA containing the CG3776 coding sequence and SV40/polyA 3' UTR. The constructs were transfected into S2 cells using electroporation (Amaxa Lonza, Basel, Switzerland). See Table 2.8 for the mutagenesis primers and realtime PCR primers.

2.6.15 Measurement of poly(A) tail length

Poly(A)-containing reads derived from a selected set of examples from the RIP-seq datasets were identified and summarized (Figure 2.19). PCR-based PAT assay was performed essentially as described (Salles and Strickland, 1995). Primers are listed in Table 2.8.

2.6.16 Analysis of U1-70 K RIP-seq data

The U1-70 K (two replicates) and Empty (four replicates) IP read files were downloaded from the modENCODE website. Reads were then mapped to the *Drosophila* genome and quantified using the TopHat/Cufflinks pipeline. For normalization of UCSC track files (wiggle, bedgraph, and so on) a given genome was divided into approximately 5,000 bins, and reads mapping to each bin were extracted from the track files. Only bins with significant read coverage were retained for subsequent analysis. The median of the ratios between the corresponding bins in two track files was used as the normalization factor.

2.6.17 RNA secondary structure and base pairing prediction

The secondary structures of the newly identified noncoding RNAs were predicted using either UNAfold or the Viena RNA Package with default parameter settings (Hofacker, 2003; Darty *et al.*, 2009). Secondary structures of the predicted RNAs were drawn using VARNA (Darty *et al.*, 2009). Structure alignment of ncRNAs was performed using LocARNA (global standard alignment) (Will *et al.*, 2007). Single stranded regions of the known snRNAs were used to screen for mRNA sequence complementarity with these regions using RNAhybrid (Rehmsmeier *et al.*, 2004). The minimum free energy was then calculated using the Vienna RNA package (Hofacker, 2003).

Abbreviations bp: Base pair; GFP: Green fluorescent protein; hTR: Human telomerase RNA; IP: Immunoprecipitation; mRNP: Messenger ribonucleoprotein; ncRNA: Non-coding RNA; PAR-CLIP: Photoactivatable-ribonucleoside-enhanced crosslinking and immunoprecipitation; PCR: Polymerase chain reaction; qRT-PCR: Quantitative reverse transcriptase PCR; RIP: RNA-immunoprecipitation; RNP: Ribonucleoprotein; scaRNA: Small Cajal body-specific RNA; snoRNA: Small nucleolar RNA; snRNA: Small nuclear RNA; snRNP: Small nuclear ribonucleoprotein; TMG: Trimethyl-guanosine; UTR: Untranslated region; VFP: Venus fluorescent protein.

Competing interests: The authors declare that they have no competing interests.

Authors' contributions ZL carried out most of the experiments and bioinformatic analyses. XG helped with the bioinformatics. CAS helped with the experiments in Figure 2.21. ZL and AGM designed the experiments and wrote the paper. All authors have read and approved the manuscript for publication.

Acknowledgements We are grateful to Mike and Becky Terns for gifts of antibodies. We thank the members of the Matera laboratory for numerous helpful comments and suggestions. We also thank Liang Shan (UNC Department of Computer Science) for advice and help with computer programming, Piotr Mieczkowski, Corbin Jones, Alicia Brandt and the UNC High Throughput Sequencing Facility for sequencing sample preparation. This work was supported by NIH grant R01-GM053034 (to AGM).

CHAPTER 3: Vicinal: a method for the analysis of chimeric reads from RNA-seq

Zhipeng Lu[1] and A Gregory Matera[2]

3.1 Abstract

Non-coding (nc)RNAs are important structural and regulatory molecules. Accurate determination of the primary sequence and secondary structure of ncRNAs is important for understanding their functions. During cDNA synthesis, RNA 3' end stem-loops can self-prime reverse transcription, creating RNA-cDNA chimeras. We found that chimeric RNA-cDNA fragments can also be detected at 5' end stem-loops, although at much lower frequency. Using the Gubler-Hoffman method, both types of chimeric fragments can be converted to cDNA during library construction, and they are readily detectable in high-throughput RNA sequencing (RNA-seq) experiments. Here, we show that these chimeric reads contain valuable infofrmation about the accurate boundaries of ncRNAs. We developed a bioinformatic method, called Vicinal, to precisely map the ends of numerous fruitfly, mouse and human ncRNAs. Using this method, we analyzed chimeric reads from over 100 RNA-seq datasets, the results of which we make available for users to find RNAs of interest. In summary, we show that Vicinal is a useful tool for determination of the precise boundaries of uncharacterized ncRNAs, facilitating further structure/function studies.

Key words: ncRNA, RNA-seq, self-priming

[1] This chapter was previously published in the journal *Nucleic Acid Research*, March 12th, 2014. Department of Biology, Integrative Program for Biological and Genome Sciences (IBGS), UNC Chapel Hill, Correspondence: zhipengluchina@gmail.com

[2] Departments of Biology and Genetics, IBGS, UNC Chapel Hill, Correspondence: matera@unc.edu

3.2 Introduction

Non-coding RNAs (ncRNAs) are functional RNA molecules that are not translated into proteins. Many categories of ncRNAs have been discovered and characterized. These include RNAs that carry out basic cellular functions such as pre-mRNA splicing (small nuclear RNAs, snRNAs) and mRNA translation (tRNAs and rRNAs) (Matera *et al.*, 2007). Also included are the small nucleolar (sno)RNAs and small Cajal body (sca)RNAs that guide post-transcriptional modification of rRNAs and snRNAs, respectively (Matera *et al.*, 2007). Not only are ncRNAs components of the core gene expression machinery, but they are also involved in multiple aspects of genetic regulation. This latter feature has been widely recognized with the discovery of microRNAs, siRNAs, piRNAs, lncRNAs etc. (Mattick and Makunin, 2006). The regulatory activities of the ncRNAs include roles in chromatin remodeling, transcription, splicing, translation, RNA stability and even the stability and translocation of proteins (Matera *et al.*, 2007; Mattick and Makunin, 2006; Kondrashov *et al.*, 2005; Yang *et al.*, 2001; Walter and Blobel, 1982). These functions usually depend upon their primary sequence and secondary structure in order to mediate interactions with proteins and other nucleic acids. Therefore, accurate determination of the RNA primary sequence is important for subsequent functional studies.

The rapid development in experimental and computational methodologies has significantly increased our ability to identify and study new ncRNAs. High throughput sequencing of the transcriptome (RNA-seq) has been widely used for its high sensitivity and nucleotide resolution, and revealed hundreds to thousands of short and long ncRNAs in various organisms in all three domains of life (Croucher and Thomson, 2010; Wang *et al.*, 2009a; Graveley *et al.*, 2011). De novo predictions based on evolutionary conservation and thermodynamic folding have also identified large numbers of ncRNAs and structured RNA elements in the genome (Washietl *et al.*, 2005; Pedersen *et al.*, 2006). However, these methods do not provide enough resolution to accurately define the ends of the ncRNAs (Will *et al.*, 2012), and ends of the most ncRNAs are not well defined. Traditional methods of RNA end determination, such as 5' RACE and 3' RACE, although accurate, are labor-intensive and suffer from very low-throughput (Scotto-Lavino *et al.*, 2006a,b). More advanced high-throughput experimental methods have been developed recently to map RNA ends, e.g. (Takahashi *et al.*, 2012; Ruan and Ruan, 2012), but many of these methods are complicated and/or require the presence of poly(A) tails. In addition, new ways of analyzing the vast amount of existing RNA-seq data will be cost-effective and useful for gaining insights into various aspects of RNA structure and processing.

The traditional method for preparing cDNA libraries was developed by Gubler and Hoffman (Gubler and Hoffman, 1983), which uses reverse transcriptase for first strand cDNA synthesis, RNase H, E. coli DNA polymerase I and DNA ligase for second strand synthesis. This method is also commonly used for RNA-seq library preparation. Within certain RNA-seq data sets whose libraries were prepared using the Gubler-Hoffman method, we have discovered that a large number of the 'unmappable' reads are chimeric. That is, these reads consist of two parts: one from the 5' or 3' end of the RNA, and the other from an internal region of the RNA, on the opposite strand. This phenomenon clearly suggests self-priming from the 3' end stem-loop, or ligation of the 5' end stem-loop during cDNA library preparation. Using the chimeric reads from existing data sets, we developed a program, called Vicinal, to precisely determine the boundaries of ncRNAs and provide support for the predicted terminal stem-loops.

3.3 Results

Previously, we carried out an RNA-immunoprecipitation sequencing (RIP-seq) analysis to identify RNAs that co-purify with Sm proteins in *Drosophila* and human cells (Lu *et al.*, 2014). During preparation of the sequencing libraries, we used either oligo-dT or random hexamer primers for first strand cDNA synthesis from RNA. Curiously, we found that both random and oligo-dT primed libraries contained large numbers of snRNA transcripts. This latter result was unexpected because snRNAs are not polyadenylated. To explain this observation, we considered the possibility that the snRNA reads detected in the oligo-dT primed libraries might be oligoadenylated RNA degradation intermediates (Nakamura *et al.*, 2008). However, manual inspection of reads derived from snRNA 3' ends did not reveal oligo(A) extensions in either the oligo-dT or random hexamer primed libraries; instead, the (non-templated) extensions appear to be products of self-priming from stem loop sequences that are typically present at the 3' ends of snRNAs. Though much less frequent, we also found reads that contain 5' extensions, which might be the result of ligation events between cDNA and RNA 5' stem loops. A diagram of possible mechanisms for the generation of 5' end and 3' end chimeric reads is presented in Figure 3.1. The 3' end stemloop can serve as a primer for first strand cDNA synthesis. cDNA fragments that are close to the 5' end stemloop can be ligated with the 5' end RNA during by DNA ligase. The two types of DNA-RNA chimera could, in principle, serve as templates for second strand synthesis. The resultant double stranded DNA could then be further ligated with adapters and sequenced. The 5' cap structures present at many ncRNAs provides one explanation for the low efficiency of the 5' ligation. However, it is not entirely clear how DNA polymerase I uses DNA-RNA chimerae as templates for second strand synthesis. The diagram in Figure 3.1 is provided for illustrative purposes, to help understand how chimeric reads might be generated.

3.3.1 The Vicinal algorithm

Initial examination of the chimeric reads derived from snRNAs reveals two important features, irrespective of whether they arose via self-priming or ligation. First, the two parts of each chimeric read map close to one another, usually within 100 nt. This distance is basically determined by the size of the terminal stem-loop. Second, the two parts of each chimera map to opposite strands of the encoding DNA, unlike reads derived from spliced RNAs, which map to the same strand. Based on these properties, we developed an analysis pipeline to identify reads that are derived from self-priming and ligation events (Figure 3.1).

For Vicinal mapping to work, the RNA-seq libraries must be prepared in a way that allows for self-priming. This is usually accomplished by cDNA synthesis prior to adapter ligation, on RNA samples that contain ncRNAs (Figure 3.1A). Because chimeric reads represent only a small portion of the total number of reads (see read mapping statistics in Table 3.1, and Figure 3.2A and 3.2B for an example), efficient processing of raw RNA-seq data is important for subsequent analysis. We used Bowtie2 for preliminary mapping, because it is fast and allows soft-clipping of the reads for local (partial) mapping (Langmead and Salzberg, 2012). The Bowtie2 mapping results provide a rough estimate of the coverage and size of transcripts. For the locally (partially) mapped reads, only the longer segment of the read is mapped to the

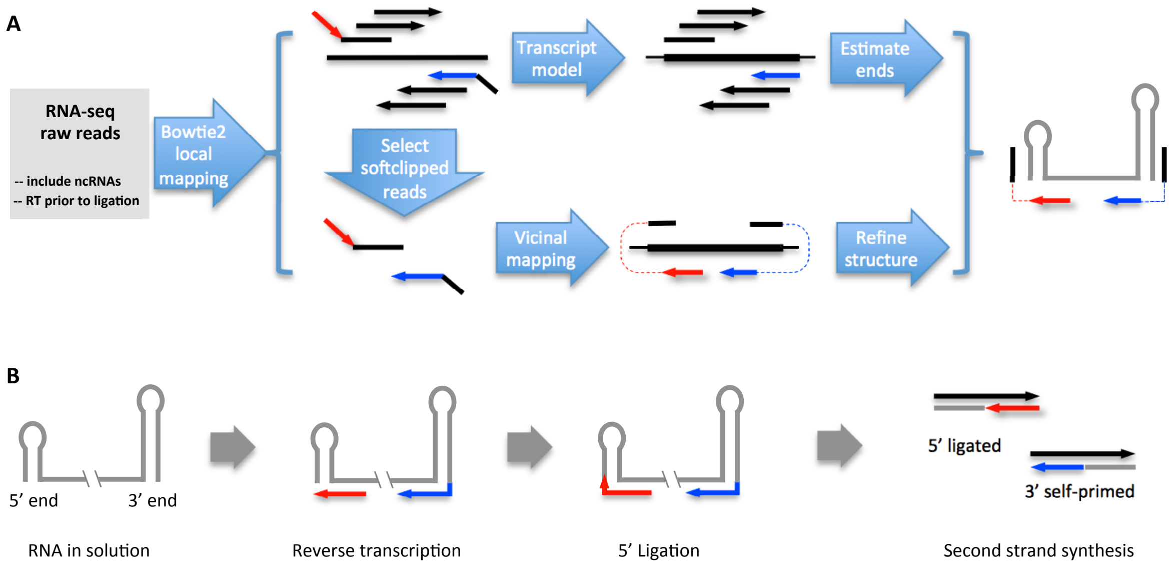


Figure 3.1: **Vicinal pipeline and possible mechanisms for the generation of chimeric reads.** (A) Flowchart of the analysis pipeline. The RNA-seq libraries are prepared such that they do not exclude ncRNAs, and the reverse transcription step precedes adapter ligation. The RNA-seq reads are first aligned to the genome using Bowtie2 in the local-mapping mode (–sensitive-local). Then partially mapped reads are selected and vicinally mapped using Vicinal. Bowtie2 mapped reads are directly used to roughly estimate the boundaries of the ncRNAs, while the Vicinal mapped reads are used to accurately determine the boundaries. Finally, the secondary structure is predicted and adjusted to fit the chimeric reads. The convention of colors is consistent throughout all the figures. (B) Many ncRNAs in solution adopt secondary structures with terminal stemloops. During reverse transcription, the 3' end stemloop can serve as primer, in addition to primers added to the solution, for cDNA synthesis (red and blue lines with arrows are cDNA fragments). After cDNA synthesis, the cDNA fragments close to the 5' end stemloop can be ligated to the stemloop. The 5' end and 3' end cDNA chimeras can further serve as templates for the second strand DNA synthesis, thus producing cDNA fragments for subsequent adapter ligation and deep sequencing.

genome, whereas the shorter segment is softclipped/ignored.

After initial mapping step, we filtered the mapped reads to select those with at least one softclipped segment longer than a defined size (e.g. 5 nt). The size of the softclipped segment is chosen so that the fragment can be uniquely mapped in the vicinity of the Bowtie2 mapped part of the read. The softclipped segments are then mapped 'vicinally', that is, mapped to a region within a certain distance (e.g. 100 nt) from the mapped segment, on the opposite strand (Figure 3.1A).

Once both segments are mapped, the junction is used to define the ends of the ncRNA,

Sample	Reference	Read length	Mappable	Softclipped	Chimeric	%Chimeric
fly_ovary_RIP	Lu et al.2014	35	87594638	11334332	990969	1.13
fly_pharate	This study	48	124544603	13919998	226981	0.18
fly_S2	Smith et al. 2011	45/50	224898608	18788217	387406	0.17
fly_larva3	Garcia et al. Unpublished	48	241674561	25221457	1294206	0.53
fly_pupa	Garcia et al. Unpublished	48	129388545	12737427	182526	0.14
mouse_ES	Smith et al. 2011	40	101615022	5573492	93137	0.09
mouse_ES	Hu et al. 2013	51	712409456	53374271	438910	0.06
mouse_satellite	Mousavi et al.2013	50	41019420	1910902	57034	0.14
human_HCT116	Hu et al. 2013	50	90190285	6695509	123856	0.14

Table 3.1: **RNA-seq datasets used in the study.**

and terminal stem-loops in the predicted secondary structures are used to explain the source of the chimeric reads (Figure 3.1A and 3.1B). However, the presence of self-primed and ligated reads does not imply that the chimera-generating terminal stem-loops are stable in vivo. It is only evidence for the presence of the terminal stem loops in solution, likely in equilibrium with other conformations (see Fig 2H and 3E, 3H, 3J and 3L). We generated lists of ncRNA genomic coordinates and used them to intersect with chimeric reads generated by Vicinal to make lists of ncRNAs with numbers of chimeric reads (see instructions and the results: https://sites.google.com/site/zhipeng0426/programming).

We have analyzed hundreds of RNA-seq datasets using Vicinal, and found 115 of them containing self-primed and ligated chimeric reads. These datasets are sorted into to 9 different groups, according to organism, tissue and/or read length (Table 3.1). These data were generated by several different laboratories, demonstrating that such chimeric reads are not specific artifacts of a single lab. Five of the nine groups are sourced from the fruitfly, three from the mouse and one from human. Statistics of the Bowtie2 local mapping, filtering and Vicinal mapping are presented. Although the fraction of chimeric reads is not high in any of the datasets, there are enough of them to determine the ends of many ncRNAs. Given the large number of starting raw reads in most RNA-seq experiments, Vicinal analysis provides users with numerous ncRNAs with sufficient chimeric read coverage. These include snRNAs, snoRNAs, scaRNAs, 5.8S rRNA, 7SK RNA, 7SL RNA, RNaseP RNA, RNaseMRP RNA, etc. (see Figures 3.2, 3.3, 3.4 and 3.5 and ncRNA lists from GSE50711). The chimeric read coverage for these ncRNAs varies greatly, from dozens to thousands of reads per RNA. However, certain types of ncRNAs, e.g. lncRNAs, miRNAs, siRNAs etc. do not have chimeric reads and thus the Vicinal program is not applicable. These results demonstrate the utility of our approach in the analysis of multiple categories of ncRNAs, with a wide range of expression levels, in different organisms. Here we show several examples of using Vicinal to analyze several known and newly discovered fly ncRNAs. More examples of Vicinal analysis on fly, mouse and human ncRNAs are presented in Figures 3.4 and 3.5.

3.3.2 Confirmation of known snRNA:U1 and RNaseP:RNA ends using Vicinal

Most snRNAs have stem-loops at their 5' and 3' ends with very short overhangs, and their sequences and secondary structures are well characterized. Sm protein immunoprecipitations enrich snRNAs and among the snRNAs, U1 is the most abundant; therefore we first analyzed chimeric reads derived from U1 snRNA as a proof of principle (Figure 3.2A-E). The chimeric reads for U1 is most abundant in the fly_ovary_RIP_35nt sample and therefore we only showed this group of RNA-seq data in the Vicinal analysis of U1.

The read coverage patterns for U1 snRNA, using Bowtie (end-to-end mapping) and Bowtie2 (local mapping) are not uniform (Figure 3.2A, and for other ncRNAs, data not shown). This is especially true near the two ends, because untemplated extensions in the reads are not mappable, and the priming and sequencing efficiency along U1 varies according to sequence and structural contexts. The non-uniformity of the read coverage makes estimation of transcript sizes difficult. However, selection of the Bowtie2 locally (partially) mapped reads clearly shows that many of the mappable fragments are justified to the left or right ends, suggesting the existence of softclipping in Bowtie2 mapping (Figure 3.2A and 3.2B). Vicinal mapping during Vicinal analysis of the chimeric reads places the unmapped segments on the opposite strands (Figure 3.2A and 3.2B) and the terminally mapped half-reads indicate the presence of terminal stem-loops (Figure 3.2C). Patterns of read coverage showed clear end justification (Figure 3.2A and 2B, the dashed blue line, see also examples in subsequent figures.).

Detailed alignment of the partially-mapped U1 reads showed clear signs of chimera formation (Fig 2C). The presence of terminal overhangs and imperfect complementarity in the stem allows

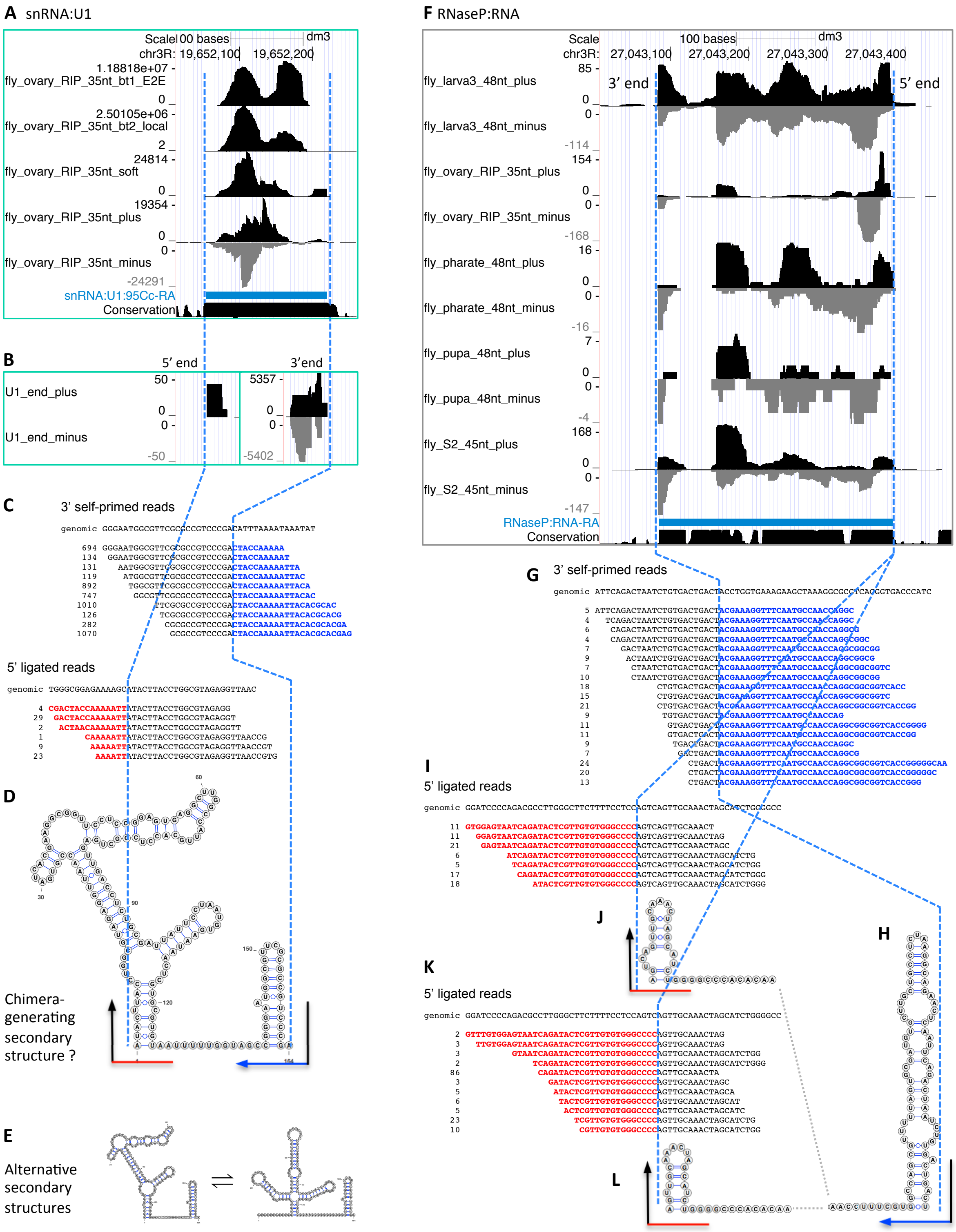


Figure 3.2: **Vicinal analysis of known fly ncRNAs, snRNA:U1 (A-E) and RNaseP:RNA (F-L).** (See figure legend on the next page.)

Figure 3.2: **Vicinal analysis of known fly ncRNAs, snRNA:U1 (A-E) and RNaseP:RNA (F-L).** (A) Five genome browser tracks for snRNA:U1:95Cc are shown from the analysis of fly_ovary_RIP_35nt datasets: bt1_E2E (Bowtie1 end-to-end mapping of raw reads), bt2_local (Bowtie2 local mapping), soft (selecting softclipped reads from bt2_local), plus and minus (reads mappable by Vicinal). The thick blue bar represents the mature U1:95Cc transcript region. The vertical dashed blue lines align the 5' and 3' ends of the U1 transcript. Note that there are five functional U1 snRNA genes in the fly genome and only one is shown here. The high peaks mapped to the middle of U1 are artifacts from mapping short softclipped parts of reads. (B) Filtering reads from fly_ovary_35nt_plus/minus for reads close to the estimated U1 snRNA ends showed clear terminal pileup of reads. Reads for all five U1 snRNA paralogs were combined. Note the difference in scale for the 5' end and 3' end chimeric reads. (C) Detailed analysis of the chimeric reads that map to both ends. The first line is the genomic DNA sequences around the 5' and 3' ends. Subsequent lines are the manually aligned chimeric reads, where black letters represent parts mapped to the ends of the transcript, blue letters represent 3' extensions mapped to the internal region on the opposite strand, and the red letters represent 5' extensions mapped to the internal region on the opposite strand. The numbers before each chimeric read sequence are read counts. Note the differences between the extended genomic DNA and the terminal extensions in the chimeric reads. Only top 10 groups of distinct reads were shown for the 3' end, and top 6 groups of distinct reads were shown for the 5' end. (D) Predicted secondary structure that explains the production of the chimeric reads. The black lines represent parts of reads mapped to the ends of U1 snRNA, whereas the red and blue lines represent terminal extensions mapped to the internal regions of U1 snRNA. (E) Potential equilibrium in solution between the chimera-generating secondary structure (on the left, the same as in D) and the well-known physiological secondary structure in U1 snRNP (on the right). The normal secondary structure is not likely to give rise to 5' end ligated reads due to the long 5' overhang. (F) Ten genome browser tracks for RNaseP:RNA are shown from the Vicinal analysis of five groups of fly RNA-seq data. The end-to-end mapping, local mapping and softclipped read tracks are not shown. Note the terminally adjusted read pileups; they are not filtered as in B for U1 snRNA. The 5' end of the RNaseP:RNA is on the right. Chimeric reads were combined from all five groups of RNA-seq data, for subsequent detailed analysis. (G) Detailed analysis of the chimeric reads that map to 3' end of RNaseP:RNA. (H) The chimera-generating secondary structure of the 3' end of RNaseP:RNA. (I and K) Detailed analysis of reads mapped the 5' end reveals two possible 5' ends that differ by 4 nucleotides. (J and L) The chimera-generating secondary structures of the 5' end of RNaseP:RNA. The chimera-generating secondary structures shown here for the 5' end and 3' end are different from the physiological secondary structure in RNaseP.

for accurate definition of the boundaries. Importantly, the abundance of the 3' end-derived chimeric reads confirmed the stable 3' end stem-loop which allows for efficient self-priming, despite the presence of imperfect complementarity. In fact, the presence of base pair mismatches made it possible to define the ends with near single-nucleotide resolution. The identity of the few additional nucleotides (usually 1-2 nucleotides) close to the end of the mature transcript may interfere with the accuracy of end determination, but most of the time, they are short enough to allow accurate determination. In contrast to the 3' end reads, there were many fewer reads derived from the U1 5' end. The relative dearth of 5' end chimeric reads is likely due to the fact that the 5' overhang in the predicted secondary structure is quite long (11nt) and that ligation to first strand cDNA is likely to be very inefficient, due to the presence of the TMG cap. Because cDNA library construction takes place on purified RNAs and not on stable RNPs, U1 may well adopt alternative secondary structures in solution. One such alternative U1 structural isomer (see Figure 3.2E) has no overhang, and might be a better substrate for generation of chimeric 5' end reads. Irrespective of the mechanism, the structure of the observed U1 chimeric reads is consistent with the known 5' and 3' ends of U1 snRNA (Lo and Mount, 1990).

RNaseP:RNA is a ribozyme that cleaves pre-tRNA 5' end leader sequences during tRNA biogenesis and is an essential RNA in all life forms. Here we present a detailed analysis of RNaseP:RNA ends using Vicinal (Figure 3.2F-L). Chimeric reads are detectable for RNaseP:RNA in all five groups of fly RNA-seq data, with varied abundance (Figure 3.2F). We combined chimeric reads from all the five groups of RNA-seq data, and Vicinal analysis revealed clear terminally justified chimeric reads for both 5' and 3' ends. The 3' end chimeric reads clearly define a single end, with a maximum of 2 ambiguous nucleotides (Figure 3.2G and H), consistent with previous report (Marquez *et al.*, 2005). The 5' ligated reads suggest two possible ends, differing by 4 nt (Figure 3.2I-L). One of the two 5' ends is consistent with previous reports (Figure 3.2I and J) (Marquez *et al.*, 2005). It is likely that the other one (Figure 3.2K and L) represents a transcript that uses a different transcription start site or is subject to alternative 5' processing. The terminal stem-loops that explain generation of 5' and 3' chimeric reads are different from the physiological secondary structure (present in the RNaseP RNP particle), but nonetheless they are very likely to exist in solution (Marquez *et al.*, 2005). Taken together, our method accurately defines the boundaries of two known ncRNAs.

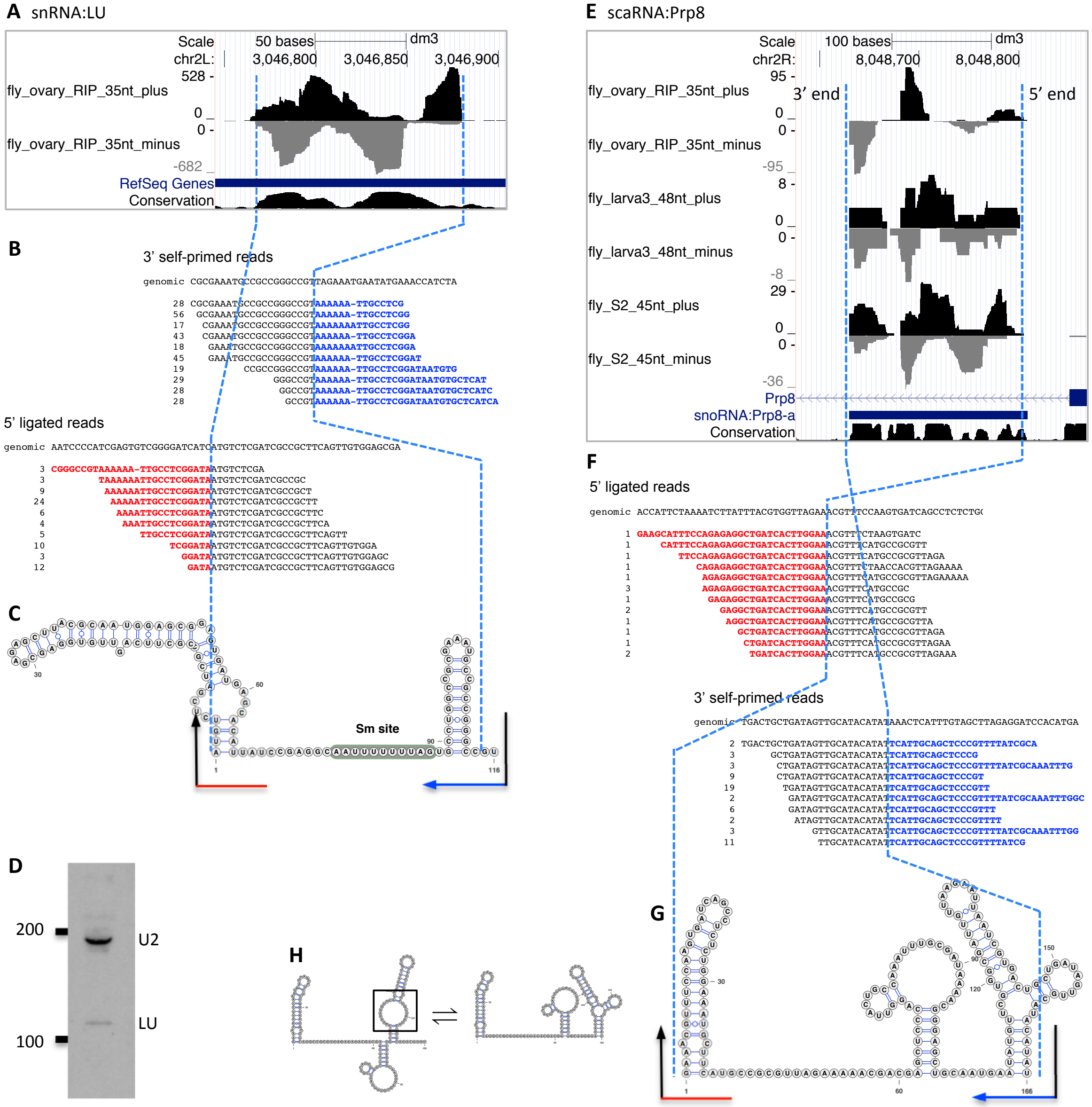
3.3.3 Vicinal analysis of newly-discovered snRNAs and sno/scaRNAs

In order to show the utility of Vicinal in analysis of novel or under-studied ncRNAs, we have examined all *Drosophila*, mouse and human ncRNAs using Vicinal and detected chimeric reads in many of them. Here we show two ncRNAs that we discovered in our previous RIP-seq analysis (17). Dozens of additional examples are presented in Figures 3.4 and 3.5.

(i) snRNA:LU. Like-U is a newly evolved Sm-class snRNA (CR43708), present only in *Drosophilid* genomes (Lu *et al.*, 2014). Vicinal analysis of LU snRNA revealed hundreds of terminally-justified fragments and internally mapped second fragments at both the 5' end and 3' end of the transcript (Figure 3.3A). We analyzed these chimeric reads, and predicted that the length of LU snRNA is 116 nt. The predicted secondary structure is shown in Figure 3.3C. Previously, Jung et al. Jung *et al.* (2010) analyzed publicly available RNA-seq data and identified a transcript from this locus, estimating its length to be 150-160 nt. To resolve this difference in length prediction, we performed northern blotting of LU snRNA, which showed a size that is consistent with our Vicinal analysis (110-120 nt, Figure 3.3D). This size is also consistent with the sequence conservation of LU orthologs among *Drosophilids* (Lu *et al.*, 2014).

(ii) scaRNA:Prp8. Another novel Sm-associated ncRNA we discovered in our RIP-seq study is scaRNA:Prp8 (CR43600; Figure 3.3C). A previous transcriptomic study estimated its size to be 178 nt (Graveley *et al.*, 2011). Here, Vicinal analysis of three groups of RNA-seq data revealed dozens of terminally justified fragments and predicting a length of 168 nt (Figure 3.3E-H). This size is also consistent with the alignment of scaRNA orthologs in insects. The chimeric reads can be explained by an alternative conformation of the secondary structure (Figure 3.3G and Figure 3.3H, right side). The other conformation (Figure 3.3H, left side) is likely to be the physiological one, due to the presence of an open pseurouridylation pocket (Figure 3.3H, black box), as reported recently by Deryusheva and Gall (Deryusheva and Gall, 2013).

The determination of ncRNA ends following Vicinal mapping requires manual alignment of the chimeric reads and fitting onto predicted secondary structures. In order to make best use of the large amounts of chimeric reads mapped using Vicinal from *Drosophila*, mouse and human RNA-seq datasets, we provide them as lists for users to identify ncRNAs of their interest. The lists contain ncRNA identifiers and numbers of chimeric reads mapped to each ncRNA. Chimeric read coverage patterns can be visualized by importing the bedgraph track files into genome browsers. The chimeric reads for each ncRNA can be extracted from the Vicinal mapped BAM files and manually aligned. In the future, additional RNA-seq datasets can be added to



3' self-primed reads

```
genomic CGCGAAATGCCGCCGGGCCGTTAGAAATGAATATGAAACCATCTA

     28 CGCGAAATGCCGCCGGGCCGTAAAAAA-TTGCCTCG
     56  GCGAAATGCCGCCGGGCCGTAAAAAA-TTGCCTCGG
     17   CGAAATGCCGCCGGGCCGTAAAAAAATTGCCTCGG
     43   CGAAATGCCGCCGGGCCGTAAAAAA-TTGCCTCGGA
     18    GAAATGCCGCCGGGCCGTAAAAAAATTGCCTCGGA
     45    GAAATGCCGCCGGGCCGTAAAAAA-TTGCCTCGGAT
     19          CCGCCGGGCCGTAAAAAA-TTGCCTCGGATAATGTG
     29               GGGCCGTAAAAAA-TTGCCTCGGATAATGTGCTCAT
     28                GGCCGTAAAAAA-TTGCCTCGGATAATGTGCTCATC
     28                 GCCGTAAAAAA-TTGCCTCGGATAATGTGCTCATCA
```

5' ligated reads

```
genomic AATCCCCATCGAGTGTCGGGGATCATCATGTCTCGATCGCCGCTTCAGTTGTGGAGCGA

      3 CGGGCCGTAAAAAA-TTGCCTCGGATAATGTCTCGA
      3         TAAAAAATTGCCTCGGATAATGTCTCGATCGCCGC
      9          AAAAAATTGCCTCGGATAATGTCTCGATCGCCGCT
     24           AAAAATTGCCTCGGATAATGTCTCGATCGCCGCTT
      6            AAAATTGCCTCGGATAATGTCTCGATCGCCGCTTC
      4             AAATTGCCTCGGATAATGTCTCGATCGCCGCTTCA
      5                TTGCCTCGGATAATGTCTCGATCGCCGCTTCAGTT
     10                     TCGGATAATGTCTCGATCGCCGCTTCAGTTGTGGA
      3                       GGATAATGTCTCGATCGCCGCTTCAGTTGTGGAGC
     12                        GATAATGTCTCGATCGCCGCTTCAGTTGTGGAGCG
```

5' ligated reads

```
genomic ACCATTCTAAAATCTTATTTACGTGGTTAGAAACGTTTCCAAGTGATCAGCCTCTCTG(

      1 GAAGCATTTCCAGAGAGGCTGATCACTTGGAAACGTTTCTAAGTGATC
      1     CATTTCCAGAGAGGCTGATCACTTGGAAACGTTTCATGCCGCGTT
      1        TTCCAGAGAGGCTGATCACTTGGAAACGTTTCATGCCGCGTTAGA
      1           CAGAGAGGCTGATCACTTGGAAACGTTTCTAACCACGTTAGAAAA
      1            AGAGAGGCTGATCACTTGGAAACGTTTCATGCCGCGTTAGAAAAA
      3            AGAGAGGCTGATCACTTGGAAACGTTTCATGCCGC
      1             GAGAGGCTGATCACTTGGAAACGTTTCATGCCGCG
      2               GAGGCTGATCACTTGGAAACGTTTCATGCCGCGTT
      1                AGGCTGATCACTTGGAAACGTTTCATGCCGCGTTA
      1                  GCTGATCACTTGGAAACGTTTCATGCCGCGTTAGA
      1                   CTGATCACTTGGAAACGTTTCATGCCGCGTTAGAA
      2                    TGATCACTTGGAAACGTTTCATGCCGCGTTAGAAA
```

3' self-primed reads

```
genomic TGACTGCTGATAGTTGCATACATATAAACTCATTTGTAGCTTAGAGGATCCACATGA

      2 TGACTGCTGATAGTTGCATACATATTCATTGCAGCTCCCGTTTTATCGCA
      3      GCTGATAGTTGCATACATATTCATTGCAGCTCCCG
      3       CTGATAGTTGCATACATATTCATTGCAGCTCCCGTTTTATCGCAAATTTG
      9       CTGATAGTTGCATACATATTCATTGCAGCTCCCGT
     19        TGATAGTTGCATACATATTCATTGCAGCTCCCGTT
      2         GATAGTTGCATACATATTCATTGCAGCTCCCGTTTTATCGCAAATTTGGC
      6         GATAGTTGCATACATATTCATTGCAGCTCCCGTTT
      2          ATAGTTGCATACATATTCATTGCAGCTCCCGTTTT
      3             GTTGCATACATATTCATTGCAGCTCCCGTTTTATCGCAAATTTGG
     11              TTGCATACATATTCATTGCAGCTCCCGTTTTATCG
```

Figure 3.3: **Vicinal analysis of snRNA:Like-U (LU) (A-D) and scaRNA:Prp8 (E-G).** Please refer to Figure 3.2 for general description of the analysis flow. (A) Only two genome browser tracks for snRNA:LU are displayed: plus and minus, whereas the end-to-end mapping, local-mapping and softclipped read tracks were not shown. Note the size predicted by Jung et al. (Jung *et al.*, 2010) (transcript model, the thick blue line) is longer than the size determined by Vicinal analysis. (B) Detailed analysis of the chimeric reads. Only top 10 groups of distinct reads were shown. The 3' ligated reads included variants, because sequencing the long stretch of adenosines unavoidably introduces errors. (C) The chimera-generating secondary structure for snRNA:LU. The Sm site is shown on the predicted secondary structure. (D). Northern blot of *Drosophila* U2 and LU snRNAs. (figure legend continued on the next page)

Figure 3.3: (E) Six genome browser tracks for scaRNA:Prp8 were shown from the analysis of three groups of RNA-seq datasets, where chimeric reads for this RNA is available. Note that the earliest annotation labeled this ncRNA as snoRNA (as shown in the gene annotation track); however, later research suggests that it is a scaRNA. (F) Detailed analysis of the chimeric reads. (G) The chimera-generating secondary structure for scaRNA:Prp8. (H) Potential equilibrium between the more likely physiological secondary structure (on the left, with the pseudouridylation pocket open in the last stemloop in a black box) and the chimera-generating secondary structure (on the right).

increase the chimeric read coverage on ncRNAs in the species analyzed, and potentially in other species as well.

3.4 Discussion

In this study, we present a new bioinformatic tool, called Vicinal, to define the ends of ncRNAs with terminal stem-loops. This method takes advantage of the self-priming and ligation property of ncRNA 3' and 5' terminal stem-loops during library preparation using the Gubler-Hoffman method (Gubler and Hoffman, 1983), and the power of massively parallel sequencing. Using Vicinal, we confirmed the boundaries of previously studied ncRNAs, and also defined the boundaries of newly discovered ncRNAs from many different RNA-seq datasets from various species.

Although other methods are available for the determination of ncRNA ends, many of them are labor-intensive and require more experiments. Our analysis method makes use of published RNA-seq data and is cost-effective. More accurate determination of ends for more ncRNAs will be available with the publication of ever increasing amount RNA-seq data.

It has long been known that 3' end self-priming of U3 snoRNA mediates pseudogene formation during the process of retrotransposition (Bernstein *et al.*, 1983). Pseudogenes derived from other highly structured ncRNAs, including U1 and U2 snRNAs, are also known to form in this manner. Furthermore, self-priming from 3' end stem-loops is a relatively common feature among certain single-stranded RNA and DNA viruses (Salzman and Fabisch, 1979; Bourguignon *et al.*, 1976; Tuiskunen *et al.*, 2010). This self-priming ability is required for proper replication of the viral genome. Moreover, the high efficiency and specificity of self-priming from terminal stem-loops has been exploited for quantification of small RNA levels by RT-PCR, wherein a stem-loop RT primer is used instead of a conventional, unstructured primer (Chen *et al.*, 2005a). In contrast to the widespread use of 3' self-priming in nature, we are not aware of previous findings regarding the phenomenon of ligation to 5' end RNA stem-loops during cDNA library construction, and further studies will be needed in order to understand the mechanism.

The use of soft-clipped reads for mapping inevitably creates artifacts. Vicinal mapping sometimes assigns reads to exon-exon junctions, due to their short length after clipping (data not shown). Other kinds of artifacts are also observed, mainly in highly expressed ncRNAs, such as ribosomal RNAs and some snRNAs (Figure 3.2A). However, such artifacts can be clearly distinguished from chimeras that are generated by self-priming and ligation. The latter have distinct features, such as terminally justified pileups. We note that certain RNA secondary structures are likely to be more favorable for chimera formation than others. Although such structures are not necessarily the most stable ones in solution or in vivo (see Figures 3.2D, H, J, L and Figure 3.3G), the boundary mapping procedure described here can easily pick up such low-frequency priming events.

In summary, the method described above enables highly sensitive analysis of ncRNA boundaries. The use of fast short-read mappers (we used Bowtie2) in combination with rapid local alignment of what would otherwise be considered 'unmappable' fragments allows for efficient processing of large datasets in a relatively short period of time. Because terminal stem-loops and internal single-stranded regions are common features of many ncRNAs, our method should prove useful for a wide variety of studies in RNA biology.

3.5 Materials and Methods

3.5.1 Total RNA-seq of pharate adult flies

Total RNA was extracted from pharate adult flies and treated with DNase I to remove DNA contamination. Ribosomal RNAs were removed from the samples using Ribo-Zero Human/Mouse/Rat kit (Epicentre). A TruSeq RNA Sample Preparation Kit v2 (Illumina) was used for barcoding for multiplexing and cDNA library preparation. The TruSeq procedure first fragments rRNA-depleted RNA samples and performs first strand synthesis using reverse transcriptase and random primers. The second strand synthesis uses DNA polymerase I and RNase H. The cDNA fragments then go through an end repair by adding a single adenosine at the ends. Adapters are ligated after repair. Paired end (2 48) sequencing was performed on an Illumina HiSeq 2000 platform. The data are deposited in Gene Expression Omnibus with accession number GSE50711 and named Fly_pharate_48nt (8 datasets: SRR347291-SRR347294). Data analysis is presented in the following sections.

3.5.2 Additional RNA-seq data used

Additional RNA-seq datasets used in this study are generated by our lab and other labs, and listed as follows, together with their SRA (Short Read Archive) accession numbers. Here we also briefly describe library preparation methods used to obtain these data, to help understand the Vicinal methodology. Fly_ovary_RIP_35nt: SRR120120-SRR120139 and SRR287104-SRR287107 (24 datasets) (17). Fly_S2_45nt: SRR345574-SRR345591 (18 datasets) (18). Fly_larva3_48nt: Garcia et al. unpublished (16 datasets). Fly_pupa_48nt: Garcia et al. unpublished (16 datasets). Mouse_ES_40nt: SRR392624-SRR392626 (3 datasets) (18). Mouse_ES_51nt: SRR915881-SRR915888 and SRR941123-SRR941140 (26 datasets) (19). Mouse_satellite_50nt: SRR953246 (1 dataset) (20). Human_HCT116_50nt: SRR901290-SRR901292 (3 datasets) (21). The fly_ovary_RIP_35nt libraries were described by our lab previously (Lu *et al.*, 2014). Briefly, Sm protein containing RNP complexes from *Drosophila* ovaries were immunoprecipitated using anti-Sm or anti GFP antibodies and the associated RNAs were purified. No polyA selection or rRNA removal was performed on the immunopurified RNA. First strand synthesis was carried out using SuperScript III kit (Life Technologies). Second strand synthesis was performed using E. coli DNA polymerase I and RNase H (Life Technologies). Double stranded cDNAs were made into libraries and sequenced using Illumina Genome Analyzer II. The fly_larva3_48nt, fly_pupa_48nt RNA-seq datasets were generated the same way as the fly_pharate_48nt described above. Shilatifard lab generated the fly_S2_45nt, mouse_ES_40nt, mouse_ES_51nt and human_HCT_116_50nt RNA-seq datasets (Smith *et al.*, 2011; Hu *et al.*, 2013b,a). Ribosomal

RNA was removed from two micrograms of DNase-treated total RNA using the Ribo-Zero kit from Epicentre, and libraries were made using the Tru-seq mRNA kit from Illumina. The generation of mouse_satellite_50nt dataset by the Sartorelli lab also followed a similar protocol (Mousavi *et al.*, 2013). All the datasets used in this study were generated from RNA containing ncRNAs.

3.5.3 Bioinformatic pipeline

To get a rough estimate of the sequencing coverage and length of ncRNAs, the RNA-seq reads were mapped to genome references using Bowtie, allowing a maximum of 2 mismatches (end-to-end mapping) (Langmead *et al.*, 2009; Langmead and Salzberg, 2012). Splicing was not considered in read mapping since only ncRNAs were investigated in this study. Then the same raw reads were also mapped to genome references using Bowtie2 in the –sensitive-local mode, which allows softclipping. The Bowtie and Bowtie2 mapped reads were used to make bedgraph files for visualization in genome browser. These bedgraph tracks were only shown in the analysis of snRNA:U1. In order to identify self-priming and ligation events, the Bowtie2 mappable reads were filtered (using samsoftfilter.py, see the software package and instructions therein) to select reads that are only partially mapped to the genome, leaving at least n nucleotides from either end that are not mappable (n >5). After filtering, the unmappable parts of the partially mappable reads were mapped again to the vicinity of the mappable parts (using Vicinal_1.0.py and Vicinal_2.0.py). This step generates a SAM file of chimeric reads and two wiggle files containing the coverage data for both the plus and minus strands. To make the method easy to use, we prepared a complete set of command line instructions. The scripts and instructions are available for download from the following website: https://sites.google.com/site/zhipeng0426/-programming.

3.5.4 Implementation of the Vicinal algorithm

Samsoftfilter.py parses the CIGAR information from the input SAM file to find reads that have at least one softclipped region longer than a defined value (default n >5). Shorter softclipped regions (n >5) could be sequencing errors or other kinds of chimeric sequences and therefore not considered in subsequent analysis. The softclipped reads were processed using Vicinal_1.0 and Vicinal_2.0.py scripts. The purpose of the Vicinal_1.0 and Vicinal_2.0.py script is to map the unmappable regions of the partially mapped reads to the vicinity of the mappable parts, on the opposite strand. We term this process 'vicinal mapping' to distinguish it from the Bowtie2 terminology of 'local mapping'. Vicinal_1.0.py stores an initialized dictionary for fast processing but is only appropriate for genomes with smaller chromosomes, e.g. fly and nematode, while Vicinal_2.0.py does not store an initialized dictionary and is slower and can be used for genomes with bigger chromosomes like mouse and human. In order to map the softclipped reads efficiently and minimize memory footprint, the reads were sorted by chromosomal position. Once the reads mapped to one chromosome are processed, the results are written to output files (file_prefix_chim.sam for chimeric reads, file_prefix_1.wig and file_prefix_2.wig for coverage). Prior to vicinal mapping, the CIGAR code from the softclipped SAM file is parsed, and accordingly each read is divided into 2 or 3 parts, S+M, M+S, or S+M+S, depending on whether there are softclipped fragments on the 5' and/or 3' end, where S represents softclipped, and M represents matched. Internal mismatches were ignored. A region around each mapped fragment (with a radius defined by the users, default is 100nt) was extracted from the reference genome on the opposite strand. Then the softclipped fragments were searched against the extracted region and a total of one match is allowed. Once a match is found, the record for that read is output to a SAM file, and the coordinates of the mapped fragments on both strands were calculated and output to the two wiggle files. The wiggle files can be further converted to smaller gzipped bedgragh files for efficient storage and transfer. See detailed instructions in the Vicinal software package: https://sites.google.com/site/zhipeng0426/programming.

3.5.5 ncRNA lists and generation of lists with chimeric read numbers.

Lists of fly, mouse and human ncRNA coordinates used in the analysis were generated as follows. The fly ncRNA list was downloaded from UCSC Genome Bioinformatics Site (http://hgdownload.soe.ucsc.edu/goldenPath/dm3/database/), matched with gene names from Flybase (http://flybase.org/static_pages/downloads/COORD.html) and rearranged according to the format described in the Vicinal software. The mouse ncRNA list was downloaded from MGI at Jackson Laboratory (ftp://ftp.informatics.jax.org/pub/reports/MGI_MRK_Coord.rpt). The human ncRNA list was downloaded from Ensembl (ftp://ftp.ensembl.org/pub/release-74/-fasta/homo_sapiens/ncrna/) and rearranged accordingly. The number of chimeric reads for each ncRNA can be obtained using the readnum.sh script, which depends on the samtools package (Li *et al.*, 2009).

3.5.6 Northern blotting

S2 cells were homogenized in TRIzol (Life Technologies) and total RNA was extracted following manufacturer's instructions. RNA was electrophoresed in 4-12% TBE-Urea polyacrylamide gels (Life Technologies), transferred to nylon membranes, and probed with 32P-labeled PCR products corresponding to the D. melanogaster U2 and LU snRNA cDNAs.

3.5.7 RNA secondary structure prediction

The secondary structures of the non-coding RNAs were predicted using either UNAfold or the Vienna RNA Package with default parameter settings (Markham and Zuker, 2008; Hofacker, 2003). Alternative secondary structures (conformers) were occasionally adjusted manually to fit the chimeric reads. Structured alignments of ncRNAs were performed using LocARNA (global standard alignment) (Will *et al.*, 2007). Secondary structures of the predicted RNAs were drawn using VARNA (Darty *et al.*, 2009).

Acknowledgements We thank the members of the Matera laboratory for numerous helpful comments and suggestions. We also thank Dr. Xiaojun Guan (Sequenom, San Diego, CA) and Liang Shan (UNC Department of Computer Science) for advice and help with computer programming. This work was supported by NIH grant R01-GM053034 (to A.G.M.).

Conflict of interest statement. None declared.

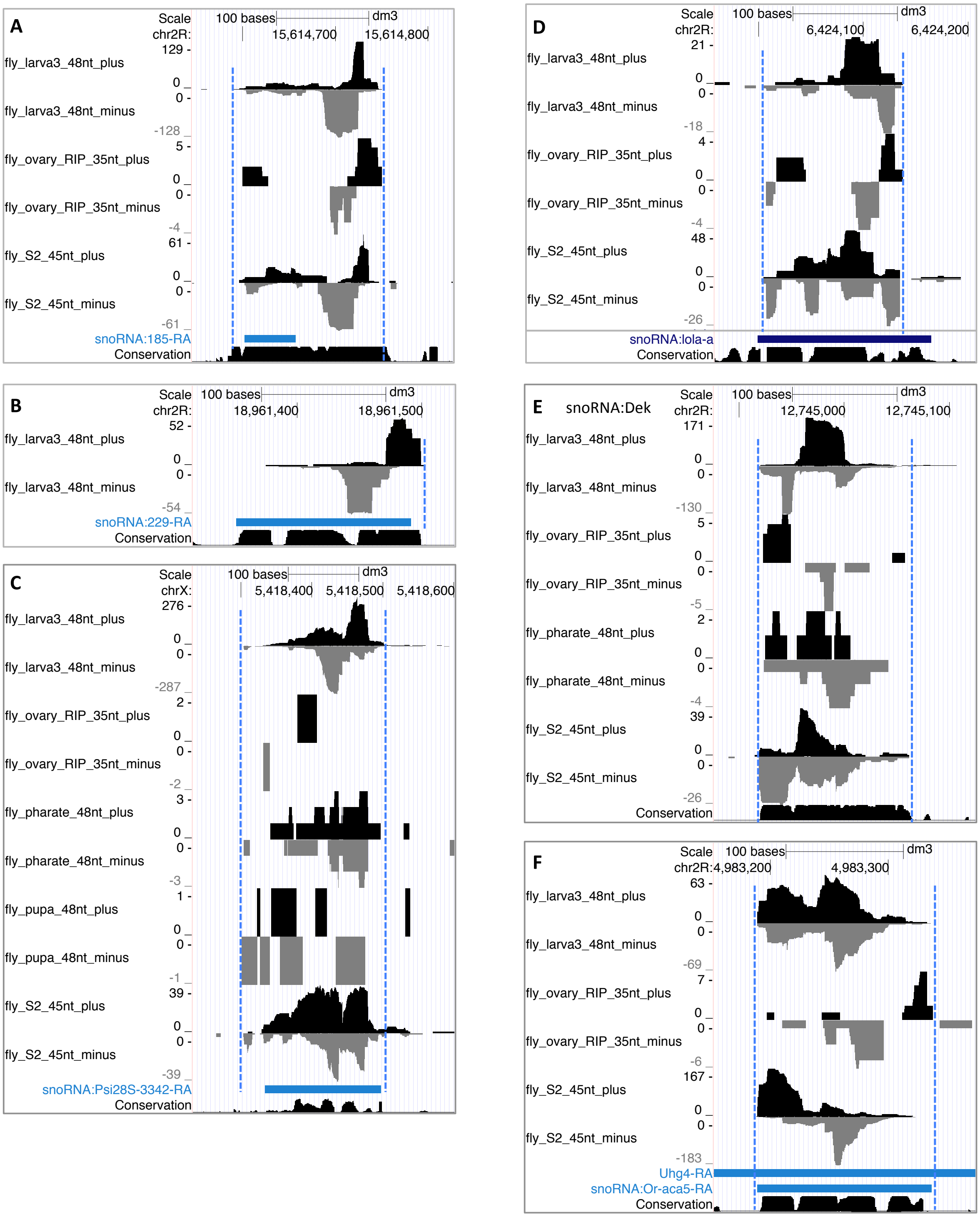


Figure 3.4: **Vicinal analysis of additional fly ncRNAs.** (continued on the next page.)

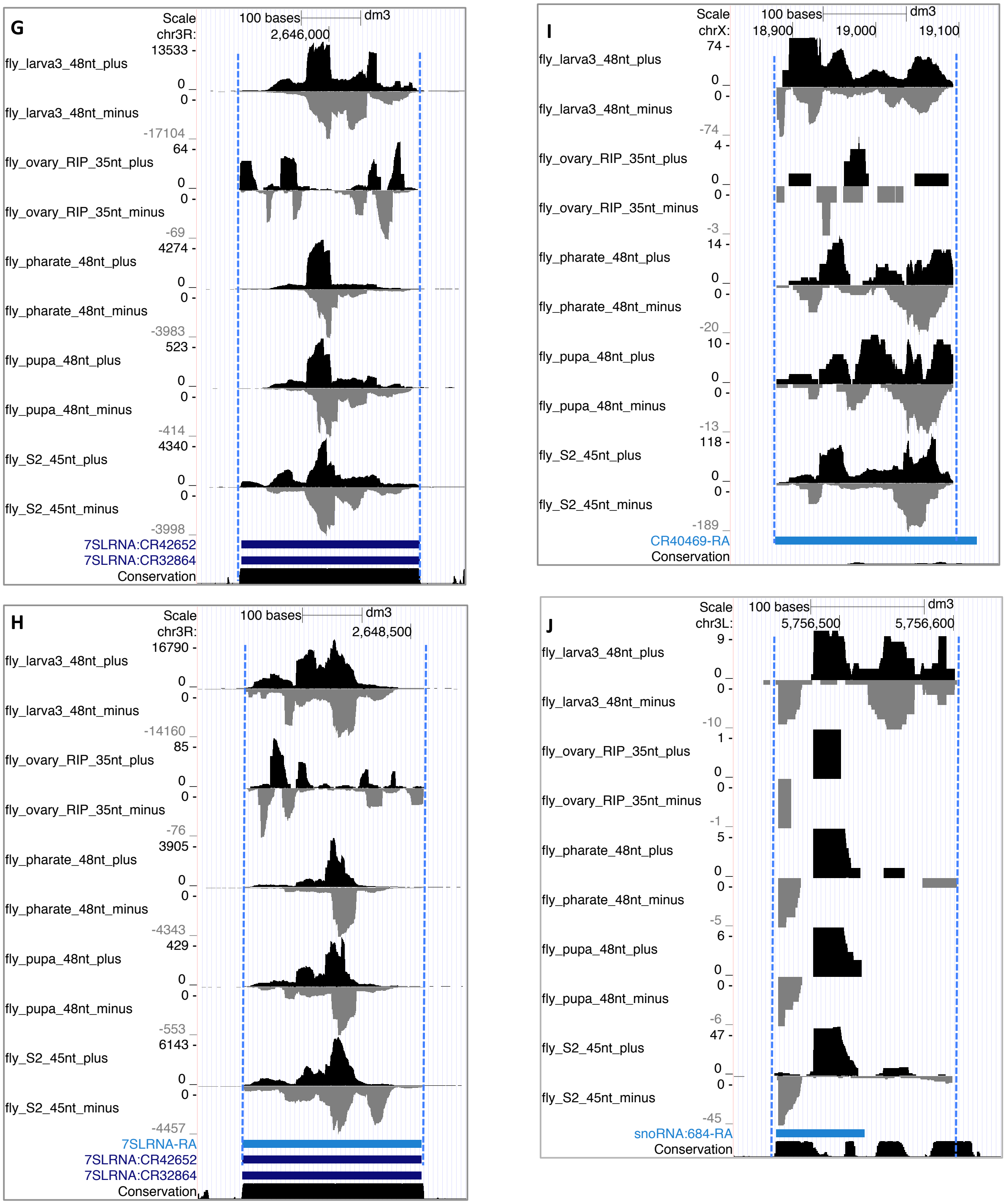


Figure 3.4: **Vicinal analysis of additional fly ncRNAs.** (continued on the next page.)

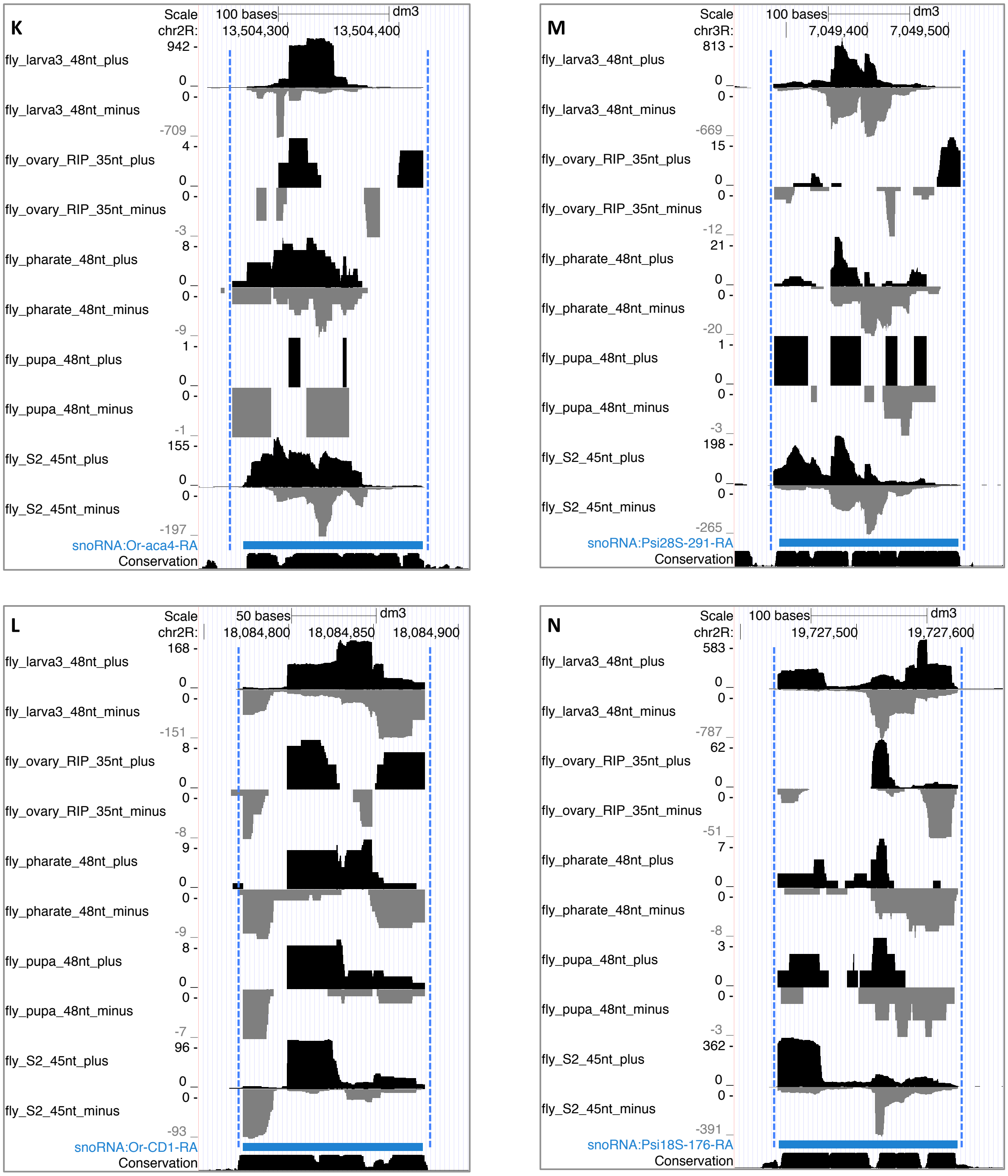


Figure 3.4: **Vicinal analysis of additional fly ncRNAs.** (continued on the next page.)

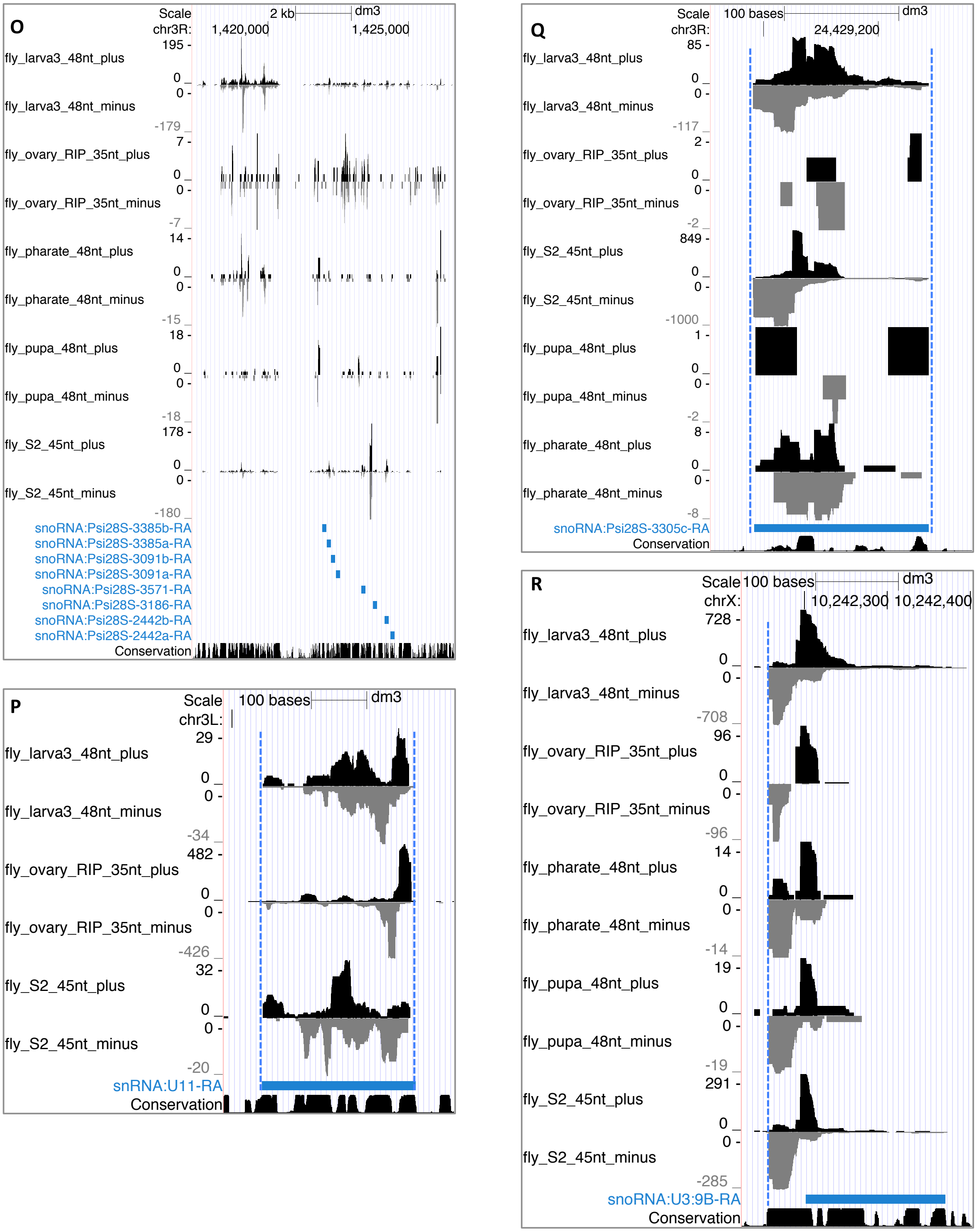


Figure 3.4: **Vicinal analysis of additional fly ncRNAs.** More examples of of fly ncRNAs that have chimeric reads are shown here. Genome browser tracks are shown for datasets that contain useful chimeric reads for these ncRNAs. Detailed analysis of the chimeric reads are not shown here. (O) shows a snoRNA cluster, where some of them have chimeric reads useful for end determination.

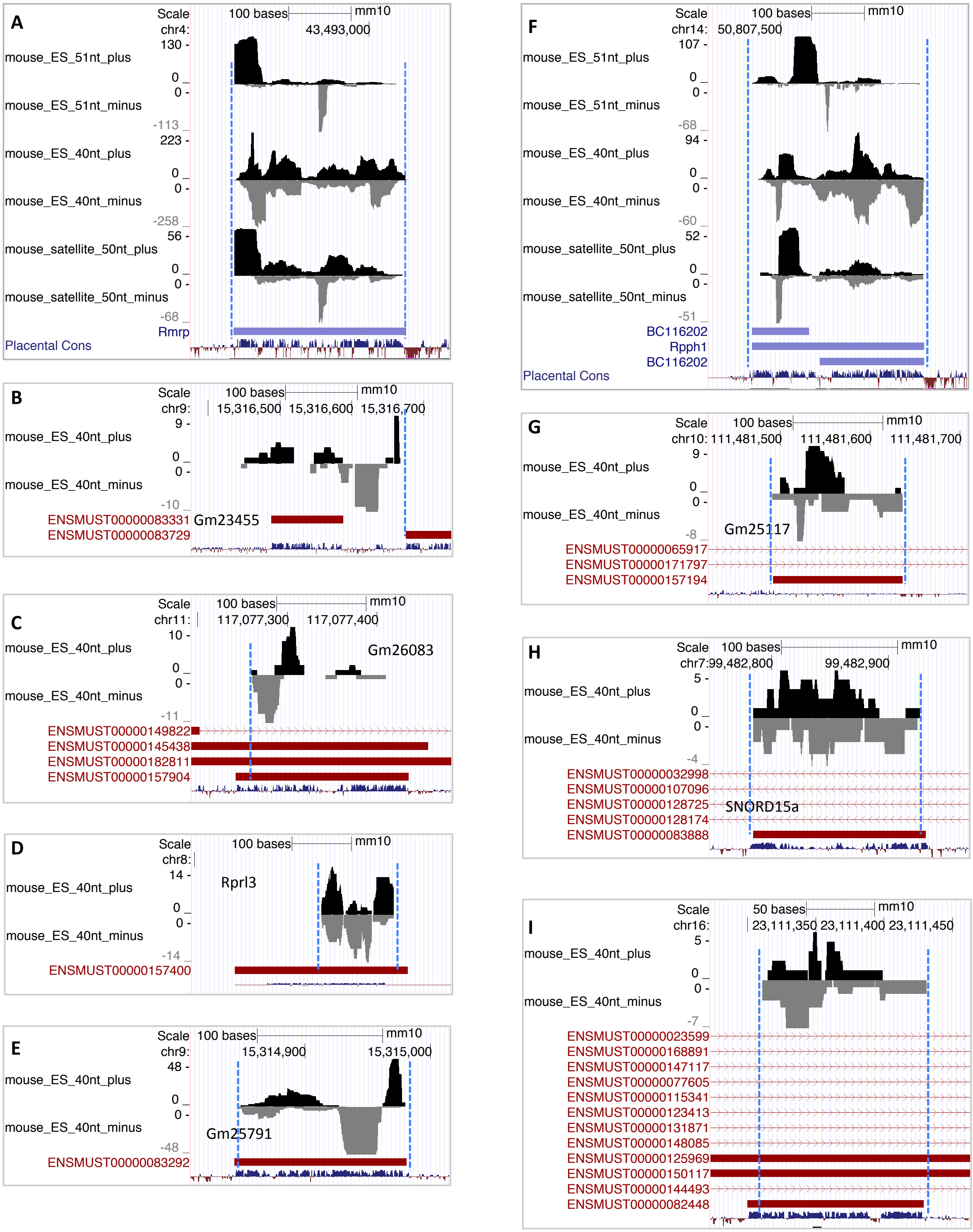


Figure 3.5: **Vicinal analysis of mouse and human ncRNAs.** (continued on the next page.)

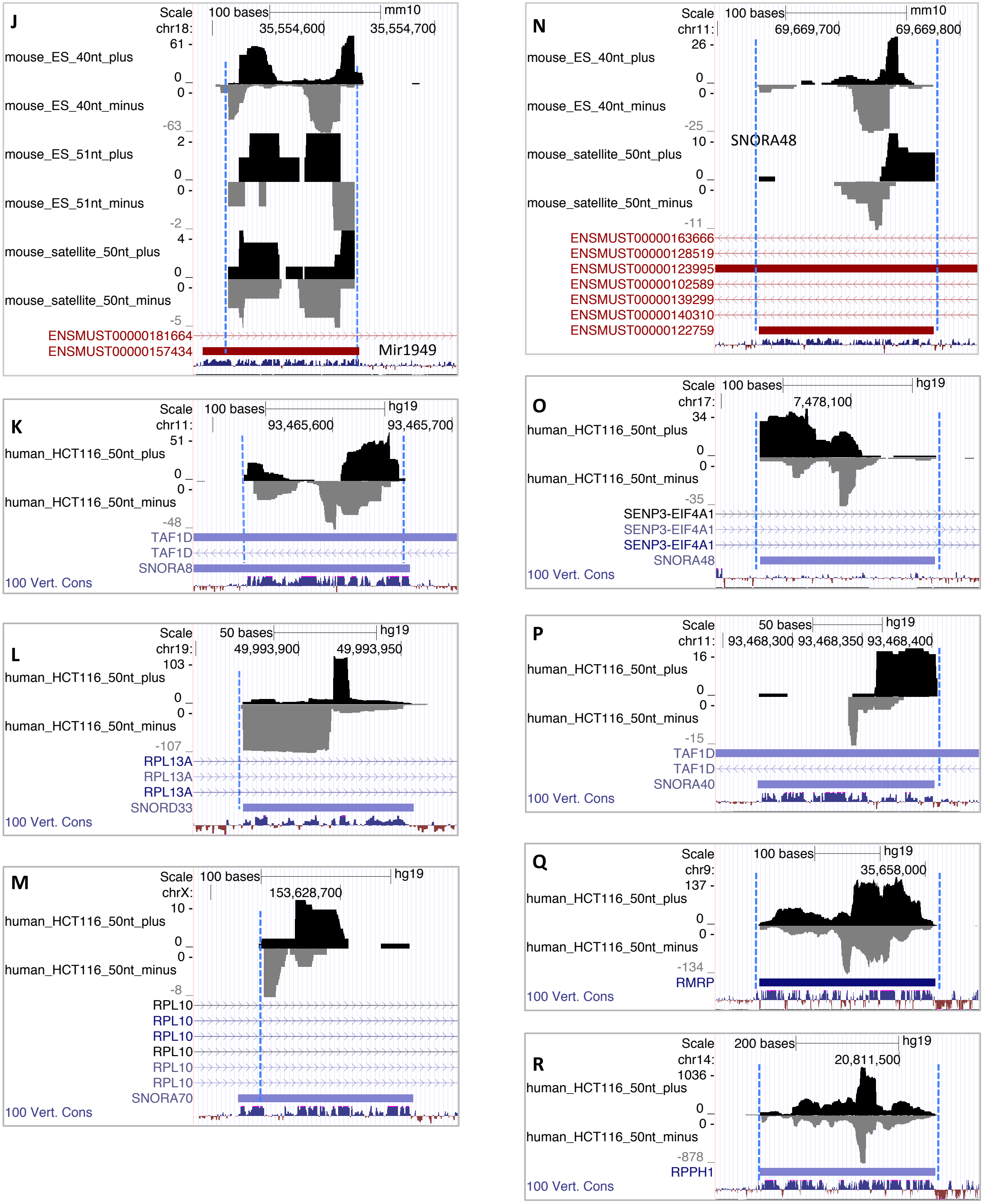


Figure 3.5: **Vicinal analysis of mouse and human ncRNAs** Examples of of mouse (A-J) and human (K-R) ncRNAs that have chimeric reads are shown here. Genome browser tracks are shown for datasets that contain useful chimeric reads for these ncRNAs. Detailed analysis of the chimeric reads are not shown here.

CHAPTER 4: Developmental switch of snRNA isoforms is conserved in evolution

Zhipeng Lu and A Gregory Matera

Note: This chapter is a manuscript currently in preparation.

4.1 Abstract

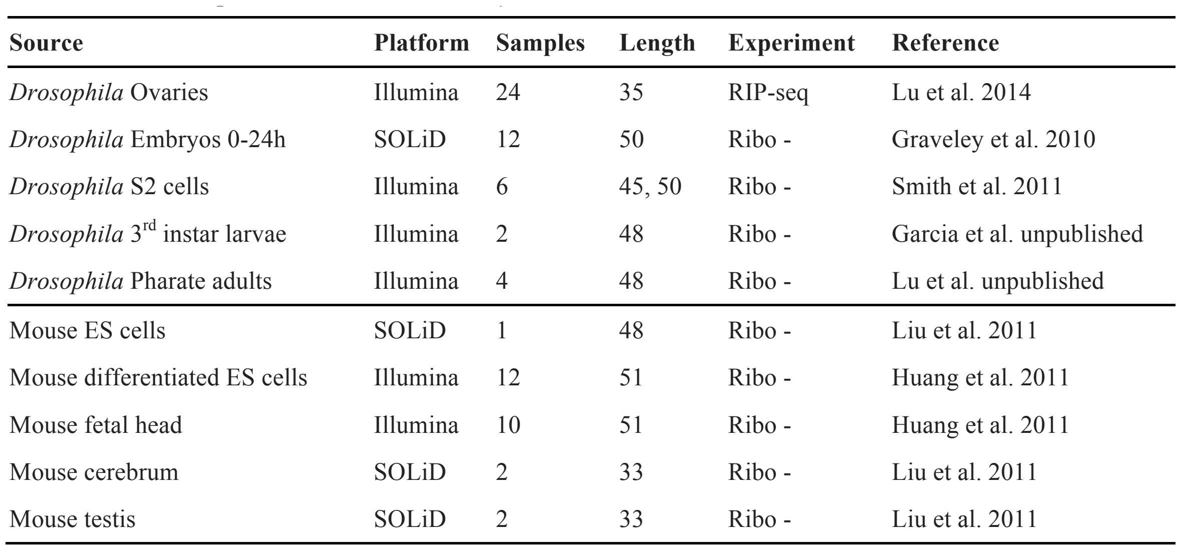
Pre-mRNA splicing is a critical step in eukaryotic gene expression that contributes to proteomic, cellular and developmental complexity. Small nuclear (sn)RNAs are core spliceosomal components, however, the extent to which differential expression of snRNA isoforms regulates splicing is completely unknown. This is partly due to difficulties in the accurate analysis of the spatial and temporal expression patterns of snRNAs. Here, we use high throughput RNA-sequencing (RNA-seq) data to precisely profile expression of four major snRNA isoforms throughout *Drosophila* development. This analysis shows that isoforms of each snRNA species have distinct expression patterns in the embryo, larva and pharate adult stages. Remarkably, expression of different isoforms is more heterogeneous during embryogenesis, and as development progresses, a single isoform from each snRNA species gradually dominates expression. Despite the lack of stable snRNA orthologous groups during evolution, we found that this developmental switching of snRNA isoforms is highly conserved in distantly related vertebrate species, such as *Xenopus*, mouse and human. Our results suggest that expression of snRNA isoforms is regulated, laying the foundation for functional studies of individual snRNA isoforms.

4.2 Introduction

Removal of introns from pre-mRNAs, a process called splicing, is an important step in the expression of eukaryotic genes. Splicing adds an important layer to the spatial and temporal regulation of gene expression, which is essential for the generation of diverse cell types from an identical genome (Chen and Manley, 2009). Splicing of most introns is catalyzed by the spliceosome, a macromolecular complex containing five small nuclear ribonucleoproteins (snRNPs) and numerous auxiliary proteins (Will and Luhrmann, 2011). Two types of spliceosomes coexist in most eukaryotic cells, the major U2-type, containing U1, U2, U4, U5 and U6 snRNPs, and the minor U12-type, containing U11, U12, U4atac, U5 and U6atac snRNPs. The U2-type splices more than 99% of all introns, whereas the U12-type splices less than 1% of all introns (Alioto, 2007).

The potential for snRNA paralogs to regulate splicing has been recognized since the early 1980s, following the discovery of heterogeneity in snRNA populations (Mattaj and Hamm, 1989). As a result, a number of studies have analyzed the expression of snRNA isoforms in a variety of organisms (Forbes *et al.*, 1984; Lund *et al.*, 1985, 1987; Lund and Dahlberg, 1987; Korf *et al.*, 1988; Lund, 1988; Lobo *et al.*, 1988; Nash *et al.*, 1989; Santiago and Marzluff, 1989; Lo and Mount, 1990; Stefanovic *et al.*, 1991; Hanley and Schuler, 1991; Sontheimer and Steitz, 1992; Sierra-Montes *et al.*, 2002, 2003; Pereira-Simon *et al.*, 2004; Sierra-Montes *et al.*, 2005; Chen *et al.*, 2005b; Smail *et al.*, 2006; Hinas *et al.*, 2006; Praveen *et al.*, 2012; Jia *et al.*, 2012; O'Reilly *et al.*, 2013). These studies showed that certain snRNA isoforms are differentially expressed during development or in various tissues. However, most of these studies used semi-quantitative methods, and could not distinguish near identical isoforms.

The contribution of snRNA variants to splicing regulation is unclear. First of all, due to the lack of a genetically tractable system, early studies were unable to demonstrate their biological relevance. Second, sequence analysis of snRNA paralogs across evolution suggests that all multi-copy snRNA genes have undergone concerted evolution, i.e. members of a given snRNA gene family are more similar within a species than between species (Pavelitz *et al.*, 1995, 1999; Mount *et al.*, 2007; Marz *et al.*, 2008). Because stable orthologous gene groups do not persist over evolutionary time (groups are usually only detectable within a genus), the possibility of significant functional divergence is in question. Surprisingly, a recent study showed that a 5nt deletion in a mouse U2 snRNA paralog caused neurodegeneration (Jia *et al.*, 2012). The mutated U2 gene is expressed primarily in the central nervous system and the mutation reportedly caused tissue-specific splicing defects (Jia *et al.*, 2012). This study suggests that the paralogs of an snRNA can quickly acquire tissue-specific expression patterns and essential functions during evolution.



Source	Platform	Samples	Length	Experiment	Reference
Drosophila Ovaries	Illumina	24	35	RIP-seq	Lu et al. 2014
Drosophila Embryos 0-24h	SOLiD	12	50	Ribo -	Graveley et al. 2010
Drosophila S2 cells	Illumina	6	45, 50	Ribo -	Smith et al. 2011
Drosophila 3^{rd} instar larvae	Illumina	2	48	Ribo -	Garcia et al. unpublished
Drosophila Pharate adults	Illumina	4	48	Ribo -	Lu et al. unpublished
Mouse ES cells	SOLiD	1	48	Ribo -	Liu et al. 2011
Mouse differentiated ES cells	Illumina	12	51	Ribo -	Huang et al. 2011
Mouse fetal head	Illumina	10	51	Ribo -	Huang et al. 2011
Mouse cerebrum	SOLiD	2	33	Ribo -	Liu et al. 2011
Mouse testis	SOLiD	2	33	Ribo -	Liu et al. 2011

Table 4.1: **RNA-seq datasets used in this study.** ‘Ribo -’ means the ribosomal RNAs are depleted from the samples.

Most vertebrate snRNAs exist in gene families consisting of dozens of nearly identical copies; therefore a reverse genetic approach to establish genotype-phenotype correlations for all the snRNA gene copies is not feasible. Compared to the vertebrates, *Drosophila* has a much smaller number of snRNA paralogs: five U1 genes, six U2, three U4, seven U5 and three U6. The other spliceosomal snRNAs are all expressed from single copy genes. The extensive genetic toolkit available for *Drosophila*, in addition to the reduced snRNA copy number, made it an ideal system for the analysis of multi-copy snRNA genes.

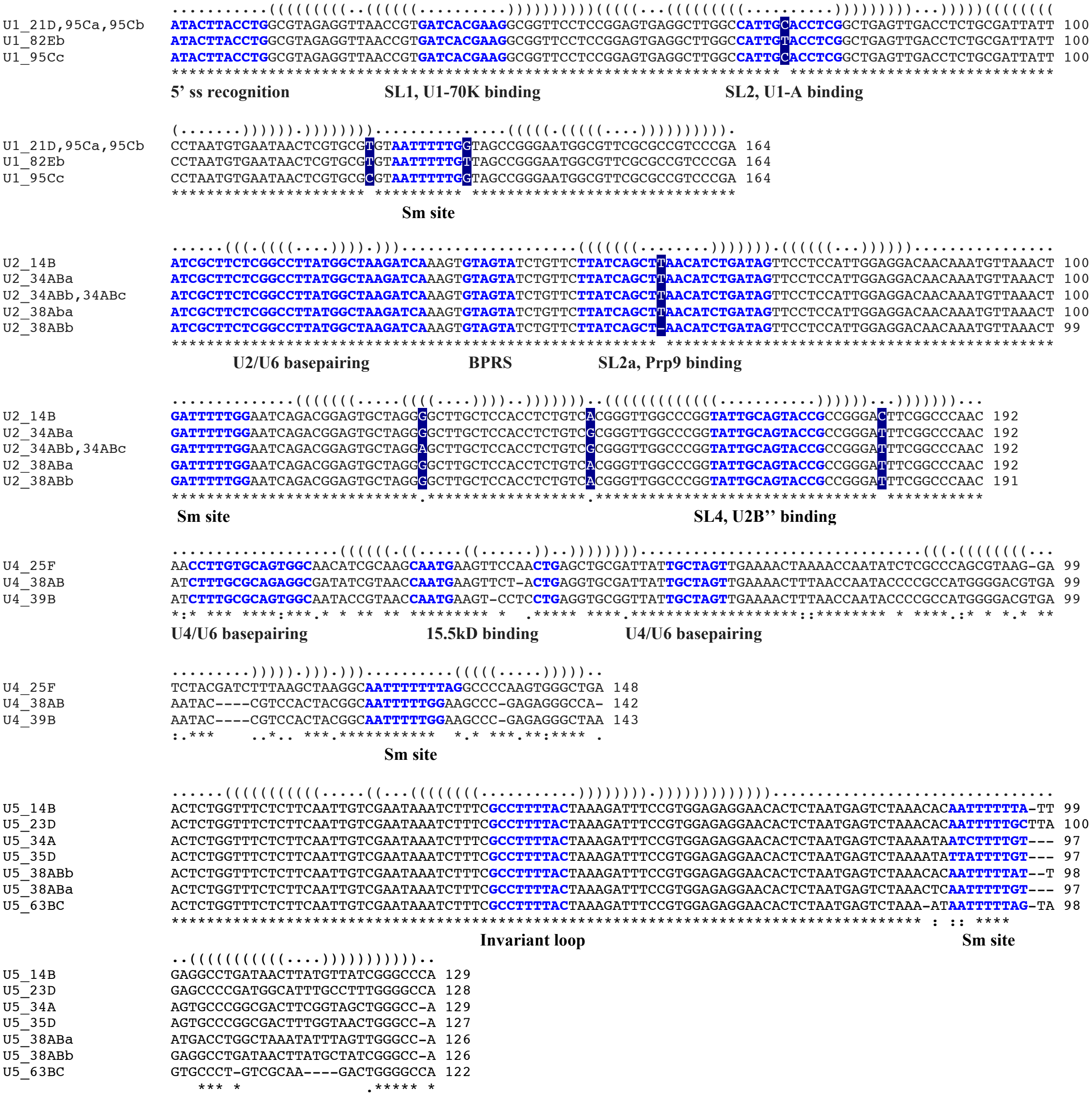
Massively-parallel transcriptome sequencing (RNA-seq) makes it possible to analyze transcripts with high accuracy and nucleotide resolution, therefore it is well suited for the analysis of highly similar snRNA paralogs. However, most RNA-seq datasets published thus far, including large-scale projects like modENCODE, are size selected to exclude abundant medium-sized (75 to 300 nt) non-coding (nc)RNAs, such as snRNAs. To analyze the expression of snRNAs, we identified available RNA-seq datasets that contain snRNAs, and carried out additional RNA-seq experiments on rRNA-depleted samples from *Drosophila* larvae and pharate adults. Using these datasets, we performed a comprehensive analysis of the expression of snRNA paralogs throughout *Drosophila* development, as well as from a few mouse tissues. We found that snRNA paralogs are differentially expressed in development. Surprisingly, the expression patterns are conserved in many other distantly related species, despite the lack of conservation in orthologous groups of snRNA genes. These data suggest that the developmental regulation of snRNA isoforms plays an important role in eukaryotic gene expression.

4.3 Results and Discussion

4.3.1 Generation and identification of appropriate RNA-seq datasets

To analyze the expression of *Drosophila* snRNA paralogs, we first collected published RNA-seq data that contain snRNAs (Table 4.1). In a previous study, we performed RNA-immunoprecipitation sequencing (RIP-seq) on *Drosophila* Sm proteins on ovarian lysates and these data were used quantify snRNA levels in ovaries (Lu *et al.*, 2014). SnRNAs not bound by Sm proteins are unstable; therefore the snRNAs recovered from Sm protein IPs accurately reflect the snRNA population (Sauterer *et al.*, 1988; Praveen *et al.*, 2012). Similarly, snRNA measurements from RNA-seq also reflect the number of functional snRNPs in vivo. The fruitfly ovary contains a mixture of somatic and germline cells. Because eggs provide most of the cellular material for early embryogenesis, for the purpose of developmental analysis, we consider the ovary as a developmental stage that is prior to the embryo. We searched public databases and found two additional RNA-seq datasets that contain snRNAs, and these data came from embryos and S2 cells (Graveley *et al.*, 2011; Smith *et al.*, 2011). S2 cells are derived from 20-24 hour late stage embryos (Schneider, 1972); therefore we compared them to late stage embryos in our subsequent analysis. In addition, we performed RNA-seq on rRNA-depleted total RNA samples from early 3rd instar larvae and pharate adults (Garcia *et al.*, 2013; Lu and Matera, 2014). In summary, our data collection covers major stages of Drosophila development: pre-embryo, embryo, larva and pharate adult (Table 4.1).

For evolutionary comparisons, we compiled RNA-seq data containing mouse snRNAs from several types of cells and tissues (Table 4.1), including embryonic stem (ES) cells, differentiated ES cells, fetal head, cerebrum and testis (Cui *et al.*, 2010; Yang *et al.*, 2011). Despite the



```
                    ...........((((((((..((((((..........))))))))))))))((((...(((((((((..........)))))))))...))))(((((((
U1_21D,95Ca,95Cb    ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_82Eb             ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGTACCTCGGCTGAGTTGACCTCTGCGATTATT 100
U1_95Cc             ATACTTACCTGGCGTAGAGGTTAACCGTGATCACGAAGGCGGTTCCTCCGGAGTGAGGCTTGGCCATTGCACCTCGGCTGAGTTGACCTCTGCGATTATT 100
                    ***************************************************************** **********************************
                    5' ss recognition        SL1, U1-70K binding                 SL2, U1-A binding

                    (.......)))))).))))))))...............(((((.(((((....))))))))))).
U1_21D,95Ca,95Cb    CCTAATGTGAATAACTCGTGCGTGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_82Eb             CCTAATGTGAATAACTCGTGCGTGTAATTTTTGTTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
U1_95Cc             CCTAATGTGAATAACTCGTGCGCGTAATTTTTGGTAGCCGGGAATGGCGTTCGCGCCGTCCCGA 164
                    ********************** ********** ******************************
                                 Sm site

                    ......(((.((((....)))).)))..................(((((((........))))))).((((((...))))))..................
U2_14B              ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_34ABa            ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_34ABb,34ABc      ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_38Aba            ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCTTAACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 100
U2_38ABb            ATCGCTTCTCGGCCTTATGGCTAAGATCAAAGTGTAGTATCTGTTCTTATCAGCT-AACATCTGATAGTTCCTCCATTGGAGGACAACAAATGTTAAACT 99
                    ******************************************************* ********************************************
                        U2/U6 basepairing        BPRS        SL2a, Prp9 binding

                    ..............(((((((.....((((....))))..)))))))..((((((((((((...........))))))...)))))))...
U2_14B              GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGACTTCGGCCCAAC 192
U2_34ABa            GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCGCGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_34ABb,34ABc      GATTTTTGGAATCAGACGGAGTGCTAGGAGCTTGCTCCACCTCTGTCGCGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_38ABa            GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 192
U2_38ABb            GATTTTTGGAATCAGACGGAGTGCTAGGGGCTTGCTCCACCTCTGTCACGGGTTGGCCCGGTATTGCAGTACCGCCGGGATTTCGGCCCAAC 191
                    ****************************.******************.******************************** ***********
                     Sm site                                                SL4, U2B'' binding

                    ...................((((((.((.....((......))..))))))))...............................(((.((((((((...
U4_25F              AACCTTGTGCAGTGGCAACATCGCAAGCAATGAAGTTCCAACTGAGCTGCGATTATTGCTAGTTGAAAACTAAAACCAATATCTCGCCCAGCGTAAG-GA 99
U4_38AB             ATCTTTGCGCAGAGGCGATATCGTAACCAATGAAGTTCT-ACTGAGGTGCGATTATTGCTAGTTGAAAACTTTAACCAATACCCCGCCATGGGGACGTGA 99
U4_39B              ATCTTTGCGCAGTGGCAATACCGTAACCAATGAAGT-CCTCCTGAGGTGCGGTTATTGCTAGTTGAAAACTTTAACCAATACCCCGCCATGGGGACGTGA 99
                    *:* *** ****:***.* * ** ** ********* *  .***** ****.*******************::******** * ****.:* * *.* **
                    U4/U6 basepairing         15.5kD binding        U4/U6 basepairing

                    .........))))).))).)))..........(((((.....)))))..
U4_25F              TCTACGATCTTTAAGCTAAGGCAATTTTTTTAGGCCCCAAGTGGGCTGA 148
U4_38AB             AATAC----CGTCCACTACGGCAATTTTTGGAAGCCC-GAGAGGGCCA- 142
U4_39B              AATAC----CGTCCACTACGGCAATTTTTGGAAGCCC-GAGAGGGCTAA 143
                    :.***    ..*.. ***.***********  *.* ***.**:**** .
                                   Sm site

                    ......((((((((((((.....((...((((((((...........))))))))).))))))))))))).............................
U5_14B              ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTTTA-TT 99
U5_23D              ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTTGCTTA 100
U5_34A              ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAAATAATCTTTTGT--- 97
U5_35D              ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAAATATTATTTTGT--- 97
U5_38ABb            ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACACAATTTTTAT--T 98
U5_38ABa            ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAACTCAATTTTTGT--- 97
U5_63BC             ACTCTGGTTTCTCTTCAATTGTCGAATAAATCTTTCGCCTTTTACTAAAGATTTCCGTGGAGAGGAACACTCTAATGAGTCTAAA-ATAATTTTTAG-TA 98
                    ************************************************************************************* : :: ****
                                                          Invariant loop                                     Sm site
                    ..((((((((((((....))))))))))))..
U5_14B              GAGGCCTGATAACTTATGTTATCGGGCCCA 129
U5_23D              GAGCCCCGATGGCATTTGCCTTTGGGGCCA 128
U5_34A              AGTGCCCGGCGACTTCGGTAGCTGGGCC-A 129
U5_35D              AGTGCCCGGCGACTTTGGTAACTGGGCC-A 127
U5_38ABa            ATGACCTGGCTAAATATTTAGTTGGGCC-A 126
U5_38ABb            GAGGCCTGATAACTTATGCTATCGGGCC-A 126
U5_63BC             GTGCCCT-GTCGCAA----GACTGGGGCCA 122
                       *** *              .***** *
```

Figure 4.1: **Alignment of *Drosophila* snRNA paralogs.** The secondary structure of each snRNA is presented on the top line of each alignment, using the dot-bracket notation. The U1 and U2 paralogs have very few variable nucleotide positions (three for U1 and four for U2) and they are highlighted with the black background and white lettering. Sequence elements that are important for base-pairing with other RNAs or interaction with proteins are indicated. U1:21D, U1:95Ca and U1:95Cb are identical. U2:34ABb and U2:34ABc are identical. 5' ss recognition: sequence recognizing pre-mRNA 5' splice site. BPRS: branch-point recognition sequence. SL1, SL2, SL2a and SL4: stem loops. U4 and U5 paralogs have significant differences among them and U5 paralogs are the most diverse. The 3' stem loop secondary structure of U5 isoforms is conserved, despite the divergence on the sequence level. Reads covering U4:25F (nucleotides 1-47), U4:38AB (1-46) and U4:39B (1-46) are unique among the three U4 paralogs. Reads covering U5:63BC (96-122) and the other six (97-end) are unique among all U5 paralogs. See Supplementary methods for details of read mapping.

fact that these samples are not derived from a single lineage, they represent the full range of development, from undifferentiated cells to terminally differentiated cells. These data can be used in comparison with the analysis of fruitfly snRNAs.

4.3.2 Structural and functional alignment of snRNA isoforms

The *Drosophila* genome encodes 27 spliceosomal snRNA genes that belong to 9 different snRNA species. The five major spliceosomal snRNAs are each expressed from multiple genes, and (with the exception of U6) have multiple nucleotide differences. Generic RNA-seq read mappers, e.g. Bowtie, randomly assign individual reads to multiple mappable locations in the genome, therefore, the measurements of snRNA isoform expression levels are not accurate. To overcome this problem, we aligned U1, U2, U4 and U5 snRNA paralogs and identified variable nucleotides and regions (Figure 4.1). Mouse and other vertebrate snRNA genes are not well characterized because they are present in multiple copies and many of them are pseudogenes (Denison *et al.*, 1981; Domitrovich and Kunkel, 2003). In order to analyze the expression of mouse snRNA isoforms, we retrieved known snRNA sequences and performed BLAST searches against genome sequence databases. Available mouse snRNA isoforms are aligned similar to their fly counterparts (Figure 4.2).

To help understand how differential expression of snRNA isoforms affects their functions *in vivo*, we superimposed the alignments of snRNA paralogs with the sequence elements known to be required for interaction with proteins and base pairing with other RNA molecules (Figures 4.1 and 4.2) (Madhani and Guthrie, 1992; Nagai *et al.*, 2001; Will and Luhrmann, 2011; Lin and Xu, 2012). Note that some of the nucleotide variations overlap with important sequence and structure motifs and are likely to affect the functions of these isoforms.

4.3.3 Developmental switching of *Drosophila* snRNA isoform dominance

To determine the relative expression of each snRNA isoform, we extracted unique sequencing reads mapped to variable regions based on the sequence alignments of fly and mouse snRNAs (Figures 4.1 and 4.2). For each RNA-seq experiment (e.g. a certain developmental stage or

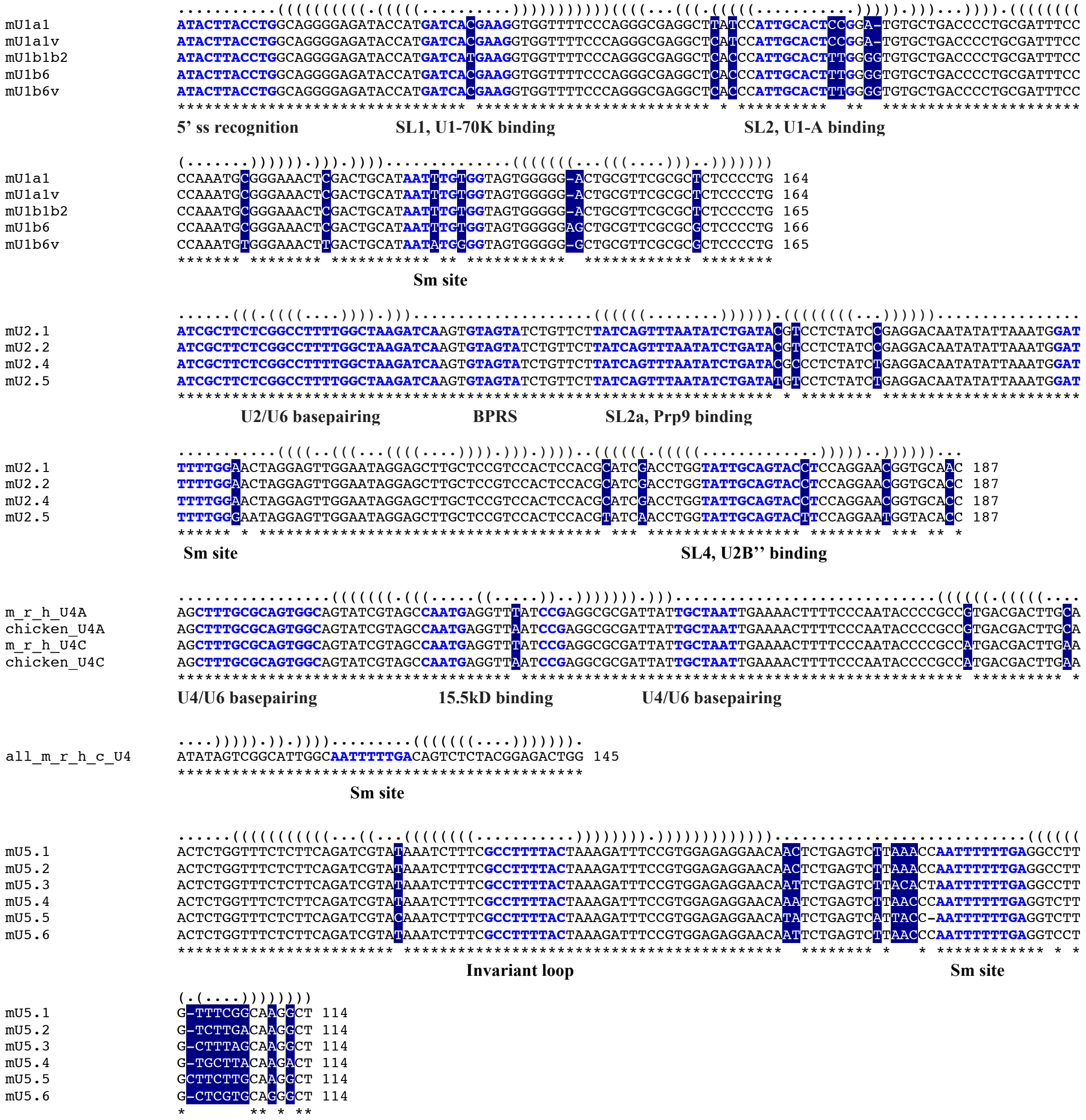


Figure 4.2: **Alignment of mouse snRNA paralogs.** Nucleotide variations are highlighted with the black background and white letters. See Fig. 1 for abbreviated motif names. Sequence elements that are important for base-pairing with other RNA species or interaction with proteins are indicated. Mouse (m), rat (r), chicken (c) and human (h) U4 snRNA paralogs are aligned together to show the two orthologous groups. Interestingly, even though the U4 snRNAs in several vertebrate species, human, mouse, rat and chicken, have only three nucleotide variations, they clearly segregate into two groups, based on the two variants in the second stem-loop. Similar to fly U5 snRNAs, mouse U5 paralogs are also the most diverse, and the variable region is confined to the 3’ end. See Supplementary methods for details of read mapping.

a cell/tissue type), we calculated the fraction of reads that each isoform uses in each snRNA group (see Supplemental Methods for details of mapping unique reads to snRNA isoforms). Surprisingly, this analysis showed that snRNAs that express multiple isoforms exhibit a developmental switch from expressing multiple isoforms during early stages to expressing a single dominant isoform in adults (Figure 4.3A and B)

U1 snRNA. Five U1 snRNA genes exist in *Drosophila*, and they express three isoforms, U1:21D/95Ca/95Cb, U1:82Eb and U1:95Cc. In all the stages analyzed, the U1:21D/95Ca/95Cb isoform is the dominant one, representing 70% to 98% of total U1 snRNA (Figure 4.3A). Expression of U1:21D/95Ca/95Cb gradually increases during development to almost 100% in adults, whereas U1:82Eb gradually decreases to barely detectable levels. A previous semi-quantitative analysis showed a similar expression pattern for these three isoforms during fly development (Lo and Mount, 1990).

U2 snRNA. The six *Drosophila* U2 snRNA genes express five distinct isoforms (Figure 4.1). The nucleotide variations allow us to analyze them in four groups because some of the variable regions are close to the snRNA ends and few reads are available to distinguish them. U2:34ABb/34ABc is the major isoform in the ovary, representing over 60% of total U2 (Figure 4.3A). Its expression dropped sharply in embryos to barely detectable levels later in development. In embryos, U2:14B and U2:38ABa are the dominant isoforms, representing over 60% of total U2. In later stages, U2:34ABb/34ABc gradually increased to more than 90% in pharate adults, becoming the dominant isoform. U2:38ABb is expressed only in the embryonic stages and barely detectable in ovary or after embryogenesis. S2 cells mainly express U2:34ABa, which is different from all the other samples. Overall, U2:14B/38ABa and U2:34ABb/34ABc display reciprocal expression trends in fly development. Their expression shows clear developmental switching, similar to U1 snRNAs.

U4 snRNA. Expression of all three *Drosophila* U4 isoforms can be measured accurately due to their divergence (Figures 4.1 and 4.3A). U4:25F is barely detectable in any of the stages analyzed, and it is likely a pseudogene or may only be expressed from a small number of cells. The expression levels of the other two isoforms, U4:38AB and U4:39B, are similar to each other in earlier stages, including the ovary and embryo. As development progresses, U4:39B gradually takes over the U4 population, generating more than 90% of the total U4 reads in pharate adults. Consistent with the fact that U4:39B is the major isoform expressed in flies, a P element insertion in U4:39B is lethal (Y. Wen and A.G. Matera, unpublished observations).

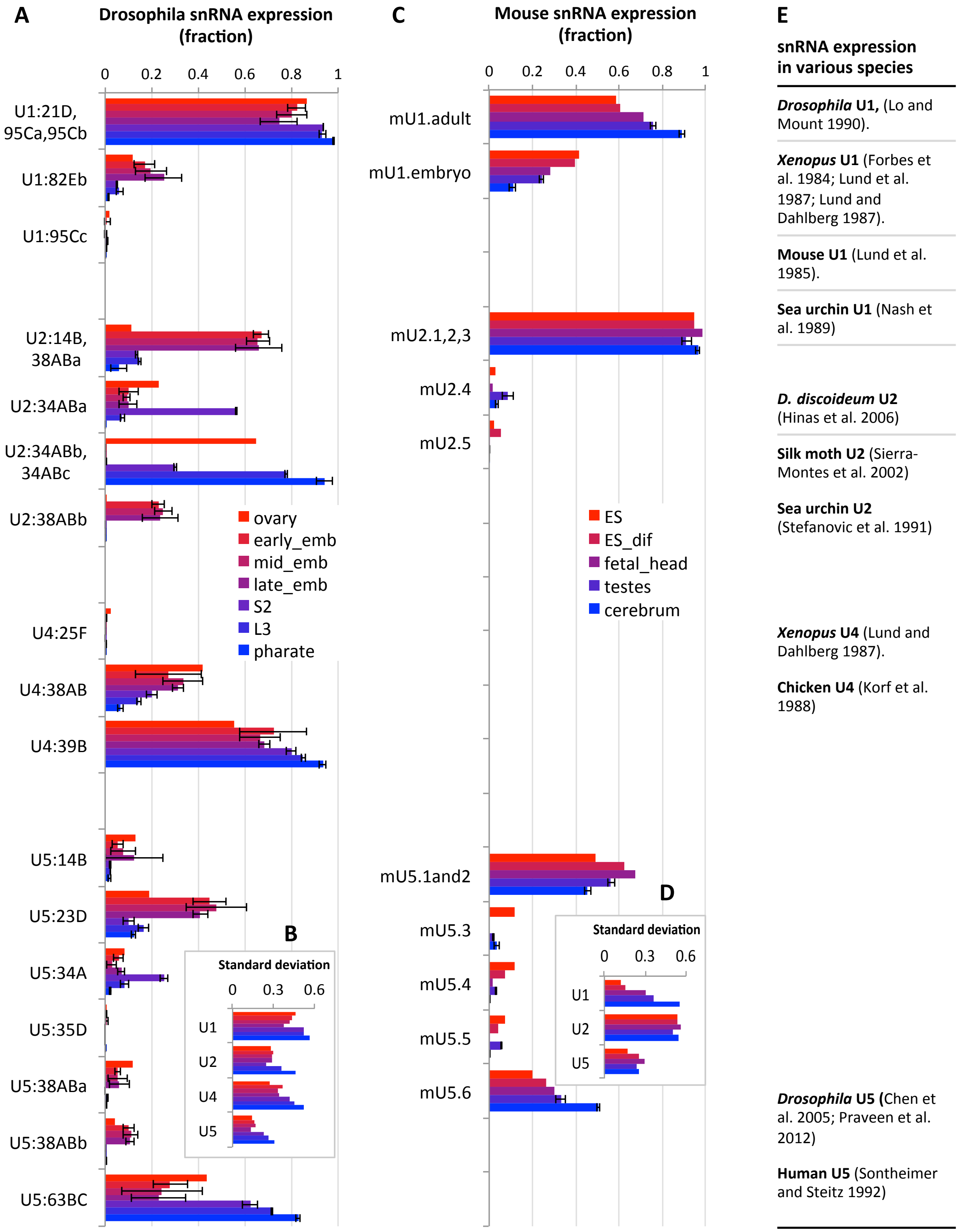
U5 snRNA. All seven *Drosophila* U5 snRNAs can be clearly distinguished from each other (Figure 4.1). Similar to U1, U2 and U4 snRNAs, our analysis showed a clear developmental switch in U5 isoform expression (Figure 4.3A). U5:14B, 34A, 35D, 38ABa and 38ABb are expressed at very low levels in all the stages analyzed. U5:23D and U5:63BC are expressed at very high levels and their expression follows a reciprocal pattern. Whereas U5:23D is the dominant isoform in embryos, accounting for more than 40% of total U5, its expression decreases to 15% after embryogenesis. U5:63BC accounts for roughly 30% of total U5 reads in embryos, and its expression increases dramatically during development to over 80% in pharate adults. These results are consistent with previous semi-quantitative analyses of U5 snRNAs (Chen *et al.*, 2005a; Praveen *et al.*, 2012).

The analysis of U1, U2, U4 and U5 snRNAs in fly RNA-seq data revealed clear developmental switching of isoform expression. We calculated the standard deviation of the fractional expression values for each group of snRNA isoforms at each developmental stage and the results are shown in Figure 4.3B. Thus, the overall trend of increasing standard deviations over developmental time provides further evidence that the dominance of one isoform in later stages of development is a common feature among different spliceosomal snRNA species.

4.3.4 Developmental switching of snRNA isoform dominance is conserved

To determine whether the consistent developmental switch in *Drosophila* snRNAs is conserved in evolution, we analyzed the expression profiles of mouse snRNAs. In addition, we compared our results to other species from published studies (Figure 4.3C, D and E).

U1 snRNA. Previous studies divided mouse U1 snRNAs into embryonic and adult isoforms, each of which are heterogeneous (Figure 4.2) (Lund *et al.*, 1985). Despite the lack of orthologous groups between mouse and fly, we observed a similar switch of isoforms during mouse development, consistent with previous reports (Figure 4.3C and E) (Lund *et al.*, 1985). Similar to flies, the standard deviation of fractions also showed an increasing trend (Figure 4.3D). The



snRNA expression in various species

***Drosophila* U1,** (Lo and Mount 1990).

***Xenopus* U1** (Forbes et al. 1984; Lund et al. 1987; Lund and Dahlberg 1987).

Mouse U1 (Lund et al. 1985).

Sea urchin U1 (Nash et al. 1989)

***D. discoideum* U2** (Hinas et al. 2006)

Silk moth U2 (Sierra-Montes et al. 2002)

Sea urchin U2 (Stefanovic et al. 1991)

***Xenopus* U4** (Lund and Dahlberg 1987).

Chicken U4 (Korf et al. 1988)

***Drosophila* U5 (**Chen et al. 2005; Praveen et al. 2012)

Human U5 (Sontheimer and Steitz 1992)

Figure 4.3: **Expression profiles of snRNA isoforms during development.** (See figure legend on the next page.)

Figure 4.3: **Expression profiles of snRNA isoforms during development.** (A and C) The fractional expression level for each snRNA paralog was calculated from reads mapping to the variable regions shown in Figures 4.1 and 4.2. The fractions for the paralogs of each snRNA species add up to 1 in each stage. (A) U2:14B and U2:38ABa are not identical, but they are lumped together due to an isufficiency in read numbers for embryos. (B and D) The standard deviation of the fractional expression values for each group of snRNA isoforms was calculated for each developmental stage. (E) Summary of previous studies on snRNA isoform expression patterns in various species.

same switching of isoform expression is recapitulated in mouse ES cell differentiation (Cheng *et al.*, 1997). Studies in *Xenopus* showed that different U1 snRNA isoforms are expressed in oocytes/embryos versus adults (Lund *et al.*, 1987; Lund and Dahlberg, 1987). Studies in sea urchins also showed that U1 isoforms are developmentally switched (Nash *et al.*, 1989). Taken together, these results reveal a conserved expression pattern that the major isoform of U1 snRNA is expressed throughout development, whereas the less abundant isoforms are primarily expressed in early embryos and switched off as development progresses.

U2 snRNA. Available mouse U2 snRNA reads only allow us to distinguish three groups (Figure 4.2). We found that mU2.1 and mU2.2 dominate U2 snRNA expression in all tissues analyzed (Figure 4.3C). The lack of developmental switching here is likely because we cannot distinguish all U2 isoforms. Nevertheless, previous analysis of U2 snRNAs in *D. discoideum* suggests that one group of U2 isoforms decreases dramatically relative to the other group during development (Hinas *et al.*, 2006). Analysis of U2 isoforms in silk moth showed a complex pattern of expression, with distinct isoforms dominating each stage (Sierra-Montes *et al.*, 2002). Studies in sea urchins showed that U2 isoforms are developmentally switched (Stefanovic *et al.*, 1991). Despite the complication of more isoforms for U2 snRNA, these results together with our analysis showed that more isoforms are expressed in earlier stages, and as development progresses, one isoform takes over the whole population.

U4 snRNA. We could not analyze the expression pattern of mouse U4 snRNAs due to the low number of mappable reads. But a very similar switching of U4 isoform expression has been shown during *Xenopus* and chicken development (Lund and Dahlberg, 1987; Korf *et al.*, 1988).

U5 snRNA. Similar expression switching of U5 isoforms is also observed in mice, although the change is not as dramatic (Figure 4.3C and D). mU5.6 and mU5.1and2 showed reciprocal expression patterns, whereas the other isoforms are expressed at very low levels. Previous analysis of human U5 snRNAs also revealed developmental isoform switching (Sontheimer and Steitz, 1992).

4.4 Conclusions

The consistent switching of snRNA isoforms suggests the functional importance in maintaining multiple genes for each snRNA species. snRNA paralogs may exert their regulatory effects on splicing through multiple mechanisms. The expression of isoforms is clearly under differential control in a tissue and/or developmental stage specific manner. Expression from different gene copies makes it possible to regulate production of specific snRNPs and therefore influence splicing. Supporting this notion, studies in mice that contain mutation in a single U2 gene copy showed that the mutation selectively affects brain function and the splicing of a specific set of mRNAs (Jia *et al.*, 2012). In addition, the different isoforms may form structurally distinct snRNPs with divergent functions. Consistent with this idea, Bach *et al.* reported that U1 snRNA isoforms have different affinities for U1 snRNP-specific proteins (Bach *et al.*, 1990). In conclusion, the comprehensive analysis of snRNA expression in a genetically tractable system provides essential information to guide future functional studies on snRNA isoforms *in vivo.*

4.5 Materials and Methods

4.5.1 RNA-seq data files

The following previously published RNA-seq data files used in this study were downloaded from modENCODE, NCBI and EMBL-EBI. Fly ovaries RIP-seq: GSE35842 (GSM876115 to GSM876134 and GSM1149490 to GSM1149493) (Lu *et al.*, 2014). Fly embryos: 12 datasets covering 0-2hr to 22-24hr embryo transcriptomes (modENCODE_4607 to modENCODE_4618) (Graveley *et al.*, 2011). S2 cells: GSE32120 (six datasets from control RNAi, SRR345578, SRR345579 and SRR345588-SRR345591) (Smith *et al.*, 2011). Fly L3 larvae: two datasets of wildtype early 3rd instar larvae (Garcia et al. unpublished). Fly pharate adults: GSE50711

(Lu *et al.*, 2014). Mouse ES cells: SRR407407 (Liu *et al.*, 2011). Mouse testis: SRR407405 and SRR407406 (Liu *et al.*, 2011). Mouse cerebrum: SRR018013 and SRR018014 (Liu *et al.*, 2011). Mouse fetal head: GSM566796-GSM566798, GSM566803-GSM566805, GSM566809-GSM566811 and GSM718983 (Huang *et al.*, 2011). Mouse CCE differentiated ES cells: GSM-566792-GSM566795, GSM566799-GSM566802, GSM566806-GSM566808 and GSM718982 (Huang *et al.*, 2011).

4.5.2 Conversion of formats

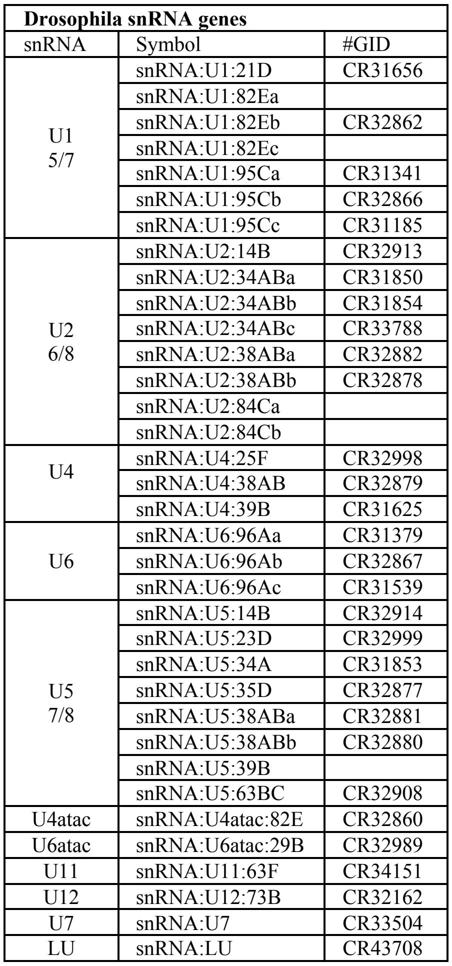
Conversion of scarf format to fastq format was performed using fq_all2std.pl with modifications on Phred score conversion, where fq_all2std.pl was originally from the MAQ package (Li *et al.*, 2008). Conversion of Phred encoding is performed using EMBOSS seqret (Rice *et al.*, 2000).

4.5.3 Extraction of uniquely mappable reads

Uniquely mappable reads were identified from U1, U4 and U5 snRNAs. Fractions of U2 paralogs were determined by a set of linear equations. Since the variance is bigger in the ovary RIP-seq datasets (due to the lengthy procedure and the use of variable conditions), all RIP-seq datasets were added up to calculate the fraction each snRNA paralog takes. The 12 embryo RNA-seq datasets were divided into three stages: early (0-8hr), mid (8-16hr) and late (16-24hr) (Graveley *et al.*, 2011). This is because we did not see significant variation in the fractions at each embryonic stage. The estimated time intervals are as follows. Ovary to early embryo: 1-2 days (later stage egg chambers contribute more to the total sequenced snRNAs). Early to middle embryo: 8 hours. Middle to late embryo: 8 hours. S2 cells were derived from late embryos (20-24 hours after egg laying), and therefore later than late embryos. Late embryo to third instar larva: 50 hours. Third instar larva to pharate adult: 140 hours. Note that the intervals among the developmental stages are not constant. Standard deviations were calculated for each stage shown in Figure 4.3. For Drosophila snRNAs: ovary, n=1; early_emb, n=4; mid_emb, n=4; late_emb, n=4; S2, n=6; L3, n=2; pharate, n=4. For mouse snRNAs: testes, n=2; cerebrum, n=2. See Supplementary methods for detailed mapping procedure along with instructions.

4.6 Supplementary methods: Assigning RNA-seq reads to snRNA isoforms

4.6.1 *Drosophila* snRNAs



Drosophila snRNA genes		
snRNA	Symbol	#GID
U1 5/7	snRNA:U1:21D	CR31656
	snRNA:U1:82Ea	
	snRNA:U1:82Eb	CR32862
	snRNA:U1:82Ec	
	snRNA:U1:95Ca	CR31341
	snRNA:U1:95Cb	CR32866
	snRNA:U1:95Cc	CR31185
U2 6/8	snRNA:U2:14B	CR32913
	snRNA:U2:34ABa	CR31850
	snRNA:U2:34ABb	CR31854
	snRNA:U2:34ABc	CR33788
	snRNA:U2:38ABa	CR32882
	snRNA:U2:38ABb	CR32878
	snRNA:U2:84Ca	
	snRNA:U2:84Cb	
U4	snRNA:U4:25F	CR32998
	snRNA:U4:38AB	CR32879
	snRNA:U4:39B	CR31625
U6	snRNA:U6:96Aa	CR31379
	snRNA:U6:96Ab	CR32867
	snRNA:U6:96Ac	CR31539
U5 7/8	snRNA:U5:14B	CR32914
	snRNA:U5:23D	CR32999
	snRNA:U5:34A	CR31853
	snRNA:U5:35D	CR32877
	snRNA:U5:38ABa	CR32881
	snRNA:U5:38ABb	CR32880
	snRNA:U5:39B	
	snRNA:U5:63BC	CR32908
U4atac	snRNA:U4atac:82E	CR32860
U6atac	snRNA:U6atac:29B	CR32989
U11	snRNA:U11:63F	CR34151
U12	snRNA:U12:73B	CR32162
U7	snRNA:U7	CR33504
LU	snRNA:LU	CR43708

Table 4.2: ***Drosophila* Sm class snRNAs.** The numbers under U1, U2 and U5 indicate numbers of real genes vs. genes reported by flybase. Some of the flybase annotations are incorrect.

There are a total of 29 genes for all the 11 *Drosophila* Sm class snRNAs (see the table below). Even though there are 33 named snRNAs, 5 of them are mistakes in annotation. U1, U2, U4 and U5 snRNA paralogs have nucleotide variations, which can be used to assign uniquely-mappable reads correctly. The three U6 snRNA paralogs are identical, and therefore it is not possible to analyze each separately. The default setting for Bowtie assign repetitive reads to mapped locations uniformly, and is thus not accurate in determining expression levels. Here we use the variable regions in the snRNA paralogs to reassign the reads. Since the patterns of the variations are different for different snRNAs, we used different strategies to extract reads mapped to these variable regions. In this manual, we only use uniquely mappable reads. You are free to explore the possibility of reassigning all reads mappable to snRNAs, based on the ratios inferred from uniquely mappable reads. However we don't think that it adds much value since it requires significant amount of unique reads for the determination of ratios reliable; and when there are large amounts of unique reads we do not care if the repetitive reads are to be added or not.

The input files are from bowtie mapping of RNA-seq reads, and the genome assembly is *D. melanogaster* Apr. 2006 (BDGP R5/dm3). The output of other mapping programs, including Bowtie2, and wrappers that use Bowtie/Bowtie2 as the engines, have variations in the format, and therefore some of the commands for subsequent analysis need to be modified accordingly. For example: the labeling of mismatches is different for them. The following analysis procedure is also useful for other purposes, such as analysis of differential expression of the paralogs in different tissues or development. Since the following analysis is designed for 35nt RNA-seq reads, modifications are needed for some of them to work optimally for reads of different lengths. Please familiarize yourself with the basics of command line interface before attempting to use these commands. Note: the samtools retrieving reads by location from indexed BAM files is very efficient, and therefore the job submission command 'bsub' for 'samtools view' is not so necessary in most occasions.

4.6.2 *Drosophila* Single copy snRNAs

Even though these snRNAs are all single copy genes (except that LU has an unexpressed pseudogene paralog, which takes away many reads from the expressed gene), we present the commands for retrieving RNA-seq reads mapped to them.

```
bsub samtools view -o Lu001.U4atac.sam Lu001_sorted.bam chr3R:1020726-1020885

bsub samtools view -o Lu001.U6atac.sam Lu001_sorted.bam chr2L:8389724-8389820

bsub samtools view -o Lu001.U7.sam Lu001_sorted.bam chr3L:3593823-3593893

bsub samtools view -o Lu001.U11.sam Lu001_sorted.bam chr3L:3893056-3893330

bsub samtools view -o Lu001.U12.sam Lu001_sorted.bam chr3L:16646869-16647106

bsub samtools view -o Lu001.LU.sam Lu001_sorted.bam chr2L:3046765-3046880

bsub samtools view -o Lu001.LUp.sam Lu001_sorted.bam chr2L:21644292-21644365

cat Lu001.LU.sam Lu001.LUp.sam > Lu001.LU.all.sam

awk ’$14 == "NM:i:0"’ Lu001.LU.all.sam > Lu001.LU.noMM.sam

rm Lu001.LU.sam Lu001.LUp.sam Lu001.LU.all.sam

wc -l Lu001.U4atac.sam Lu001.U6atac.sam Lu001.U7.sam

wc -l Lu001.U11.sam Lu001.U12.sam Lu001.LU.all.sam
```

4.6.3 *Drosophila* U1 snRNA

There are 5 U1 snRNAs in *Drosophila*: U1:21D, U1:82Eb, U1:95Ca, U1:95Cb and U1:95Cc. Three nucleotide variations at 70, 123 and 134 separate them into three different groups: U1:82Eb, U1:95Cc and U1:21D/U1:95Ca/U1:95Cb (see alignment of U1 paralogs below). The variable nucleotides at position 123 and 134 are close to each other and can be used to distinguish all three groups, therefore we searched for these fragments covering 123-134 in reads mapped to all U1 paralogs using grep or awk (grep is much faster than awk in general searching). Since the variable region used for analysis is only 12nt long, we cannot use the ‘samtools view coordinate’ method to retrieve reads covering this region. These 3 variant fragments are unique in U1 sequence and not anywhere else in the U1 gene. The following analysis of U1 snRNAs is not dependent on the size of the RNA-seq reads.

U1 paralogs	**strand**	**genomic locations**	**fragments used**
U1:21D	-	chr2L:901491-901654	TGTAATTTTTGG
U1:82Eb	-	chr3R:773655-773818	TGTAATTTTTGT
U1:95Ca	-	chr3R:19685189-19685352	TGTAATTTTTGG

U1:95Cb	+	chr3R:19653592-19653755	TGTAATTTTTGG
U1:95Cc	+	chr3R:19652056-19652219	CGTAATTTTTGG

Start from bowtie mapped bam files:

```
bsub samtools sort Lu001.bam Lu001_sorted bsub samtools index Lu001_sorted.bam
```

Extract reads mapped to all five paralogs (optional, compare it to uniquely mappable reads):

```
bsub samtools view -o Lu001.U1_21D.sam Lu001_sorted.bam chr2L:901491-901654

bsub samtools view -o Lu001.U1_82Eb.sam Lu001_sorted.bam chr3R:773655-773818

bsub samtools view -o Lu001.U1_95Ca.sam Lu001_sorted.bam chr3R:19685189-19685352

bsub samtools view -o Lu001.U1_95Cb.sam Lu001_sorted.bam chr3R:19653592-19653755

bsub samtools view -o Lu001.U1_95Cc.sam Lu001_sorted.bam chr3R:19652056-19652219
```

Merge paralogs and remove mismatches:

```
cat Lu001.U1_* > Lu001.U1.sam awk ’$14 == "NM:i:0"’ Lu001.U1.sam > Lu001.U1.noMM.sam

rm Lu001.U1_* Lu001.U1.sam
```

Extract reads covering the variable region by grepping:

```
grep CCAAAAATTACA Lu001.U1.noMM.sam > Lu001.U1_21D.U1_95Ca.U1_95Cb.sam

grep TGTAATTTTTGG Lu001.U1.noMM.sam >> Lu001.U1_21D.U1_95Ca.U1_95Cb.sam

grep ACAAAAATTACA Lu001.U1.noMM.sam > Lu001.U1_82Eb.sam

grep TGTAATTTTTGT Lu001.U1.noMM.sam >> Lu001.U1_82Eb.sam

grep CCAAAAATTACG Lu001.U1.noMM.sam > Lu001.U1_95Cc.sam

grep CGTAATTTTTGG Lu001.U1.noMM.sam >> Lu001.U1_95Cc.sam
```

Using bsub to submit these grep jobs (Note: grep with bsub produces a header of 30 lines that include the job description and log, make sure subtract these when counting the lines of output)

```
bsub -o Lu001.U1_21D.U1_95Ca.U1_95Cb1.sam grep CCAAAAATTACA Lu001.U1.noMM.sam

bsub -o Lu001.U1_21D.U1_95Ca.U1_95Cb2.sam grep TGTAATTTTTGG Lu001.U1.noMM.sam

bsub -o Lu001.U1_82Eb1.sam grep ACAAAAATTACA Lu001.U1.noMM.sam

bsub -o Lu001.U1_82Eb2.sam grep TGTAATTTTTGT Lu001.U1.noMM.sam

bsub -o Lu001.U1_95Cc1.sam grep CCAAAAATTACG Lu001.U1.noMM.sam
bsub -o Lu001.U1_95Cc2.sam grep CGTAATTTTTGG Lu001.U1.noMM.sam
```

4.6.4 *Drosophila* U2 snRNA

Drosophila has 6 U2 paralogs: U2:14B, U2:34Aba, U2:34ABb, U2:34ABc, U2:38Aba and U2:38ABb. A total of 4 nucleotide variations separate them into 5 groups (see alignment of U2 paralogs). However these mismatches are scattered around the whole transcript, therefore we cannot use a single region to distinguish all of them.

In order to assign all U2 snRNA reads to the paralogs, we first determine the fraction each paralog takes, based on the 4 nucleotide variations. Assuming each paralog taking a fraction: *a*, *b*, *c*, *d*, *e* and *f*, we can establish the following system of 6 linear equations and solve each fraction:

Equations:

$$a + b + c + d + e + f = 1 \qquad \text{(adds up to 1)} \tag{4.1}$$

$$c = d \qquad \text{(34ABb and 34ABc are identical)} \tag{4.2}$$

$$f/(a + b + c + d + e) = r \qquad \text{(measured ratios at variation 1)} \tag{4.3}$$

$$(a + e + f) : b : (c + d) = x : y : z \qquad \text{(measured ratios at variation 2 and 3)} \tag{4.4}$$

$$a/(b + c + d + e + f) = s \qquad \text{(measured ratios at variation 4)} \tag{4.5}$$

Solutions:

$$a = s/(s + 1) \tag{4.6}$$

$$b = y/(x + y + z) \tag{4.7}$$

$$c = d = z/2(x + y + z) \tag{4.8}$$

$$e = x/(x + y + z) - s/(s + 1) - r/(r + 1) \tag{4.9}$$

$$f = r/(r + 1) \tag{4.10}$$

In case there are not enough reads mapped to the last variant nucleotide (e.g. in the embryonic stages RNA-seq using the SOLiD platform, modENCODE), we have to lump U2:14B and U3:38ABa (a and e) together, and treat them as equal. In fact, theses two isoforms are not very different in expression levels, as can be seen in the data. In this case, the solutions become:

$$a = e = (x/(x + y + z) - r/(r + 1))/2 \tag{4.11}$$

$$b = y/(x + y + z) \tag{4.12}$$

$$c = d = z/2(x + y + z) \tag{4.13}$$

$$f = r/(r + 1) \tag{4.14}$$

These are the commands used to extract reads covering these three regions and calculate the ratios to be used for determining the distribution of all reads (not just uniquely mappable reads).

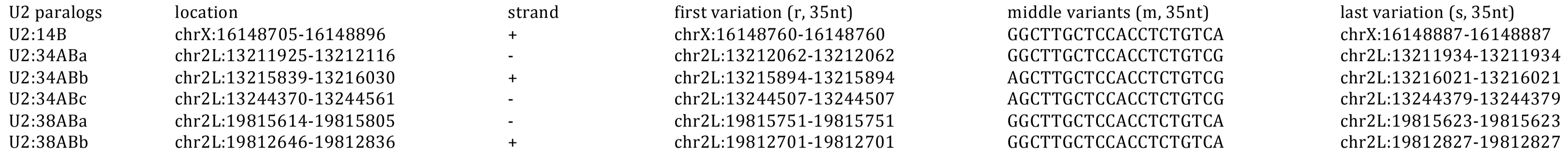
Since the distances among the 4 variable nucleotides are not all very big (73, 19 and 33) (Figure 4.4), we have to determine the intervals that can be used to determine each ratio used in the equations. The 1st and last variants can be retrieved using ‘samtools view coordinate’, whereas the middle variants have to be retrieved using grep. Read lengths affect the variant regions used for read assignment to the 1st and last regions. Solutions to this problem for 35 nt reads are presented below:

Commands used to retrieve reads mapped to all U2 paralogs

```
bsub samtools view -o Lu001.U2_14B.sam Lu001_sorted.bam chrX:16148705-16148896

bsub samtools view -o Lu001.U2_34ABa.sam Lu001_sorted.bam chr2L:13211925-13212116

bsub samtools view -o Lu001.U2_34ABb.sam Lu001_sorted.bam chr2L:13215839-13216030

bsub samtools view -o Lu001.U2_34ABc.sam Lu001_sorted.bam chr2L:13244370-13244561

bsub samtools view -o Lu001.U2_38ABa.sam Lu001_sorted.bam chr2L:19815614-19815805

bsub samtools view -o Lu001.U2_38ABb.sam Lu001_sorted.bam chr2L:19812646-19812836
```

Merge paralogs and remove mismatches:

```
cat Lu001.U2_* >  Lu001.U2.sam

awk ’$14 == "NM:i:0"’ Lu001.U2.sam > Lu001.U2.noMM.sam rm Lu001.U2_* Lu001.U2.sam
```



U2 paralogs	location	strand	first variation (r, 35nt)	middle variants (m, 35nt)	last variation (s, 35nt)
U2:14B	chrX:16148705-16148896	+	chrX:16148760-16148760	GGCTTGCTCCACCTCTGTCA	chrX:16148887-16148887
U2:34ABa	chr2L:13211925-13212116	-	chr2L:13212062-13212062	GGCTTGCTCCACCTCTGTCG	chr2L:13211934-13211934
U2:34ABb	chr2L:13215839-13216030	+	chr2L:13215894-13215894	AGCTTGCTCCACCTCTGTCG	chr2L:13216021-13216021
U2:34ABc	chr2L:13244370-13244561	-	chr2L:13244507-13244507	AGCTTGCTCCACCTCTGTCG	chr2L:13244379-13244379
U2:38ABa	chr2L:19815614-19815805	-	chr2L:19815751-19815751	GGCTTGCTCCACCTCTGTCA	chr2L:19815623-19815623
U2:38ABb	chr2L:19812646-19812836	+	chr2L:19812701-19812701	GGCTTGCTCCACCTCTGTCA	chr2L:19812827-19812827

Figure 4.4: ***Drosophila*** **U2 snRNA variations.**

Commands used to obtain the ratio r (by taking reads overlapping this nucleotide):

```
bsub samtools view -o Lu001.U2_r14B.sam Lu001_sorted.bam chrX:16148760-16148760

bsub samtools view -o Lu001.U2_r34ABa.sam Lu001_sorted.bam chr2L:13212062-13212062

bsub samtools view -o Lu001.U2_r34ABb.sam Lu001_sorted.bam chr2L:13215894-13215894

bsub samtools view -o Lu001.U2_r34ABc.sam Lu001_sorted.bam chr2L:13244507-13244507

bsub samtools view -o Lu001.U2_r38ABa.sam Lu001_sorted.bam chr2L:19815751-19815751

bsub samtools view -o Lu001.U2_r38ABb.sam Lu001_sorted.bam chr2L:19812701-19812701
```

Remove mismatches

```
awk ’$14 == "NM:i:0"’ Lu001.U2_r14B.sam > Lu001.U2_r14B.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U2_r34ABa.sam > Lu001.U2_r34ABa.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U2_r34ABb.sam > Lu001.U2_r34ABb.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U2_r34ABc.sam > Lu001.U2_r34ABc.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U2_r38ABa.sam > Lu001.U2_r38ABa.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U2_r38ABb.sam > Lu001.U2_r38ABb.noMM.sam
```

Remove intermediate files

```
rm Lu001.U2_r14B.sam Lu001.U2_r34ABa.sam Lu001.U2_r34ABb.sam

rm Lu001.U2_r34ABc.sam Lu001.U2_r38ABa.sam Lu001.U2_r38ABb.sam wc -l *U2_r*
```

Commands used to obtain values x, y and z (by grepping the 20nt regions):

```
bsub samtools view -o Lu001.U2_14B.sam Lu001_sorted.bam chrX:16148705-16148896
bsub samtools view -o Lu001.U2_34ABa.sam Lu001_sorted.bam chr2L:13211925-13212116
bsub samtools view -o Lu001.U2_34ABb.sam Lu001_sorted.bam chr2L:13215839-13216030
bsub samtools view -o Lu001.U2_34ABc.sam Lu001_sorted.bam chr2L:13244370-13244561
bsub samtools view -o Lu001.U2_38ABa.sam Lu001_sorted.bam chr2L:19815614-19815805
bsub samtools view -o Lu001.U2_38ABb.sam Lu001_sorted.bam chr2L:19812646-19812836
cat Lu001.U2_14B.sam Lu001.U2_34ABa.sam Lu001.U2_34ABb.sam Lu001.U2_34ABc.sam
    Lu001.U2_38ABa.sam Lu001.U2_38ABb.sam > Lu001.U2.sam
awk ’$14 == "NM:i:0"’ Lu001.U2.sam > Lu001.U2.noMM.sam
rm Lu001.U2_14B.sam Lu001.U2_34ABa.sam Lu001.U2_34ABb.sam
rm Lu001.U2_34ABc.sam Lu001.U2_38ABa.sam Lu001.U2_38ABb.sam Lu001.U2.sam

grep GGCTTGCTCCACCTCTGTCA Lu001.U2.noMM.sam > Lu001.U2_x.sam

grep TGACAGAGGTGGAGCAAGCC Lu001.U2.noMM.sam > Lu001.U2_x.sam
grep GGCTTGCTCCACCTCTGTCG Lu001.U2.noMM.sam > Lu001.U2_y.sam
grep CGACAGAGGTGGAGCAAGCC Lu001.U2.noMM.sam > Lu001.U2_y.sam
grep AGCTTGCTCCACCTCTGTCG Lu001.U2.noMM.sam > Lu001.U2_z.sam
grep CGACAGAGGTGGAGCAAGCT Lu001.U2.noMM.sam > Lu001.U2_z.sam
```

Commands used to obtain ratios s (by taking reads overlapping this nucleotide, sparing the neighbor variants):

```
bsub samtools view -o Lu001.U2_s14B.sam Lu001_sorted.bam chrX:16148887-16148887
bsub samtools view -o Lu001.U2_s34ABa.sam Lu001_sorted.bam chr2L:13211934-13211934
bsub samtools view -o Lu001.U2_s34ABb.sam Lu001_sorted.bam chr2L:13216021-13216021
bsub samtools view -o Lu001.U2_s34ABc.sam Lu001_sorted.bam chr2L:13244379-13244379
bsub samtools view -o Lu001.U2_s38ABa.sam Lu001_sorted.bam chr2L:19815623-19815623
bsub samtools view -o Lu001.U2_s38ABb.sam Lu001_sorted.bam chr2L:19812827-19812827
```

Remove mismatches

```
awk ’$14 == "NM:i:0"’ Lu001.U2_s14B.sam > Lu001.U2_s14B.noMM.sam
awk ’$14 == "NM:i:0"’ Lu001.U2_s34ABa.sam > Lu001.U2_s34ABa.noMM.sam
awk ’$14 == "NM:i:0"’ Lu001.U2_s34ABb.sam > Lu001.U2_s34ABb.noMM.sam
awk ’$14 == "NM:i:0"’ Lu001.U2_s34ABc.sam > Lu001.U2_s34ABc.noMM.sam
awk ’$14 == "NM:i:0"’ Lu001.U2_s38ABa.sam > Lu001.U2_s38ABa.noMM.sam
awk ’$14 == "NM:i:0"’ Lu001.U2_s38ABb.sam > Lu001.U2_s38ABb.noMM.sam
```

Remove intermediate files

```
rm Lu001.U2_s14B.sam Lu001.U2_s34ABa.sam Lu001.U2_s34ABb.sam
    Lu001.U2_s34ABc.sam Lu001.U2_s38ABa.sam Lu001.U2_s38ABb.sam
```

4.6.5 *Drosophila* U4 snRNA

Drosophila has 3 U4 paralogs: U4:25F, U4:38AB and U4:39B. Of these 3 paralogs, U4:25F has no fragments longer than 34 that are identical to the other 2 paralogs. The first 51 nucleotides are different among all of them and thus can be used to separate the 3 paralogs. To reassign the reads, first obtain the reads mapped to these three locations, then remove reads with mismatches, finally take reads that overlap the first 51 nucleotides. Read lengths do not affect the commands used for U4 snRNA paralogs. U4 snRNA genomic coordinates of the variable regions are as follows:

U4 snRNAs	**genomic coordinates**	**strand**	**variable region (35nt)**
U4:25F	chr2L:5565619-5565766	+	chr2L:5565619-5565665
U4:38AB	chr2L:19810734-19810875	+	chr2L:19810734-19810779
U4:39B	chr2L:21215036-21215178	-	chr2L:21215133-21215178

Start from bowtie mapped bam files:

```
bsub samtools sort Lu001.bam Lu001_sorted bsub samtools index Lu001_sorted.bam
```

Extract reads mapped to all U4 snRNA paralogs (optional, useful for comparing with reads mapped to only variable regions)

```
bsub samtools view -o Lu001.U4_25F.sam Lu001_sorted.bam chr2L:5565619-5565766

bsub samtools view -o Lu001.U4_38AB.sam Lu001_sorted.bam chr2L:19810734-19810875

bsub samtools view -o Lu001.U4_39B.sam Lu001_sorted.bam chr2L:21215036-21215178
```

Merge paralogs and remove mismatches:

```
cat Lu001.U4_* > Lu001.U4.sam awk ’$14 == "NM:i:0"’ Lu001.U4.sam > Lu001.U4.noMM.sam

rm Lu001.U4_* Lu001.U4.sam
```

Extract reads mapped to variable regions:

```
bsub samtools view -o Lu001.U4_u25F.sam Lu001_sorted.bam chr2L:5565619-5565665

bsub samtools view -o Lu001.U4_u38AB.sam Lu001_sorted.bam chr2L:19810734-19810779

bsub samtools view -o Lu001.U4_u39B.sam Lu001_sorted.bam chr2L:21215133-21215178
```

Remove reads with mismatches:

```
awk ’$14 == "NM:i:0"’ Lu001.U4_u25F.sam >Lu001.U4_u25F.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U4_u38AB.sam >Lu001.U4_u38AB.noMM.sam

awk ’$14 == "NM:i:0"’ Lu001.U4_u39B.sam >Lu001.U4_u39B.noMM.sam
```

Remove intermediate files:

```
rm Lu001.U4_u25F.sam Lu001.U4_u38AB.sam Lu001.U4_u39B.sam
```

4.6.6 *Drosophila* U5 snRNA

Drosophila has 7 U5 paralogs: U5:14B, U5:23D, U5:34A, U5:35D, U5:38ABa, U5:ABb and U5:63BC. The 5' part of U5 is identical in all of them, but the 3' end is very different among them (see alignment of U5 paralogs). Reads spanning the 3' end variable region can be used to distinguish all of them from each other. In fact all reads must overlap the highlighted nucleotide ('s'), and this is enough for the retrieval of most if not all unique reads. There is no need for any adjustment of the coordinates for reads of different lengths. U5 snRNA genomic coordinates of variable regions are as follows.

U5 snRNAs	**genomic coordinates**	**strand**	**variable region**
U5:14B	chrX:16148019-16148150	–	chrX:16148052-16148052
U5:23D	chr2L:3048701-3048831	+	chr2L:3048797-3048797
U5:34A	chr2L:13244848-13244974	+	chr2L:13244944-13244944
U5:35D	chr2L:15751557-15751682	–	chr2L:15751587-15751587
U5:38ABa	chr2L:19811948-19812074	–	chr2L:19811978-19811978
U5:38ABb	chr2L:19816414-19816540	+	chr2L:19816509-19816509
U5:63BC	chr3L:3090801-3090923	+	chr3L:3090895-3090895

Start from bowtie mapped bam files:

```
bsub samtools sort Lu001.bam Lu001_sorted bsub samtools index Lu001_sorted.bam
```

Extract reads mapped to all U5 snRNA paralogs (optional, useful for comparing with reads mapped to only variable regions)

```
bsub samtools view -o Lu001.U5_14B.sam Lu001_sorted.bam chrX:16148041-16148150

bsub samtools view -o Lu001.U5_23D.sam Lu001_sorted.bam chr2L:3048701-3048831

bsub samtools view -o Lu001.U5_34A.sam Lu001_sorted.bam chr2L:13244848-13244974

bsub samtools view -o Lu001.U5_35D.sam Lu001_sorted.bam chr2L:15751557-15751682

bsub samtools view -o Lu001.U5_38ABa.sam Lu001_sorted.bam chr2L:19811948-19812074

bsub samtools view -o Lu001.U5_38ABb.sam Lu001_sorted.bam chr2L:19816414-19816540

bsub samtools view -o Lu001.U5_63BC.sam Lu001_sorted.bam chr3L:3090801-3090923
```

```
cat Lu001.U5_14B.sam Lu001.U5_23D.sam Lu001.U5_34A.sam Lu001.U5_35D.sam
    Lu001.U5_38ABa.sam Lu001.U5_38ABb.sam Lu001.U5_63BC.sam > Lu001.U5.sam
```

Remove mismatches

```
awk '$14 == "NM:i:0"' Lu001.U5.sam > Lu001.U5.noMM.sam
```

Remove intermediate files

```
rm Lu001.U5_14B.sam Lu001.U5_23D.sam Lu001.U5_34A.sam Lu001.U5_35D.sam
    Lu001.U5_38ABa.sam Lu001.U5_38ABb.sam Lu001.U5_63BC.sam Lu001.U5.sam
```

Extract reads mapped to the variable regions only

```
bsub samtools view -o Lu001.U5_u14B.sam Lu001_sorted.bam chrX:16148052-16148052
bsub samtools view -o Lu001.U5_u23D.sam Lu001_sorted.bam chr2L:3048797-3048797
bsub samtools view -o Lu001.U5_u34A.sam Lu001_sorted.bam chr2L:13244944-13244944
bsub samtools view -o Lu001.U5_u35D.sam Lu001_sorted.bam chr2L:15751587-15751587
bsub samtools view -o Lu001.U5_u38ABa.sam Lu001_sorted.bam chr2L:19811978-19811978
bsub samtools view -o Lu001.U5_u38ABb.sam Lu001_sorted.bam chr2L:19816509-19816509
bsub samtools view -o Lu001.U5_u63BC.sam Lu001_sorted.bam chr3L:3090895-3090895
```

Remove reads with mismatches

```
awk '$14 == "NM:i:0"' Lu001.U5_u14B.sam >Lu001.U5_u14B.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u23D.sam >Lu001.U5_u23D.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u34A.sam >Lu001.U5_u34A.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u35D.sam >Lu001.U5_u35D.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u38ABa.sam >Lu001.U5_u38ABa.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u38ABb.sam >Lu001.U5_u38ABb.noMM.sam
awk '$14 == "NM:i:0"' Lu001.U5_u63BC.sam >Lu001.U5_u63BC.noMM.sam
```

Remove intermediate files

```
rm Lu001.U5_u14B.sam Lu001.U5_u23D.sam Lu001.U5_u34A.sam Lu001.U5_u35D.sam
    Lu001.U5_u38ABa.sam Lu001.U5_u38ABb.sam Lu001.U5_u63BC.sam
```

4.6.7 *Drosophila* U6 snRNA

Drosophila has 3 identical U6 paralogs: U6:96Ca, U6:96Cb and U6:96Cc. However, the Bowtie mapping procedure randomly assign reads to each location, therefore may create heterogeneity in the expression levels. Here we treat them as one gene, and use the total number of reads to analyze enrichment.

```
bsub samtools view -o Lu001.U6_96Ca.sam Lu001_sorted.bam chr3R:20381810-20381916

bsub samtools view -o Lu001.U6_96Cb.sam Lu001_sorted.bam chr3R:20382414-20382520

bsub samtools view -o Lu001.U6_96Cc.sam Lu001_sorted.bam chr3R:20382937-20383043

cat Lu001.U6_* > Lu001.U6.sam awk '$14 == "NM:i:0"' Lu001.U6.sam > Lu001.U6.noMM.sam

rm Lu001.U6_* Lu001.U6.sam
```

4.6.8 Mouse U1 snRNA

Mammalian genomes usually contain multiple genes for each major spliceosomal snRNA. They are usually not well characterized due to their repetitive nature. To analyze U1, use the first 5 equations to solve the fractions, then use the last two to verify, because the last two equations are less reliable. Note bowtie mapping prints out all possible mapping locations, while tophat randomly selects one mapped location if multiple locations are found.

Equations:

$$a + b + c + d + e = 1 \tag{4.15}$$

$$c/(a + b + d + e) = m \tag{4.16}$$

$$a/(b + c + d + e) = n \tag{4.17}$$

$$(a + b)/(c + d + e) = r \tag{4.18}$$

$$e/(a + b + c + d) = s \tag{4.19}$$

$$(a + b + c) : d : e = t1 : t2 : t3 \tag{4.20}$$

$$(d + e)/(a + b + c) = u \tag{4.21}$$

Solutions:

$$a = n/(1+n) \tag{4.22}$$

$$b = 1-(a+c+d+e) \tag{4.23}$$

$$c = m/(1+m) \tag{4.24}$$

$$d = 1/(r+1) - m(1+m) - s/(1+s) \tag{4.25}$$

$$e = s/(1+s) \tag{4.26}$$

Mouse U1_m

```
samtools view file1_sorted.bam mU1a1:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1a1v:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b1b2:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b6:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b6v:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b1b2:33 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1a1:33 > file1_sorted_m.sam

samtools view file1_sorted.bam mU1a1v:33 >> file1_sorted_m.sam

samtools view file1_sorted.bam mU1b6:33 >> file1_sorted_m.sam

samtools view file1_sorted.bam mU1b6v:33 >> file1_sorted_m.sam

cut -f1 file1_sorted_m.sam | sort -g | uniq | wc -l rm file1_sorted_m.sam
```

Mouse U1_n

```
samtools view file1_sorted.bam mU1a1:60 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1a1v:60 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b1b2:60 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b6:60 | grep 'NM:i:0' | wc -l

samtools view file1_sorted.bam mU1b6v:60 | grep 'NM:i:0' | wc -l
samtools view file1_sorted.bam mU1a1:60 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1v:60 > file1_sorted_n.sam

samtools view file1_sorted.bam mU1b1b2:60 >> file1_sorted_n.sam

samtools view file1_sorted.bam mU1b6:60 >> file1_sorted_n.sam

samtools view file1_sorted.bam mU1b6v:60 >> file1_sorted_n.sam

cut -f1 file1_sorted_n.sam | sort -g | uniq | wc -l rm file1_sorted_n.sam
```

Mouse U1_r

```
samtools view file1_sorted.bam mU1a1:62-78 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1v:62-78 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b1b2:62-78 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6:62-78 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6v:62-78 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1:62-78 > file1_sorted_r1.sam

samtools view file1_sorted.bam mU1a1v:62-78 >> file1_sorted_r1.sam

cut -f1 file1_sorted_r1.sam | sort -g | uniq | wc -l

samtools view file1_sorted.bam mU1b1b2:62-78 > file1_sorted_r2.sam

samtools view file1_sorted.bam mU1b6:62-78 >> file1_sorted_r2.sam

samtools view file1_sorted.bam mU1b6v:62-78 >> file1_sorted_r2.sam

cut -f1 file1_sorted_r2.sam | sort -g | uniq | wc -l
```

Mouse U1_s

```
samtools view file1_sorted.bam mU1a1:107-131 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1v:107-131 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b1b2:108-132 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6:108-132 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6v:108-132 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1:107-131 > file1_sorted_s.sam

samtools view file1_sorted.bam mU1a1v:107-131 >> file1_sorted_s.sam
samtools view file1_sorted.bam mU1b1b2:108-132 >> file1_sorted_s.sam

samtools view file1_sorted.bam mU1b6:108-132 >> file1_sorted_s.sam

cut -f1 file1_sorted_s.sam | sort -g | uniq | wc -l

samtools view file1_sorted.bam mU1b6v:108-132 | grep ’NM:i:0’ | wc -l
```

Mouse U1_t

```
grep mU1 file1.sam > file1_U1.sam grep GGACT file1_U1.sam | grep ’NM:i:0’ | wc -l

grep GGAGC file1_U1.sam | grep ’NM:i:0’ | wc -l grep GGGCT file1_U1.sam | grep ’NM:i:0’ | wc -l
```

Mouse U1_u

```
samtools view file1_sorted.bam mU1a1:156 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1a1v:156 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b1b2:157 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6:158 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU1b6v:157 | grep ’NM:i:0’ | wc -l
```

4.6.9 Mouse U2 snRNA

mU2.2 and mU2.3 are identical. mU2.1 and mU2.2 only differs at nt186, thus not practical to distinguish. Therefore mU2.1, mU2.2 and mU2.3 are considered as one isoform. The first three nucleotide variations are used to separate them into three isoforms.

```
samtools view file1_sorted.bam mU2.1 > file1_U2.sam

samtools view file1_sorted.bam mU2.2 >> file1_U2.sam

samtools view file1_sorted.bam mU2.3 >> file1_U2.sam

grep CGTCCTCTATCC file1_U2.sam | grep ’NM:i:0’ | cut -f1 | sort -g | uniq | wc -l

samtools view file1_sorted.bam mU2.4 | grep CGCCCTCTATCT | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU2.5 | grep TGTCCTCTATCT | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU2.1:186-186 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU2.2:186-186 | grep ’NM:i:0’ | wc -l
samtools view file1_sorted.bam mU2.3:186-186 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU2.4:186-186 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU2.5:186-186 | grep ’NM:i:0’ | wc -l
```

4.6.10 Mouse U4 snRNA

```
samtools view file1_sorted.bam mU4a:89-100 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU4b:89-100 | grep ’NM:i:0’ | wc -l
```

4.6.11 Mouse U5 snRNA

```
samtools view file1_sorted.bam mU5.1:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.2:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.3:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.4:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.5:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.6:103-114 | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.1 > file1_U5.12.sam

samtools view file1_sorted.bam mU5.2 >> file1_U5.12.sam

grep CTCTGAGTCTTAA file1_U5.12.sam | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.3 | grep TTCTGAGTCTTAC | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.4 | grep ATCTGAGTCTTAA | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.5 | grep ATCTGAGTCATTA | grep ’NM:i:0’ | wc -l

samtools view file1_sorted.bam mU5.6 | grep TTCTGAGTCTTAA | grep ’NM:i:0’ | wc -l
```

CHAPTER 5: SMN body formation caused by a block in snRNP assembly

Zhipeng Lu and A Gregory Matera

Note: This chapter is a manuscript currently in preparation. Even though the overall framework is complete, some of the results presented in this chapter are not conclusive, and further studies are required to make the conclusions. The experiments needed to finish this chapter are indicated in the manuscript and also summarized at the end of this chapter.

5.1 Abstract

Subcellular organelles compartmentalize the cell to provide spatial separation of cellular processes. Organelles can be divided into two major categories, depending on the presence or absence of membranes. Advances in microscopy and *in situ* labeling of molecules in the past decades enabled the discovery of many non-membrane-bound organelles, most of which are large assemblies of ribonucleoprotein (RNP) complexes. Here we describe a novel RNP granule in the *Drosophila* primary spermatocytes. This organelle, termed the SMN body, contains components of the Survival Motor Neuron (SMN) complex, and canonical Sm proteins, suggesting a role for these granules in the cytoplasmic steps of snRNP biogenesis. Despite their compositional similarity to the previously described egg chamber U bodies, these RNP granules do not contain snRNAs, and act differently from U bodies. Detailed analysis of the SMN body suggests that they are distinct from all previously known RNP granules, e.g. P bodies, piNG bodies etc. We also provide evidence that the previously identified U bodies are not present in other tissues or species, and are likely to be a specialized structure in *Drosophila* female germline. Using live imaging of egg chambers, we show that the U bodies are transport particles for snRNPs. It has been shown previously that perturbations of the snRNP assembly pathway lead to the formation of Cajal bodies in the nuclei of cells that do not normally have them (David Stanek, personal communication). Interestingly, there exist in the literature a few reports about U

body-like granules in somatic cells after perturbation of the snRNP biogenesis pathway. We propose that the formation of cytoplasmic U bodies/ SMN bodies is a result of the altered snRNP assembly dynamics.

5.2 Introduction

Organelles are fundamental organizers of the cellular components and functions. Subcellular compartments in the cells separate different cellular processes and chemical reactions. This spatial separation prevents unwanted interference among different processes and concentrates certain molecules for more efficient reactions. Lipid membranes surround most organelles, such as the nuclei, mitochondria, Golgi apparati, lysosomes, etc. In addition to the membrane-bound organelles, there are also non-membrane-bound organelles, and many of them are ribonucleoprotein (RNP) granules, for example, P bodies, P granules, stress granules, neuronal granules, Cajal bodies, PML bodies, etc (Schisa, 2012; Ramaswami *et al.*, 2013). Compared to membranous organelles, much less is known about the mechanisms and principles governing the formation of RNP granules, and how the formation of the RNP granules affects gene expression.

Spliceosomal small nuclear ribonucleoproteins (snRNPs) are the core components of eukaryotic mRNA splicing machinery and their biogenesis involves multiple steps in both the cytoplasm and the nucleus (Matera *et al.*, 2007; Matera and Wang, 2014). There are two types of spliceosomes in most eukarytoes, the major U2 type and the minor U12 type, each consisting of 5 snRNPs. The major spliceosome contains the U1, U2, U4, U5 and U6 snRNPs, whereas the minor spliceosome contains the U11, U12, U4atac, U5 and U6atac snRNPs. Two classes of snRNAs exist in most eukaryotes, the pol II type snRNAs, including U1, U2, U4, U5, U11, U12 and U4atac, and the pol III type snRNAs, U6 and U6atac (Patel and Steitz, 2003).

After pol II snRNAs are transcribed in the nucleus, they are modified on the 5' end with a m7G cap and processed on the 3' end by the integrator complex (Baillat *et al.*, 2005; Eliceiri and Sayavedra, 1976; Neuman de Vegvar and Dahlberg, 1990; Mattaj, 1986). Certain snRNP-specific proteins are assembled onto the snRNAs before export to the cytoplasm by Phax-Crm1-CBC complex with the help of Ran-GTP (Ohno *et al.*, 2000). After export, the SMN complex assembles the canonical Sm ring onto the snRNAs (Fischer *et al.*, 1997; Pellizzoni *et al.*, 2002). Trimethyl-guanosine synthase 1 (Tgs1) then hypermethylates the m7G cap to generate the trimethyl-guanosine (TMG) cap (Mouaikel *et al.*, 2003). Sm ring assembly and TMG capping stimulates Snurportin-Importin beta (Spn-ImpB) dependent import back into the nucleus (Palacios *et al.*, 1996, 1997; Narayanan *et al.*, 2002). snRNPs undergo further maturation steps in the nucleus. snRNPs are extensively modified by psudouridylation and 2'-O-methylation in the Cajal body and other snRNP-specific proteins are assembled (Jady *et al.*, 2003; Kiss, 2001; Darzacq *et al.*, 2002). Compared to the pol II type snRNPs, assembly of the pol III type snRNPs is less complicated and occurs entirely inside the nucleus (Vankan *et al.*, 1990; Terns *et al.*, 1993; Boelens *et al.*, 1995; Pante *et al.*, 1997; Spiller *et al.*, 2007).

The SMN (Survival Motor Neuron) complex contains SMN and at least seven other proteins Gemin2-7 and Unrip, and plays an important role in the assembly of Sm proteins onto pol II snRNAs in the cytoplasm and in snRNP import into the nucleus (Liu *et al.*, 1997; Pellizzoni *et al.*, 2002; Yong *et al.*, 2002; Massenet *et al.*, 2002). The SMN protein is the product of a disease-determining gene, the loss of which causes a neurodegenerative disease spinal muscular atrophy (SMA) (Lefebvre *et al.*, 1995). Consistent with its major role in snRNP assembly in the cytoplasm, SMN is mainly diffusely localized in the cytoplasm. A small portion of the SMN proteins also localize to the Cajal bodies in the nucleus (Liu *et al.*, 1997).

The Cajal body, a subnuclear organelle, was discovered by Santiago Ramon y Cajal over one hundred years ago, and it exists mainly in proliferative cells and neurons. The Cajal body, which is molecularly defined by the presence of coilin, has been implicated in a number of nuclear processes, including the assembly of snRNPs, telomerase, RNA polymerase etc. However, the structure *per se* is not required for these functions. The formation of proper Cajal bodies has been shown to depend not only on the presence of coilin, but also on the methylation of the coilin RG box (Hebert *et al.*, 2002). It was recently shown that blocking tri-snRNP assembly in the nucleus causes formation of Cajal bodies in primary cells that normally do not have Cajal bodies (Novotny et al. unpublished data).

Even though much is known about the steps in assembly of snRNPs, little is known about the spatial organization of the steps in the cytoplasm. In most cell types, the localization of snRNP components and snRNP assembly machineries in the cytoplasm is diffuse. However, certain perturbations of the snRNP assembly pathway cause formation of cytoplasmic foci containing some of the snRNP assembly components (Pellizzoni *et al.* (1998); Mouaikel *et al.* (2003); Takata *et al.* (2012), unpublished results by Boysen and Gruss, and unpublished results from our lab). Several recent studies showed that in certain cell types, SMN complex and the snRNP components are concentrated in granular structures in the cytoplasm (Liu and Gall, 2007; Lee *et al.*, 2009; Cauchi *et al.*, 2010). These results suggest that snRNP assembly in the cytoplasm may also involve RNP granule formation, similar to the Cajal bodies. However, it is still not clear whether these cytoplasmic snRNP-containing granules exist in different cell types, and how and why they form.

In order to study the organization of the cytoplasmic steps of snRNP assembly, we analyzed the SMN and snRNP-containing granules in detail in many different cell types. We found that SMN/U bodies do not exist in most cell types. The most prominent SMN/U bodies are only detected in *Drosophila* ovaries and testes. The female germline U bodies contain most of the known cytoplasmic snRNP assembly components and are snRNP transport particles during oogenesis. The SMN bodies in the testes contain SMN complex components, Sm proteins, but not snRNPs. A comprehensive analysis of previously reported SMN/U body-like structures suggests that the formation of these structures is a consequence of the imbalance in snRNP assembly. These results revealed an important principle in the formation of RNP assembly (but not storage) granules, that is, RNP granules form as a result of imbalance between influx and efflux.

5.3 Results

Survival Motor Neuron (SMN) is an essential protein in almost all eukaryotes and one of its best-characterized functions is assembling small nuclear ribonucleoproteins (snRNPs), which are the building blocks of spliceosomes. Several subcellular SMN-containing structures have been described, including nuclear Cajal bodies and Gems that are present in many cell types, and cytoplasmic U bodies that are present in *Drosophila* ovaries (Carvalho *et al.*, 1999; Gall, 2000; Liu *et al.*, 2006; Liu and Gall, 2007). Both of these structures are likely involved in some steps of snRNP biogenesis. In order to study the spatial organization of snRNP assembly in

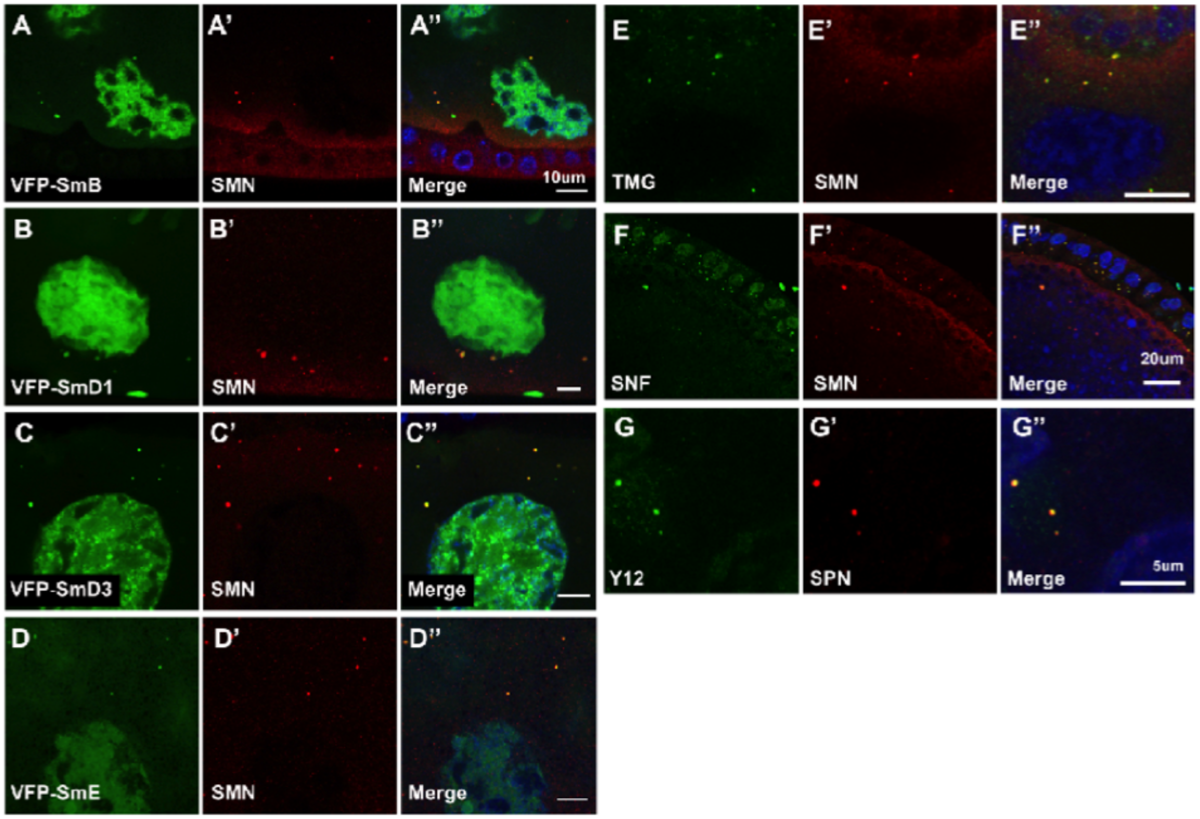


Figure 5.1: **VFP-tagged Sm proteins are properly localized in U bodies in the female germline.** Fly ovaries over-expressing VFP-tagged Sm proteins (*nos-Gal4 VFP-SmB, nos-Gal4 VFP-SmD3, nos-Gal4 VFP-SmE* and *da-Gal4 VFP-SmD1*) were stained with SMN antibody (A-D). Wild type (Oregon R) ovaries were stained with antibodies as shown on the figure (E-G). Scale bars in (A-E) are the same. The giant nuclei are the nurse cell nuclei, whereas the smaller ones (in A, B and F) are the follicle cell nuclei.

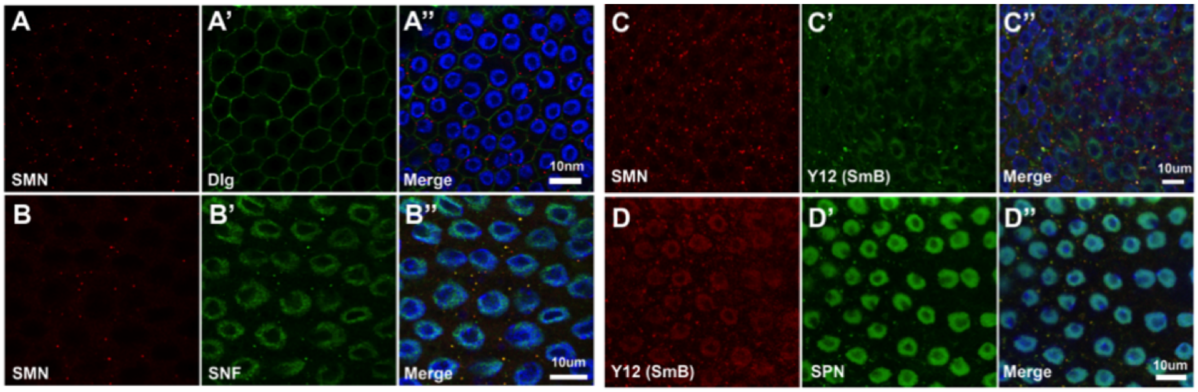


Figure 5.2: **Follicle cells contain U bodies.** Wild type (Oregon R) fly ovaries were stained with antibodies as shown on the figure. Images were taken on stage 8-10 follicle cells. The column on the right merges the red, green and blue (DAPI) channels. Dlg (A-A', Discs large) marks the cytoplasmic membrane. SPN: snurportin1.

the cytoplasm, we examined the localization of SMN in many different *Drosophila* cell/tissue types using immunofluorescence.

5.3.1 SMN containing cytoplasmic RNP granules are not ubiquitous

Whereas in most tissues and cell types SMN staining is diffuse and homogeneous in the cytoplasm, we noticed punctate cytoplasmic staining in the *Drosophila* ovary nurse cells, oocyte and follicle cells, and testis spermatocytes (Figures 5.1, 5.2 and 5.6). Contrary to what was reported previoiusly by Liu and Gall (Liu and Gall, 2007), we did not see U body like structures in other cell types. Antibodies against SMN have been used extensively in immunofluorescence on many cell/tissue types in various organisms for two decades; however, none of these studies have reported U body like structures in the cytoplasm, for example, human fetal tissues (Burlet *et al.*, 1998), HeLa cells, rat supraoptic nuclei, rat trigeminal ganglia (Carvalho *et al.*, 1999), fly malphigian tubules (Natalizio and Matera, 2013), fly S2 cells (unpublished observations from our lab), fly body wall muscle and gastric caecal cells (Cauchi, 2011) and mouse testis. These studies suggest that U bodies are not ubiquitous RNP granules in all cell types.

A previous study by Liu et al. (Liu and Gall, 2007) suggested that U bodies are present in many different kinds of cells and tissues, in several different organisms, including human HeLa cells. However, the extensive use of SMN and Sm protein (e.g. Y12) antibodies in immunofluorescence studies in our lab and others did not reveal cytoplasmic granules like the U bodies, except in the *Drosophila* ovaries and the testes (Gonsalvez *et al.* (2010); Rajendra *et al.* (2007); Takata *et al.* (2012), etc.). In fact, SMN has been shown to be present in stress granules in several cell types, and the stress granules do not contain snRNP components (Hua and Zhou, 2004b,a). Also the localization patterns of Sm-class snRNPs have been extensively characterized using anti-sense probes and antibodies in many cell types (e.g. 3T3 mouse fibroblasts, BRL rat liver cells, Vero African green monkey kidney cells), and none of these studies have shown evidence of cytoplasmic aggregates like U bodies (Carmo-Fonseca and Hurt, 1991; Matera and Ward, 1993; Frey and Matera, 1995). The detection of YFP-Lsm11 containing cytoplasmic dots in fly larval brain suggests the U bodies might be present in other cell types (Liu *et al.*, 2006) (further studies will be needed to show whether these granules contain snRNPs, for example using TMG, SNF antibodies and performing in situ hybridizations against snRNAs). Therefore we conclude that the U bodies, even if present in cells other than fly gonads, are not common RNP granules.

5.3.2 The ovarian U bodies are passive snRNP transport particles

Consistent with previous reports, we detected cytoplasmic U bodies in *Drosophila* egg chambers (Liu and Gall, 2007; Lee *et al.*, 2009; Cauchi *et al.*, 2010) (Figures 5.1 and 5.2). Liu and Gall showed that the germline U bodies in *Drosophila* egg chambers contain SMN, SmB, Lsm10, Lsm11, snRNAs (including U6), Gemin2, Gemin3 and Gemin5. All of these RNA and proteins are involved in the snRNP biogenesis process.

To further characterize the U bodies in ovaries, we labeled U bodies with VFP-tagged Sm proteins and performed immunofluorescence analysis against multiple components of the snRNP assembly pathway. The transgenic VFP-Sm proteins are properly localized to the nuclei and the cytoplasmic U bodies (Figures 5.1 and 5.3). In addition, we also detected SNF (U1A/U2B"), SPN (snurportin), U2 and U6 snRNAs, TMG cap and Tgs1 in the U bodies. These results suggest that the snRNPs in female germline U bodies are already assembled in the cytoplasm and ready for import. Then we first determined the stages during which U bodies can be seen. We determined the stage of the egg chambers by the appearance of the nurse cell nuclei and the

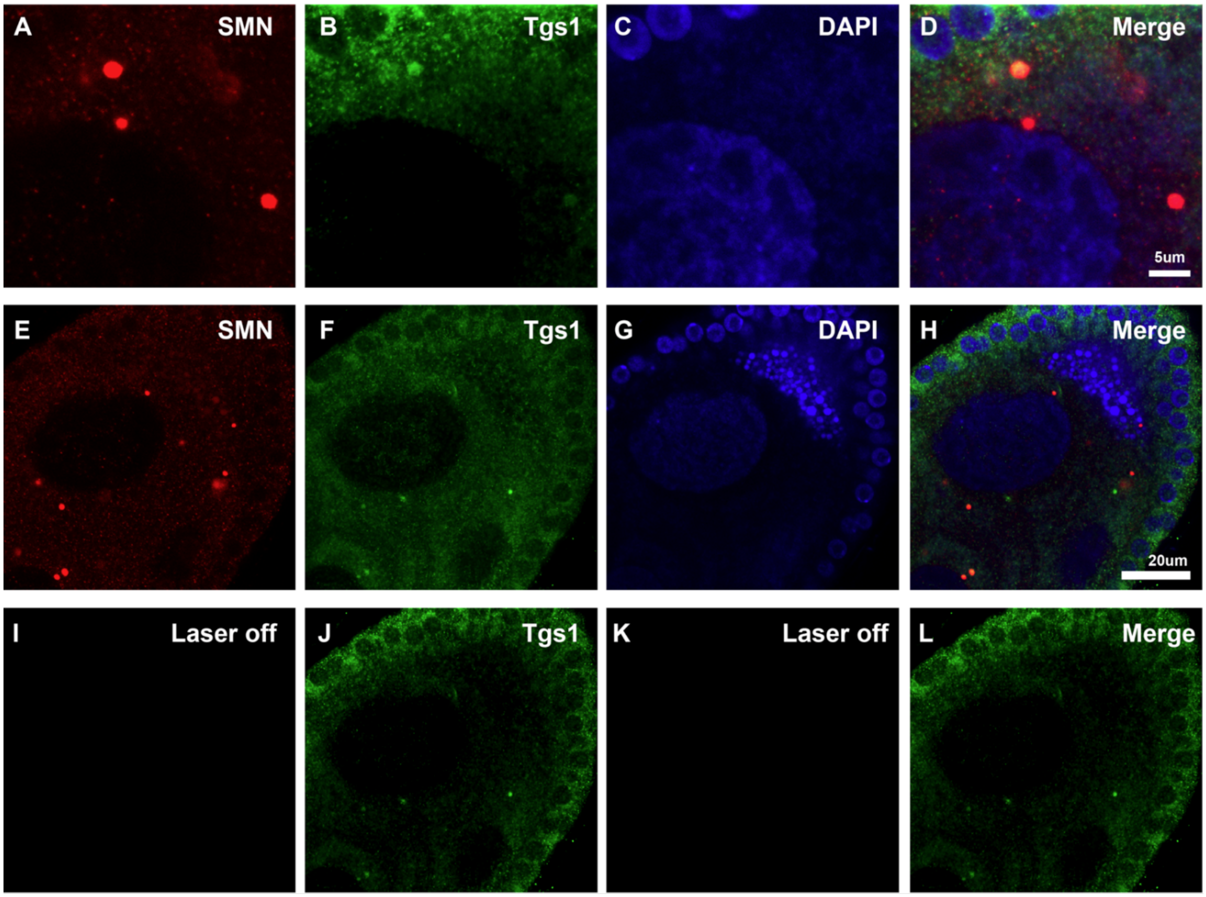


Figure 5.3: **Tgs1 is localized to U bodies.** Wild type ovaries were stained with SMN and Tgs1 antibodies. The Tgs1 antibody does not work well for immunofluorescence since only very weak signal is observed. Panels A-D and E-H show two examples of colocalization of SMN and Tgs1. Only some of the granules overlap, probably due to the weakness of the Tgs1 antibody. Panels I-L are the same field as panels E-H, except that the lasers for the red (SMN) and blue (DAPI) channels were turned off. This was done to show that the weak signal from the green (Tgs1) channel is not due to bleed through.

size of the egg chambers. We stained wild type fly ovaries with an antibody against *Drosophila* SMN, and we found that the U bodies appear as early as stage 5 (Figure 5.4). Stage 5 is characterized by the de-condensation of the polytene chromosomes (Dej and Spradling (1999), Bate and Arias *Development of Drosophila melanogaster*). Nurse cell chromosomes are polytene and blob-like until stage 5, and by stage 6, they are dispersed.

U6 snRNA is generally considered to be nuclear restricted, and its assembly does not involve a cytoplasmic phase, whereas other spliceosomal snRNAs (e.g. U1, U2, U4, U5, U11 and U12) have a cytoplasmic phase in their life cycle (Hamm and Mattaj, 1989; Vankan *et al.*, 1990; Terns

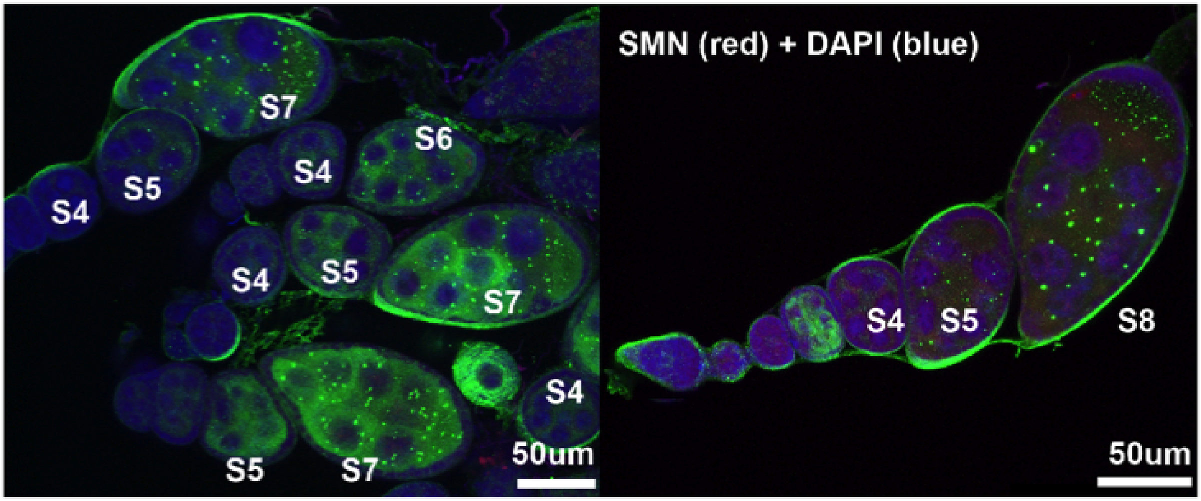


Figure 5.4: **U bodies appear in egg chambers around stage 5.** Wild type (Oregon R) fly ovaries were stained with SMN and DAPI (blue). U bodies are only visible from stage 5 onward. Assignment of egg chamber stages was performed according to Bate and Arias (*The Development of Drosophila melanogaster*, 1993 CSHL Press). Note: overexpression of Sm proteins causes the U bodies to appear earlier than stage 5, therefore, we only used staining on endogenously expressed proteins for determining the stage of U bodies.

et al., 1993; Boelens *et al.*, 1995; Pante *et al.*, 1997; Spiller *et al.*, 2007). Evidence for several mechanisms has been presented to explain the nuclear retention of U6 snRNA, including the lack of export signals (Terns *et al.*, 1993), or the binding of La protein (Boelens *et al.*, 1995), and the Lsm2-8 complex (Spiller *et al.*, 2007). Regardless, the presence of U6 snRNA in cytoplasmic U bodies suggests that U bodies are not simply cytoplasmic factories for U6 snRNP assembly, since nuclear-retained U6 snRNA is assembled just fine.

Early embryogenesis in *Drosophila* depends on the maternal deposition of signals and nutrients into the oocytes and eventually into the fertilized eggs. However the oocyte nucleus is mostly transcriptionally inactive; the majority of the constituents is supplied by the polyploid nurse cells through intercellular openings, called ring canals (Spradling 1993, *Developmental Genetics of Oogenesis*). Transport of these cellular components occurs in two phases, a slow phase during which specific molecules are gradually passed into the oocyte (either actively or passively), and a later rapid phase, during which the nurse cells empty their contents into the oocytes in a short time (Clark *et al.*, 2007; Bullock and Ish-Horowicz, 2001) (Dumping of nurse cell cytoplasmic contents occurs during stage 11, and apoptosis of nurse cells occurs at stage

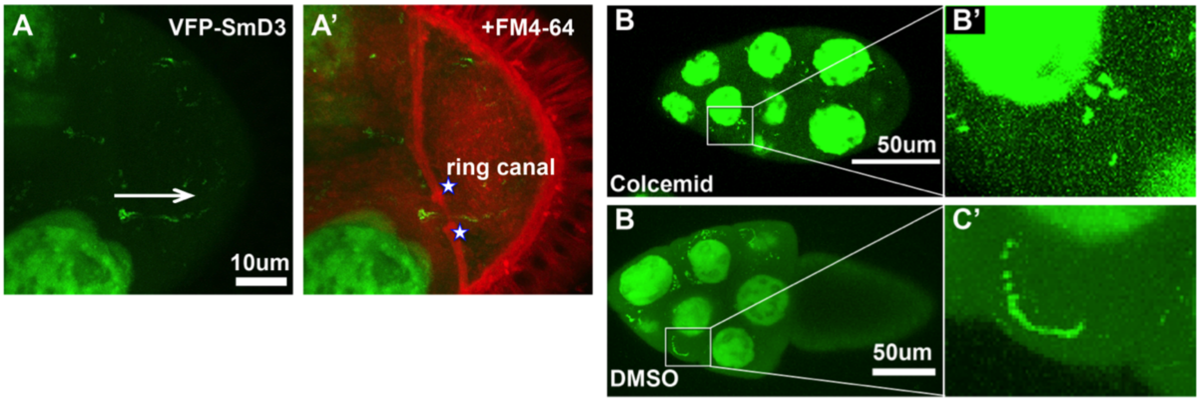


Figure 5.5: **U bodies are non-essential passive snRNP transport particles.** Experiments were performed on *now-Gal4 VFP-SmD3* ovaries. (A and A') Time lapse live imaging of U body movement from the nurse cell compartment to the oocyte compartment. FM4-64 labels membranes. The arrow indicates the direction of the U body movement. The two stars indicates the edges of a ring canal connecting a nurse cell and the oocyte. (B, B', C and C') Movement of U bodies is inhibited by microtubule depolymerizer Colcemid (B and B'), but not the solvent DMSO (C and C'). The images were overlaid time lapse images from live imaging.

12) (Foley and Cooley, 1998; Guild *et al.*, 1997). Therefore we considered the possibility that the U bodies are transport particles for snRNPs.

In order to study the dynamic behavior of U bodies during oogenesis, we performed live imaging of the ovaries expressing VFP-Sm proteins. Interestingly, the U bodies are highly mobile. Even though their movement does not follow a uniform direction, we found many of U bodies cross ring canals between nurse cells and from nurse cells to the oocyte. These transport events occur as early as stage 5, long before the dumping phase (stage 11) (Figure 5.5 and supplementary movie). To determine the mechanism of the U body movement, we depolymerized the microtubules using colcemid, and the directional movements ceased (Figure 5.5), suggesting that the U body movement depends on the cytoskeleton. We also determined the speed of the directional movement of U bodies. The U bodies move at 0.08um/sec, similar to the speed of cytoplasmic streaming, and much slower than active transport on cytoskeleton.

These results suggest that the U bodies in the cytoplasm of Drosophila female germline are snRNP transport particles. The absence of U bodies caused by SmD3 and Dart5 mutations suggest that U bodies are not essential for the survival of *Drosophila* or snRNP transport.

5.3.3 *Drosophila* testes contain SMN bodies

Drosophila testis is the only other tissue where we see U body like structures based on SMN staining. In order to further characterize these SMN positive granules, we first examined the cell types where they appear (Figure 5.6). Similar to the ovarioles, the *Drosophila* testis contains germline cells at different stages of spermatogenesis in a roughly sequential manner (Figure 5.6A). The tip of the testis, also known as the germinal proliferation center, contains a hub consisting of a few somatic apical cells, and several male germ line stem cells (number varies from 5 to 18) that surround and directly contact the hub (Bate and Arias, The Development of *Drosophila melanogaster*). The germline stem cells undergo asymmetrical cell divisions to give rise to stem cells for renewal and primary spermatogonia. Each primary spermatogonial cell is the mitotic founders of a cluster of synchronously dividing secondary spermatogonia (Figure 5.6A, the ones with the dense dark blue small nuclei). The secondary spermatogonia undergo the premeiotic S phase after which the cells are now called the primary spermatocytes. The primary spermatocytes enter an extended G2 phase, when the cells grow 25 times in volume and transcribe large numbers of genes that are required for the growth and differentiation (Figure 5.6A, the ones with the light blue large nuclei). During the growth phase, cells transit from polar to apolar spermatocytes.

Interestingly, even though SMN is highly expressed in the earlier stages of spermatogenesis including the spermatogonial stages, the SMN-containing granules are present only in primary spermatocytes, not earlier or later (Figure 5.6B and C). More detailed examination suggest that the SMN-containing granules are present in both the polar and apolar primary spermatocytes, after the cells enlarged (see for example, Figure 5.7). Unlike the female germline where 15 nurse cells support the development of a single oocyte in each egg chamber, the male germ cells all develop to sperms, therefore, there is no need for mass transfer of materials from one cell to another. The SMN-containing granules in the primary spermatocytes must be different from the U bodies in ovaries.

The primary spermatocyte stage of spermatogenesis is characterized by the high transcriptional activity, which is important for the rapid growth of the spermatocytes and subsequent differentiation to sperms. It makes sense that larger amounts of snRNPs are needed for efficient splicing.

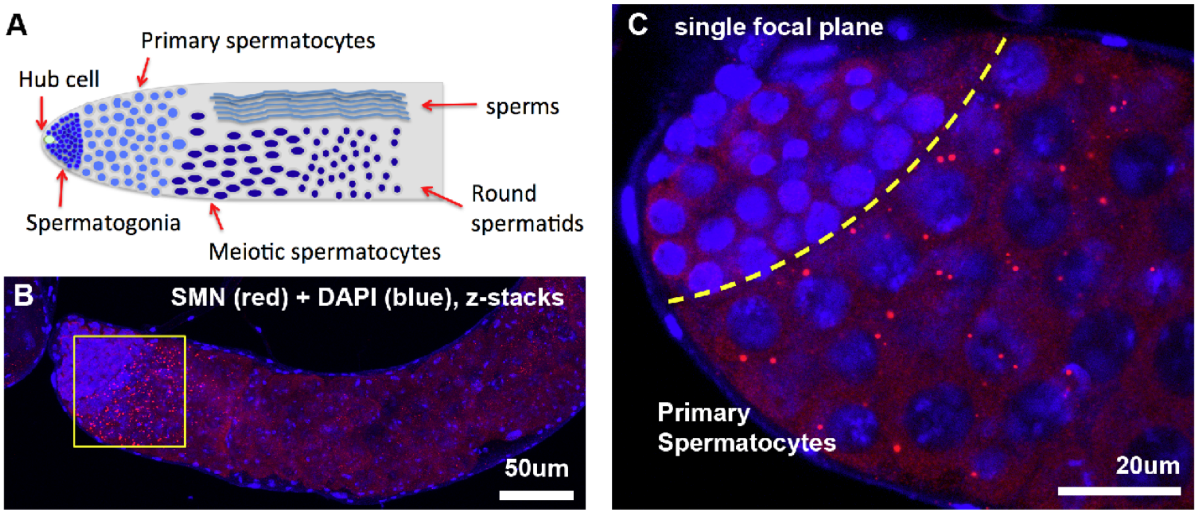


Figure 5.6: **SMN bodies are present only in primary spermatocytes.** (A). Diagram of the apical tip of a *Drosophila melanogaster* testis, showing the somatic hub cell, the germline spermatogonia (with smaller cell nuclei), primary spermatocytes (bigger nuclei and cell volume), meiotic spermatocytes, round spermatids and sperms. (B). An overview of the SMN staining in the apical tip of a testis. SMN is in red and DNA staining (DAPI) is in blue. Ten z-stacks were overlaid. The yellow box is shown in C. (C). Magnified view of the apical tip of the Drosophila testis, showing the spermatogonia and primary spermatocytes only. A single focal plane is shown here. Note that the SMN bodies are in the cytoplasm of primary spermatocytes.

We examined the localization of several components of the snRNP assembly pathway in *Drosophila* testes using immunofluorescence and in situ hybridization (Figure 5.7). We observed colocalization of SMN, Gemin2, SmB and SmD1 in the primary spermatocytes in SMN containing granules, however, we did not see U2 snRNA, SNF or TMG staining in cytoplasmic granules, even though they are highly abundant in the nucleus. These results suggest that the SMN-containing granules do not contain snRNPs, and are thus different from the U bodies in female germline. To assess whether these SMN-containing granules correspond to any other previously reported RNP granules, we performed immunofluorescence and in situ hybridizations against known markers of the other RNP granules, including Btz (Barentsz, nuage, P body and stress body component), PABP (poly(A) binding protein, P body and stress body component), PABP2 (nuclear isoform of PABP), Pacman (P body component), Fmr1 (P body and stress body component), Lsm10/Lsm11 (Histone Locus body (HLB) and U body component), Y10B

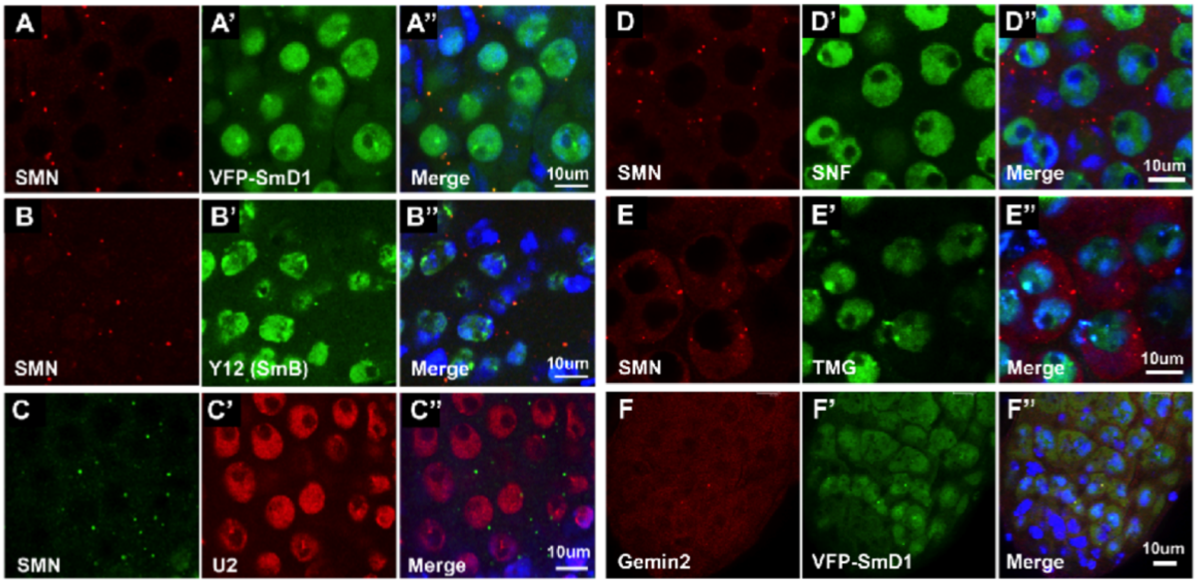


Figure 5.7: **SMN bodies contain SMN complex, Sm proteins but not snRNPs.** (A-A" and F-F") *da-Gal4 VFP-SmD1* fly testes were stained with SMN and Gemin2 antibodies. (B-B", D-D" and E-E") Oregon R (wild type) fly testes were stained with SMN, Y12, SNF and TMG antibodies. (C-C") Oregon R (wild type) fly testes were stained with SMN antibody and hybridized with U2 snRNA probe. The double prime panels are the overlay of the left and middle panels.

(ribosome component), poly(A) tails (oligo-dT in situ hybridization, P body and stress body component) (Figure 5.8). The SMN-containing granules do not contain any of these protein or RNA components. The SMN-containing granules are not the recently reported Yb bodies because the Yb bodies are only present in the somatic cells in the testes (Szakmary *et al.*, 2009; Qi *et al.*, 2011). These results suggest that the SMN-containing granule is a new type of RNP granule, and we named it the SMN body to differentiate it from the U body, since it does not contain snRNAs.

In order to see whether the testis SMN bodies are conserved in evolution, we performed immunofluorescence on mouse testis sections, using SMN and coilin antibodies (data not shown). We observed nuclear granules positive for coilin, suggesting that Cajal bodies are present in the nuclei, however, we did not see any obvious cytoplasmic granules positive for SMN. This result suggest that the SMN bodies are not present in mouse testes.

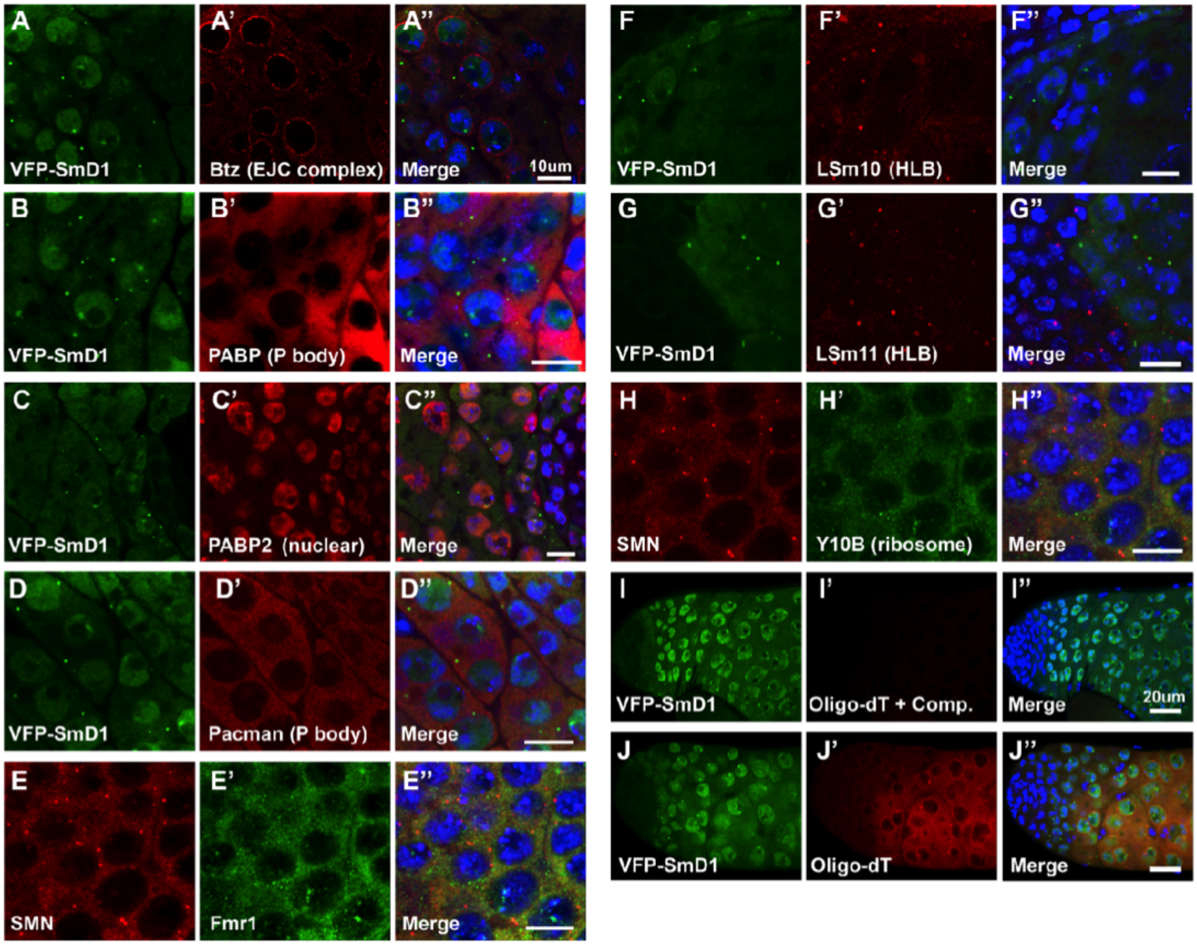


Figure 5.8: **SMN bodies are different from other RNP granules.** Oregon R (wild type, E and H) and *da-Gal4 VFP-SmD1* (A, B, C, D, F and G) fly testes were stained with the respective antibodies as indicated on the figure. (I' and J') were hybridized with Cy5-oligo(dT)20 probe, and 10 fold more unlabeled oligo(dT)20 was added as competitor in I'. The VFP-SmD1 panels are the fluorescence signals of the VFP. The double prime panels are the overlay of the left and middle panels. Scale bars in A-H are all 10um, whereas scale bars in I and J are 20um.

5.3.4 The SMN body is not affected by SMA-associated mutations

To further study how the SMN bodies are organized, we analyzed fly mutants for a few genes that are likely involved in SMN body formation (Figures 5.9, 5.12 and 5.10). Recently our lab generated an array of fly lines expressing SMN mutations identified from human SMA patients, in the fly *Smn* null background (Praveen *et al.* (2012), and Praveen et al. unpublished). These mutations disrupt different domains of the SMN, and a subset of these mutations are homozygous viable with testes. The D20V mutation is in the domain required for Gemin2 binding; G73R and I93F are in the Tudor domain required for Sm protein binding; whereas the G210C is in the YG box required for SMN oligomerization. We examined the localization of SMN mutant proteins in the *Smn* null background in fly testes. All of them are properly localized in the SMN bodies (further studies are required to examine how these mutations affect Sm protein localization to the SMN bodies).

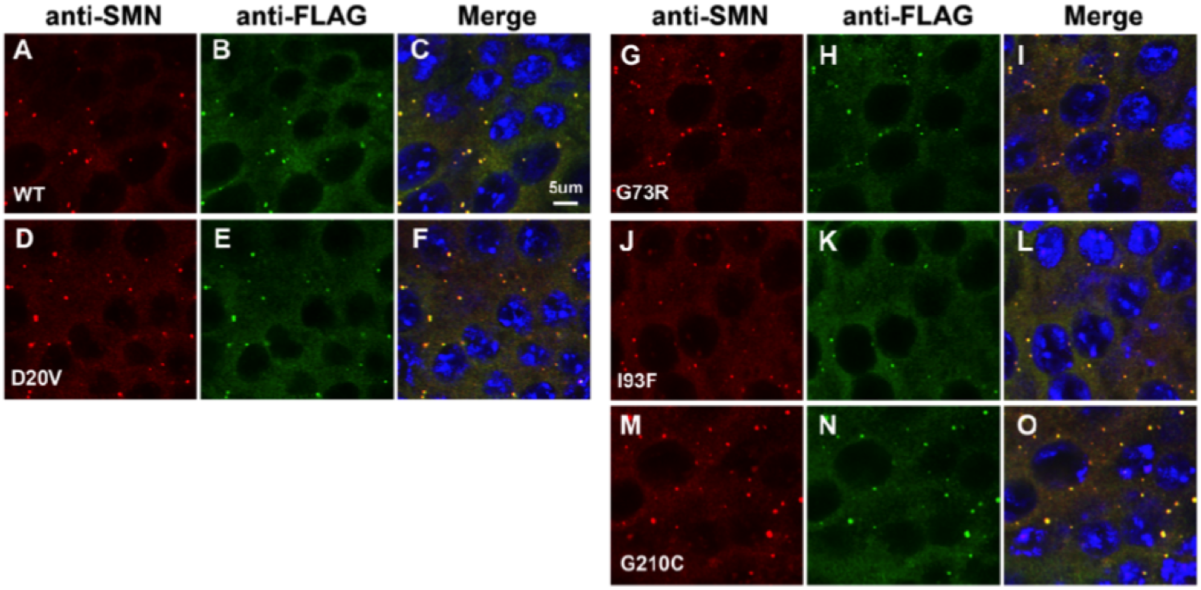


Figure 5.9: **SMN localization in SMN bodies is not affected by SMN mutations.** FLAG-tagged SMN transgenes were expressed in the *Smn*[X7/D] background under the control of endogenous SMN promoter. The transgenes include wild type (WT, A-C), D20V (D-F), G73R (G-I), I93F (J-L), and G210C (M-O). SMN still forms granules in all of these SMN mutants testes. All the panels have the same scale as shown in panel C. The right side panels are the overlay of the left and middle panels and DAPI staining.

It has been shown by our lab that symmetric arginine dimethylation of coilin RG box is required for interaction between coilin and SMN in the Cajal body (Hebert *et al.*, 2002). Three of the Sm proteins, SmB, SmD1 and SmD3, are also known to be extensively methylated in the RG box (Brahms *et al.*, 2001). GFP insertion in the middle of SmD3 protein significantly reduced the methylation of both SmB and SmD3 (Gonsalvez *et al.*, 2010). Dart5, the *Drosophila* homolog of PRMT5 (protein arginine methyltransferase 5) is required for methylation of Sm proteins (Gonsalvez *et al.*, 2007). Consistent with the fact that the Sm-binding Tudor domain mutations did not affect SMN localization to SMN bodies, loss of Sm protein methylation in the *Smd3pt* and *Dart5-1* background did not affect SMN localization either (Figure 5.10). More interestingly, even though loss of Sm protein methylation does not have noticeable effects on snRNP biogenesis, the *Smd3pt* mutant protein failed to localize to the SMN bodies. (Further immunofluorescence experiments are needed to show whether the *Dart5-1* mutation affects Sm protein localization to SMN bodies in the testes). These results suggest that posttranslational modification is generally involved in granule formation for both nuclear and cytoplasmic RNP granules.

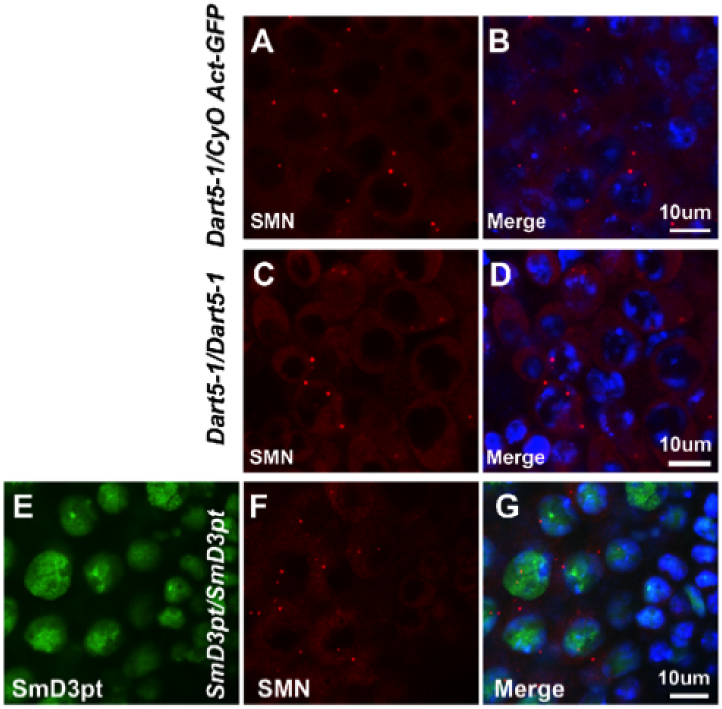


Figure 5.10: **SMN localization in SMN bodies does not require RG box methylation.** (A-D) SMN localization to SMN bodies in *Drosophila* testes is not affected by loss of Sm protein methylation in *Dart5-1* mutants. (A-B) are *Dart5-1* heterozygotes whereas (C-D) are homozygotes. (E-G) SmD3pt failed to localize to SMN bodies in *SmD3pt/SmD3pt* fly testes.

Previous studies showed that an Ago2 mutation *Ago2[51B]* causes loss of U bodies in the nurse cell compartment of the egg chambers but not in the oocytes (Liu and Gall, 2007). Our analysis of the same mutant showed a more severe phenotype (Figure 5.11). *Ago2[51B]* heterozygous ovaries have U bodies that are restricted to the oocyte compartment only, whereas homozygous ovaries completely lost the U bodies. Our Analysis of the snRNP levels did not reveal any obvious defects in snRNP biogenesis (data not shown). We analyzed the effects of siRNA pathway mutations on SMN body formation in the *Drosophila* testes. Interestingly, *Ago2[51B]* mutant testes lost SMN bodies completely in the testes, whereas the dcr-2 and r2d2 mutations did not affect SMN bodies. We cannot make any conclusions about the specific relationship between the RNAi pathway and the formation of SMN bodies based on these results.

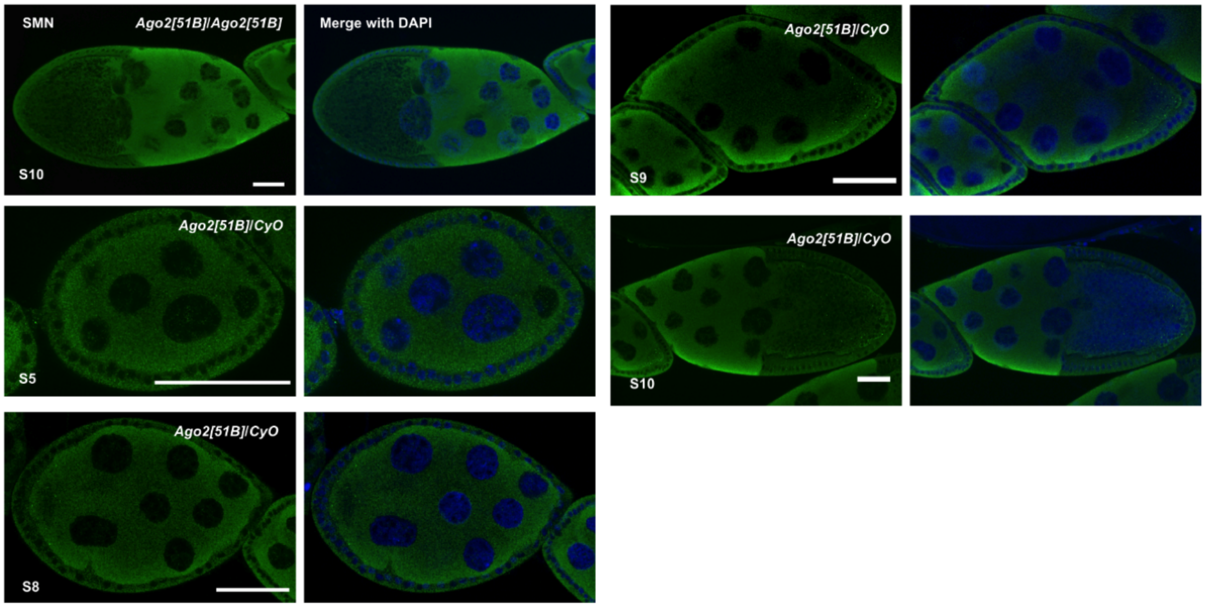


Figure 5.11: **U bodies are lost in *Ago2* mutant ovaries.** U bodies are lost in *Ago2[51B]* homozygous ovaries (first row). *Ago2[51B]* heterozygous ovaries show U bodies restricted to the oocyte compartment. Stages of the egg chambers were labeled on the figure. Scale bars are 50um.

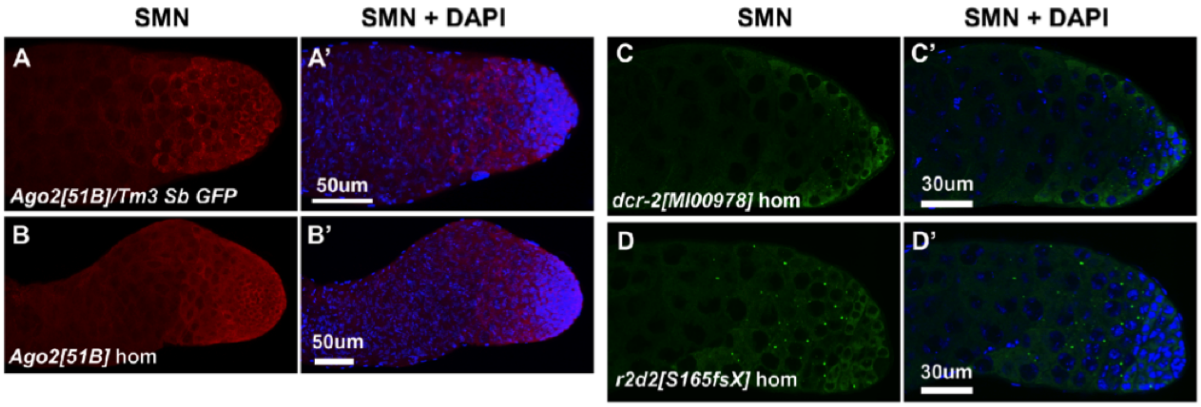


Figure 5.12: **SMN bodies are lost in *Ago2*, but not affected in *r2d2* or *dcr-2* mutants.** SMN and DAPI staining were performed on the testes with genetic backgrounds as shown on the figure. hom stands for homozygous.

5.4 Discussion

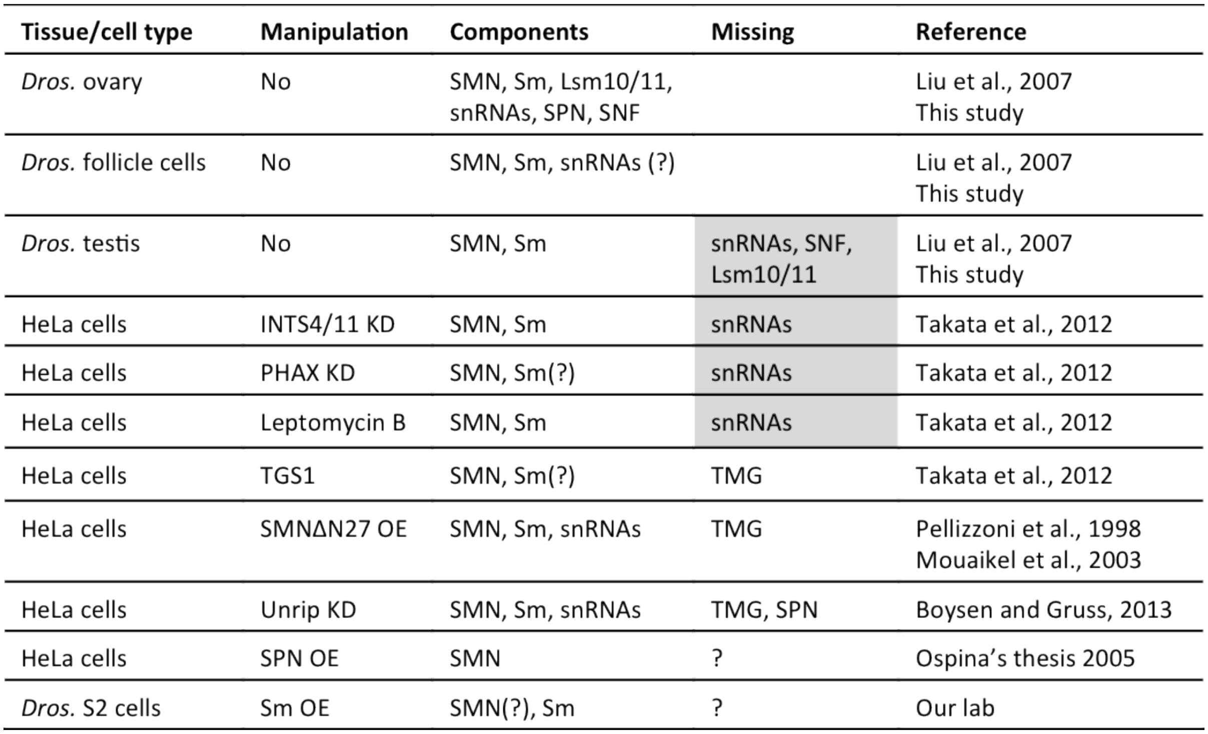
In this study, we characterized the U bodies in fly female germline and report the discovery of a related RNP granule, SMN body, in the male germline. Our results showed that the U bodies are the transport particles for snRNPs during oogenesis, even though formation of the microscope structure per se is not required for transport. We showed that surprisingly the SMN bodies contain SMN complex, Sm proteins, but not snRNPs, and SMN bodies are distinct from all the other known RNP granules.

5.4.1 Imbalanced cytoplasmic snRNP assembly causes U/SMN body formation

Previous studies have shown several requirements for the formation of U bodies, e.g. Ago2, SMN, symmetrical dimethylation of Sm protein RG boxes, etc (Liu and Gall (2007); Lee *et al.* (2009) and the current study Figures 5.10, 5.12 and 5.11). The surprising finding that the SMN bodies in *Drosophila* primary spermatocytes contain Sm proteins but not snRNPs suggests that SMN/U body formation could be caused by an imbalance in snRNP assembly, in cells where intercellular transport is not necessary. A number of previous studies investigating snRNP assembly reported U body like structures after perturbation of the assembly pathway in cells where U/SMN bodies are normally not visible. Here we summarized and examined these anecdotal observations and compared them to the U bodies in ovaries and SMN bodies in testes (Table 5.1).

Pellizzoni et al. and Mouaikel et al. reported that overexpression of a dominant negative form of SMN, SMNdeltaN27 causes formation of cytoplasmic granules that contain SMN, Tgs1, Sm proteins and snRNAs, but the snRNAs are not TMG-capped (Pellizzoni *et al.*, 1998; Mouaikel *et al.*, 2003). However, overexpression of wildtype SMN does not lead to granule formation. The presence of Tgs1 but lack of TMG cap for snRNAs suggests that the snRNP assembly is blocked, probably at the ring formation step. Without TMG caps, snRNPs cannot be imported back into the nuclei, thus accumulating in the cytoplasm. High concentrations of mis-assembled snRNPs further lead to granule formation.

Takata et al. recently showed that knockdown of the integrator complex (INTS4/11) in HeLa



Tissue/cell type	Manipulation	Components	Missing	Reference
Dros. ovary	No	SMN, Sm, Lsm10/11, snRNAs, SPN, SNF		Liu et al., 2007 This study
Dros. follicle cells	No	SMN, Sm, snRNAs (?)		Liu et al., 2007 This study
Dros. testis	No	SMN, Sm	snRNAs, SNF, Lsm10/11	Liu et al., 2007 This study
HeLa cells	INTS4/11 KD	SMN, Sm	snRNAs	Takata et al., 2012
HeLa cells	PHAX KD	SMN, Sm(?)	snRNAs	Takata et al., 2012
HeLa cells	Leptomycin B	SMN, Sm	snRNAs	Takata et al., 2012
HeLa cells	TGS1	SMN, Sm(?)	TMG	Takata et al., 2012
HeLa cells	SMNΔN27 OE	SMN, Sm, snRNAs	TMG	Pellizzoni et al., 1998 Mouaikel et al., 2003
HeLa cells	Unrip KD	SMN, Sm, snRNAs	TMG, SPN	Boysen and Gruss, 2013
HeLa cells	SPN OE	SMN	?	Ospina's thesis 2005
Dros. S2 cells	Sm OE	SMN(?), Sm	?	Our lab

Table 5.1: **Formation of U body like structures by blocking cytoplasmic snRNP assembly.** A summary of related studies. Dros.: D. melanogaster. The highlighted region indicates cases where even snRNAs are missing from the U body like structures.

cells, which processes the 3' end of snRNAs, caused disruption of Cajal bodies (Takata *et al.*, 2012). In addition, knockdown of PHAX, the snRNA export factor, TGS1, the TMG capping enzyme, or inhibition of export using leptomycin B all caused Cajal body defects. Interestingly, all these treatments also result in the formation of cytoplasmic RNP granules containing SMN and Sm proteins, similar to the SMN bodies in primary spermatocytes. Specifically, the cytoplasmic granules formed through PHAX, INTS4/11 knockdown and inhibition of export do not contain snRNAs, whereas the ones formed through TGS1 knockdown do not contain TMG caps. In all of these cases, a subset of the snRNP assembly factors accumulates in the cytoplasm, but they cannot be fully assembled and imported back into the nuclei. These results suggest that blocking the efflux of snRNPs from cytoplasmic snRNP assembly centers lead to granule formation.

Boysen and Gruss recently showed that blocking snRNP assembly by knocking down Unrip, a component of the SMN complex, also lead to cytoplasmic granule formation (unpublished results). Smolinski et al. showed that periodic expression of Sm proteins accompanied formation of nuclear Cajal bodies and cytoplasmic snRNP-rich bodies that are similar to SMN bodies (Smolinski *et al.*, 2011). We also showed that overexpression of Sm proteins, which would lead to imbalance in the ratio of Sm protein over snRNAs, causes formation of cytoplasmic Sm protein containing granules (unpublished results). Even overexpression of Snurportin (SPN) can cause SMN-containing granule formation.

Based on the results summarized here, we propose a model for the formation of RNP granules that function as RNP assembly centers, but not the final destination of RNPs. The RNP assembly centers are nucleated by certain proteins that can form scaffolds through self-oligomerization and they are normally beyond the resolution of light microscopes. When the influx of components overwhelms the efflux of assembled products, either because of direct block of assembly (expression of SMNdeltaN27, Unrip KD and TGS1 KD), lack of certain components (snRNAs in the case of INTS4/11 KD, PHAX KD and treatment using leptomycin B), or oversupply of other components (Sm protein overexpression), the components accumulate to form granules visible under the microscope.

The formation of U bodies in the female germline and SMN bodies in the male germline both may be explained using this model. Since a subset of the snRNPs in the nurse cells are to be transported to the oocytes, the import of assembled snRNPs may be blocked, thus leading to higher concentrations of snRNPs in the cytoplasm, and therefore the formation of U bodies. The block could be caused by the lack of certain unknown missing assembly steps that are required for import, or by factors actively preventing the U bodies from being imported back into the nurse cell nuclei. However, the embryos develop properly from ovaries without U bodies, suggesting that the U bodies *per se* are not essential, rather, the components of these bodies are. This situation is similar to that of the P bodies, which arise as a consequence of, but not required for, gene silencing (Eulalio *et al.*, 2007). In the case of the SMN bodies in the male germline, the lack of snRNAs in the SMN bodies clearly suggest an imbalance in the production of Sm proteins and snRNAs (this should be tested using various methods), which leads to the accumulation of Sm proteins. This imbalance could be because of lack of coordination of transcription and translation of snRNAs and Sm proteins, which might be a regulatory mechanism.

The stages-specific appearance of U bodies and SMN bodies also support the model we proposed. U bodies appear around stage 5 of egg chamber development, which is characterized by the global decondensation of chromatin. Chromatin decondensation is generally associated with active transcription. It is likely that production of snRNP components is dramatically upregulated at stage 5, thus increasing the cytoplasmic concentration of snRNPs and formation of U bodies. Similarly the primary spermatocyte stage of spermatogenesis (not before or after) is also characterized by highly active transcription. There might be a lag in the transcription and processing of snRNAs, compared to the Sm proteins, which could explain why Sm proteins accumulate in granules without snRNAs. These hypotheses should be tested in a more rigorous way, by quantitative analysis of snRNP components in situ through the developmental process of oogenesis and spermatogenesis.

5.4.2 Cytoplasmic SMN-containing granules

Besides U bodies and SMN bodies, several other cytoplasmic RNP granules have also been reported to contain SMN. However, it is not well understood what roles SMN plays in these structures. Stress granules form in cells when they are challenged with stress. Hua et al. reported that overexpression of SMN can sometimes cause formation of stress granules, and SMN colocalizes with known stress granule markers like TIA-1/R and G3BP (Hua and Zhou, 2004b). Neuronal granules have been reported to contain SMN, and the SMN-containing neuronal granules do not colocalize with Sm proteins (Y12 staining), even though Sm proteins are also reported to be present in some neuronal granules (Zhang *et al.*, 2006). However, these observations are different from the SMN bodies and U bodies. More detailed studies are needed to investigate the relationship among these structures.

5.4.3 Comparison between U bodies and P bodies

Comparison between U bodies and P bodies brings interesting insights into the assembly of RNP granules. P bodies are cytoplasmic granules that contain components involved in multiple posttranscriptional processes, such as RNAi-mediated gene silencing, mRNA degradation, translational repression and NMD-mediated mRNA decay. Eulalio et al. showed that even though blocking siRNA or miRNA silencing pathways prevents P body formation, posttranscriptional processes remain functional in cells lacking detectable microscopic P bodies (Eulalio *et al.*, 2007). This is similar to U bodies and other RNP granules. It is thought that many of the RNP granules form to concentrate RNA processing factors in a small volume and increase the efficiency, but these structures *per se* are not really essential for the functions.

We propose that the SMN/U bodies are generally not essential organelles and formation of microscopic granules is a consequence of altered flow in snRNP assembly in the cytoplasm. Interestingly, a similar interpretation has been proposed for the formation of Cajal bodies (Novotny et al. 2013, unpublished results), the nuclear snRNP assembly and maturation centers. Cajal bodies are not present in all cells (Spector *et al.*, 1992; Rajendra *et al.*, 2010). Novotny et al. showed that disruption of tri-snRNP assembly causes incomplete snRNPs to aggregate and form Cajal bodies in cells that do not have them normally.

5.5 Materials and Methods

5.5.1 Fly stocks

Flies (*Drosophila melanogaster*) were raised on standard cornmeal food, supplemented with yeast, maintained at room temperature (22-24C). The following fly strains were used: *Oregon Red* (as wild type), *nos-Gal4 VFP-SmB, nos-Gal4 VFP-SmD3, nos-Gal4 VFP-SmE, da-Gal4 VFP-SmD1* (Gonsalvez *et al.*, 2010), *Smd3pt* (from L. Cooley lab, generated by Quinones-Coello *et al.* (2007b); Gonsalvez *et al.* (2010), *Dart5-1/CyO actin-GFP* (Gonsalvez *et al.*, 2006), *Ago2[51B]/Tm3 Sb GFP* (Xu *et al.*, 2004), *dcr-2[MI00978]* (Wang *et al.* (2006), Bloomingon Stock Center), *r2d2[S165fsX]* (Bloomington Stock Center), *Smn[X7/D] FLAG-Smn*(wt or mut) (Praveen *et al.*, 2012).

5.5.2 Antibodies and probes

The following primary antibodies and dilutions were used for immunofluorescence: SMN (against fly SMN, rabbit polyclonal, 1:400, (Praveen *et al.*, 2012)), SMN (for mouse immunofluorescence, mouse monoclonal, clone 8, BD Biosciences), coilin (R288, rabbit polyclonal, (Andrade *et al.*, 1993)), Y12 (anti-SmB, mouse monoclonal, from Joan Steitz lab), Lsm10 and Lsm11 (rabbit polyclonal, 1:50, Schumperli lab, (Pillai *et al.*, 2003)), TMG (K121, mouse monoclonal, 1:100), Pacman (rabbit polyclonal, 1:100, from Sarah Newbury lab, (Grima *et al.*, 2008)), Fmr1 (mouse monoclonal, 6A15, Developmental Studies Hybridoma Bank), SNF antibody (4G3, mouse monoclonal, 1:100, Helen Salz lab, (Flickinger and Salz, 1994)), PABP2 (rabbit polyclonal, 1:100, (Benoit *et al.*, 1999)), PABP (rabbit polyclonal, 1:100, from Sonenberg lab, (Roy *et al.*, 2004)), Btz (rabbit polyclonal, 1:100, D. St Johnston lab, (van Eeden *et al.*, 2001)), Dlg (Discs large, mouse monoclonal, Developmental Studies Hybridoma Bank), Gemin2 (mouse monoclonal, Garcia et al., unpublished), Y10B (mouse monoclonal, 1:100, anti-rRNA, from Joan Steitz), SPN (rabbit polyclonal, 1:100, (Natalizio and Matera, 2013)). Secondary antibodies were goat anti-mouse IgG or goat anti-rabbit IgG labeled with Alexa 488 or 594 (Life Technologies, 1:200).

5.5.3 Immunofluorescence and in situ hybridizations

Male and female flies were transferred daily to new bottles after eclosion, and 2-4 day old adults were dissected for gonads. Immunofluorescence and in situ hybridizations were performed as described with some modifications (Gonsalvez *et al.*, 2010). For in situ hybridization using oligo-dT probe (Cy5-(dT)20, 26-4420-02 from GeneLink), the prehybridization and hybridization steps were performed at room temperature (22-24C), in PBST (PBS with 0.1% Tween 20). The probe was added at 1ng/ul to the hybridization buffer, and incubated for 2 hours. As a control, unlabeled oligo-(dT)20 was added to the prehybridization buffer at 20ng/ul (20-fold excess) before adding Cy5-(dT)20, in order to compete it off. Note: Antifade (1% DABCO (Sigma-Aldrich) in 90% glycerol and 20mM Tris-HCl pH7.0) is used for normal immunofluorescence and in situ hybridization samples, but NOT suited for oligo-(dT)20 hybridized samples, because it contains 90% glycerol, which lowers the Tm too much for oligo-(dT)20 to bind to mRNAs.

5.5.4 Live imaging of ovaries and particle tracking

Live imaging of fly ovaries was performed essentially as described (Prasad *et al.*, 2007). VFP-tagged Sm proteins were used to mark the U bodies. To mark the cytoplasmic membrane, FM4-64 (T13320, Life Technologies) was applied to the culture media 5 min before imaging began. To analyze the function of microtubules in U body movement, colcemid (D7385, Sigma-Aldrich) was added to the culture media at 20ug/ml 5 min before live imaging.

Acknowledgements We would like to thank Mike and Becky Terns, Joan Steitz, Dan St. Johnston, Fen-Biao Gao, Dan Schumperli, Helen Salz, Martine Simonelig, Nahum Sonenberg, Isabel Palacios and Sarah Newbury for their generous gifts of fly stocks and antibodies. We thank the members of the Matera laboratory for numerous helpful comments and suggestions. This work was supported by NIH grant R01-GM053034 (to A.G.M.).

CHAPTER 6: Conclusions and future directions

Just like chromatin is for DNA, ribonucleoprotein (RNP) complexes are where the message of RNA is carried and executed. Detailed analysis of the molecules and interactions in RNP complexes is the prerequisite for any functional and mechanistic studies about RNA metabolism, and gene expression in general. In this dissertation, I presented my work on the comprehensive analysis of eukaryotic RNP complexes that contain Sm proteins, which are one of the most highly conserved RNA binding protein families. In this chapter, I summarize the methodologies I developed, discoveries I made, and at the same time discuss the implications and future perspectives. Due to the nature of this dissertation, where all the previous chapters are manuscripts that are relatively independent of each other, I organized this chapter as sections, each part discussing one topic.

6.1 RNA-seq applications and analysis

It is not simply curiosity, but also powerful tools that drive scientific discoveries. In the past decades, we have seen many examples on how revolutionary technologies lead to exciting findings, and occasionally, completely new areas of scientific inquiries. For example, the invention of DNA microarrays made it possible, for the first time, for us to study DNA and RNA molecules easily in large scale (Schena *et al.*, 1995). The introduction of massively parallel sequencing technologies further increased our ability to detect nucleic acid sequences to an unprecedented level (Wang *et al.*, 2009b).

Sm proteins are a highly conserved family of RNA binding proteins present in all three domains of life(Valentin-Hansen *et al.*, 2004; Salgado-Garrido *et al.*, 1999). Understanding the functions of the Sm proteins requires the knowledge of their target RNAs. In order to study Sm-containing RNPs, I developed a multi-targeting RIP-seq strategy (Lu *et al.*, 2014). Comparison between the RIP-seq profiles of multiple subunits of the same complex (in this case the Sm subunits) ensures that the consensus set of RNAs are more likely to be *bona fide* targets of the protein complex. The increased variations in the experimental conditions of multi-target RIP-seq, such as the differences in genetic backgrounds, tagging methods and antibodies, help minimize potential artifacts. For the enrichment analysis of pairs of control-IP experiments, I used Gaussian mixture modeling of the enrichment ratios, assuming that there are two populations of RNAs in the IP experiments, ones that stick to the beads non-specifically, and the ones that are bound by the antibody-antigen complexes. I showed that the experimental and analytical procedures are highly reproducible and robust. These strategies can be easily adapted to the analysis of other RNA-binding protein complexes, where proteins cooperate in RNA binding and function.

6.2 New types of Sm-class snRNAs

The Sm class snRNAs include the 9 spliceosomal snRNAs (U1, U2, U4, U5, U6, U11, U12, U4atac and U6atac), the spliced leader snRNAs, U7 snRNA, the yeast telomerase RNA, and viral HSUR snRNAs. These snRNAs all have essential functions in diverse cellular processes. In this dissertation, I report the discovery of a new Sm-class snRNA, LU (Chapter 2)(Lu *et al.*, 2014). The LU snRNA gene has a normal snRNA like pol II promoter, and the transcript has an snRNA-like TMG cap, is localized to the nucleus (though further in situ hybridizations should be carried out to determine whether it is localized to the Cajal bodies or not).

The LU snRNA is only present in a handful of sequenced species in the melanogaster group of *Drosophilids*, however, it is highly conserved among them. Sequence alignment showed that many complementary mutations already occurred within about 15 million years of its evolution. The lack of clear sequence homology to the snRNAs suggests that during the initial evolution it may have undergone a phase of rapid mutations, and the product was then fixed due to strong functional constraints. The LU snRNA may serve as a good example for studying the origination and evolution of ncRNA genes.

While our genetic analysis suggests that the LU snRNA is not essential for survival, it nevertheless may have important functions at the molecular level. Preliminary examination revealed a potential base pairing region between LU snRNA and the 5' end of U6 snRNA, suggesting that it may regulate splicing. Direct crosslinking experiments using psoralen should be performed on LU and U6 snRNAs to test the base pairing interaction. Transcriptome analysis of LU snRNA mutant showed that the expression of many genes are altered, however, it is still hard to identify the direct consequences of the loss of LU snRNA (see Appendix C).

6.3 Sm-associated scaRNAs

Small Cajal-body specific RNAs (scaRNAs) localize to the Cajal bodies and play important roles in guiding 2'-O-methylation and pseudouridylation of snRNAs (Richard *et al.*, 2003; Jady *et al.*, 2004). While the physical location and the functional relevance to snRNP biogenesis may make it seem not surprising that Sm proteins associate with scaRNAs (Cajal bodies are enriched in Sm proteins, and snRNPs contain Sm proteins), almost nothing is known about the interaction between these two important classes of molecules, Sm proteins and scaRNAs. (Even though the yeast telomerase RNA also localizes to the Cajal body and directly binds Sm proteins, we do not include is in this section because it is not a guide RNA for snRNA modifications (Seto *et al.*, 1999; Jady *et al.*, 2004).)

We identified a number of scaRNAs in flies and human cells that associate with Sm proteins (Chapter 2). Our results suggest the entire Sm ring may associate with the scaRNAs, however, not all scaRNAs are associated with Sm proteins. We failed to identify any Sm site like sequences in the scaRNAs, or any other sequence elements that could distinguish the Sm-associated vs. non-Sm-associated scaRNAs. Even though Fu and Collins showed evidence that the CAB box of scaRNAs bind Sm proteins directly, this has not been reproduced, and later experiments by Steitz and colleagues suggested that the CAB box actually binds WDR79 directly (Fu and Collins, 2006; Tycowski *et al.*, 2009).

For future research, it would be important to first establish whether Sm proteins bind scaRNAs directly using crosslinking assays (using 4-thiouridine or conventional UV crosslinking), keeping in mind that the interaction may be refractory to crosslinking even if they are within crosslinking radius. An alternative situation would be an indirect interaction mediated by core proteins of the scaRNPs. Once binding sites are identified, we could test the functional implications of the interaction by mutating the binding interface, either on the scaRNA or on the protein-protein interaction interfaces.

6.4 snRNP-mRNA interactions

The most surprising finding in our RIP-seq experiments is the identification of large number of mature mRNAs associated with Sm proteins. We further showed that the interactions are mediated, in part, by snRNPs. Specifically, we showed that the 5' end splice site recognition sequence is required for U1 snRNP to bind the CG3776 mature mRNA in the middle of the second exon. This interaction, 12 perfect base pairs, is different from pre-mRNA-spliceosome interactions, which is weaker and more transient (for good reasons, since the snRNPs and pre-mRNAs have to be remodeled dynamically, and thus a tight interaction may become a energetic trap (Freund *et al.*, 2005)). We showed that splicing is not required for snRNP binding to mature mRNAs, since some of the targets do not have introns, and removing the introns from CG3776 does not affect the interaction.

We showed that there is no overlap between the mRNA targets of Sm proteins in fly ovaries and human HeLa cells. This is consistent with other studies showing that regulatory relationships evolve rapidly, since the binding sites can be easily gained or lost after simple mutations. One prominent example is the histone mRNAs. We identified human, but not fly, mature replication-dependent histone mRNAs as Sm-associated. More interestingly, Friend et al. already reported that human pre-mature histone mRNAs are bound by U2 snRNPs, which contain Sm proteins (Friend *et al.*, 2007). It is very likely that the Sm-histone mRNA interaction is also mediated by U2 snRNPs. The difference between these two reports is that U2 binds pre-mature or mature histone mRNAs. We argue that since the level of mature histone mRNAs are higher and sequencing pre-mature histone mRNAs might be hindered by the presence of the 3' end stem loop, it is easier to identify the mature mRNAs as targets. In addition, the identification of mature histone mRNAs as targets of Sm proteins/U2 snRNPs suggest that U2 snRNPs have additional functions beyond stimulating U7-based processing of pre-histone mRNAs. U2 snRNPs may play a role in subsequent translation and stability of histone mRNAs.

Based on the results presented in Chapter 2, I propose a model that the spliceosomal snRNPs regulate different subsets of mRNAs. This model is further supported by the Lsm11 RIP-seq results shown in Appendix B. Lsm11 RIP-seq in *Drosophila* ovaries surprisingly identified U11 and U12, and many mRNAs. Some of these mRNAs overlap with the ones associated with canonical Sm proteins, while others do not. Even though these results need further validation, they suggests the possibility that Lsm11 containing snRNPs may associate distinct sets of mRNAs to regulate their metabolism.

Even though we do not know the function of the snRNPs in binding mature mRNAs, identifying the binding sites (e.g. in CG3776 mRNA) makes it possible, for the first time, to study these functions. To study the function of the snRNP-mature mRNA interaction, we could mutate the snRNP binding sites on the mRNAs, without affecting the coded protein sequence (already achieved using the mutations in Figure 2.21 in Chapter 2). Mutating the mRNAs is better than mutating the snRNAs since snRNAs are essential for splicing. Various potential effects can be examined after mutating the binding sites, such as mRNA stability, translation, etc.

6.5 Developing new methods to study snRNP-mRNA interactions

Even though we reproducibly identified mRNAs associated with Sm proteins/snRNPs, we could not reliably identify the binding sites using the methods we developed. The identification of interaction interfaces in this unique type of interaction requires new methods. Here I propose two strategies for the identification of snRNA-mRNA interaction sites and also for studying general RNA-RNA interactions.

Psoralen is a planar molecule that can be intercalated between stacking base pairs in double stranded nucleic acids (Calvet and Pederson, 1979, 1981). UVB (365nm) irradiation of psoralen-nucleic acid complexes crosslinks the two strands of the nucleic acids at sites of intercalation. Derivatives of psoralen have been used to study RNA-RNA, RNA-DNA and DNA-DNA double helixes (Shen *et al.*, 1977; Calvet and Pederson, 1979, 1981; Thompson and Hearst, 1983; Lipson and Hearst, 1988; Skripkin *et al.*, 1996; Sastry *et al.*, 1997). It is known that several of the snRNPs base pair, or have the potential to base pair with mRNAs (or even other types of RNAs) (Matera *et al.*, 2007; Matera and Wang, 2014). For example U1, U2, U5, U6, U11, U12, U6atac, and U7 have all been shown to base pair with pre-mRNAs during certain pre-mRNAs processing steps. To identify the targets of snRNPs with nucleotide resolution, we could culture cells in the presence of psoralen, perform UV crosslinking, limited RNase treatment, Sm protein immunoprecipitation and reversal of crosslinking. The immunopurified RNA fragments should be the regions that directly base pair with snRNPs. The Sm protein IPs can be substituted with IPs against snRNP-specific proteins (Figure 1.7), and then the fragments purified would be the ones that associate with that particular snRNP. The use of crosslinking enables identification of both stable and transient base pairing interactions between snRNPs and mRNAs. In addition, if we omit the crosslinking step, we could identify the RNA fragments that bind snRNPs very tightly, like the one we identified in the RIP-seq analysis presented in Chapter 2. This method could be extended using antibodies against proteins of other ncRNPs to study the RNA targets regulated by the ncRNPs.

Proximity ligation is a powerful strategy for studying inter- or intra-molecular interactions for both DNA and RNA (see section 1.11 in Chapter 1 on a brief review about methods for studying DNA-RNA-protein interactions). The chromatin conformation capture (3C) method developed by Dekker et al. is a good example of this strategy (Dekker *et al.*, 2002). Recently, Tollervey and colleagues developed a similar method, called CLASH (crosslinking, ligation and sequencing of hybrids) (Granneman *et al.*, 2009; Kudla *et al.*, 2011; Helwak *et al.*, 2013; Helwak and Tollervey, 2014). However, most of the studies on RNA-RNA interactions have been focused on snoRNA/rRNA, scaRNA/snRNA and miRNA/mRNA interactions because these are the well-known interactions that rely on base pairing. Combining the CLASH method with Sm protein or snRNP specific protein IPs would make it possible to identify the binding sites of snRNPs on mRNAs in a global scale with nucleotide resolution.

Extending upon the proximity ligation and the CLASH techniques, I also propose to develop a method for global analysis of inter- or intramolecular base pairing interactions for the entire transcriptome. Briefly, we can culture cells in the presence of psoralen, perform UV crosslinking, lyse cells, treat lysate with RNase, dilute the lysate and perform proximity ligation, reverse crosslinking, remove proteins using proteases, and sequence the hybrids. No IP is needed for this procedure, but we can deplete the rRNAs, in order to increase the coverage other RNAs. Using this method, we can identify previously unknown RNA-RNA interactions. Since the ligation efficiency is not high for RNA-RNA ligations, we need to sequence the libraries much more deeply in order to identify interactions between lowly expressed RNAs.

6.6 Vicinal mapping of abnormal reads

Since the development of massively parallel sequencing technologies, many experimental variations and analysis methods have been introduced for a variety of purposes. RNA-seq data contain rich information, not only about the expression level of RNA species, but also about their sequence variations, like editing, alternative transcription start site, alternative splicing and alternative polyadenylation. In addition, other kinds of information can also be obtained through the analysis of the data (see section 1.12 in Chapter 1 about the different kinds of 'abnormal' RNA-seq reads and how these can be used to study particular biological questions).

While analyzing the RIP-seq data presented in Chapter 2, we noticed that some reads are chimeric. These reads are likely generated through the self-priming of the 3' end and ligation of the 5' ends of ncRNAs that have terminal stemloops. Even though these reads are artifacts of library preparation, they contain information about the precise ends of the ncRNAs. Therefore, we developed a computational method for the identification of these chimeric reads from a large number of RNA-seq datasets, and used these data to identify the ends of numerous ncRNAs (Chapter 2)(Lu and Matera, 2014). The analysis of 'abnormal reads' has attracted more attention in recent years. More and more bioinformatics methods are being developed to extract these reads and study the biology behind them.

6.7 snRNA functions in alternative splicing

Gene duplication is a widespread phenomenon in almost all life forms, and provides a source for evolutionary innovation. The duplicated genes create redundancy and in certain situations the redundant copies become pseudogenes and lost, whereas in other situations, these redundant copies are needed to provide larger amounts of transcripts, or acquire new expression patterns and functions. It has been known for a long time that there are multiple copies of each of the major spliceosomal snRNA genes in higher eukaryotes. However, it is not clear whether they only provide higher transcriptional activity or have divergent functions.

In Chapter 4, I presented a comprehensive analysis of the expression of snRNA paralogs during *Drosophila* development, and summarized previous research on snRNA expression profiles. Our analysis showed that despite the fact that the snRNA ortholog groups are not stable in evolution, the changes in dominance of specific snRNA paralogs are conserved. More isoforms are expressed in early stages of development, whereas in later stages of development, one isoform typically take over. Even though we do not know why this happens, our results make it possible to use genetics to study the functions of snRNA isoforms during development. Many P element insertions are available for *Drosophila* snRNA genes, and detailed analysis of the defects of these mutants in the future would be important for understanding the functions of snRNA paralogs. In order to facilitate further studies of these snRNA isoforms in flies, I compiled a list of available *Drosophila* mutants, shown in Table 6.1. To illustrate the usefulness of these expression profile data, our lab recently showed that an U4:39B allele is embryonically lethal, consistent with the expression profile data that U4:39B is the major U4 isoform. It remains unclear whether the U4:38AB isoform is essential and where and when is it required during development.

6.8 snRNP assembly and RNP granules

Subcellular structures, also known as organelles, compartmentalize the cell, making it possible to arrange all the complex chemical reactions in a organized manner. RNA-protein aggregates are a distinct class of organelles that have been suggested to play important roles in gene regulation (some of which are still in question). Mechanisms of formation and functions of these structures are hard to study because of the difficulty in establishing causative relationships between the components of the structure and the structure itself.

In Chapter 5, I presented some detailed analysis of a specific class of RNP granules involved in snRNP biogenesis, the U body and SMN body. Since the initial report in 2007, rigorous analysis of the U bodies has been lacking (Liu and Gall, 2007). It is unclear whether U bodies are ubiquitous and what kind of function they have. Our comprehensive analysis of SMN/Sm/snRNA stainings in various cell types suggest that the U body is not present in most

snRNA	Allele	Mutation	Balanced	Stock
U1:21D	y^{1} w^{67c23}; P{EPgy2}snRNA:U1:21D^{EY19137}	Promoter	No	BDSC 22300
U1:82Eb	P{XP}snRNA:U1:82Eb^{d09954}	Transcript	No	Exelixis
U1:82Eb	w^{1118}; P{SUPor-P}snRNA:U1:82Eb^{KG00155} ry^{506}/ TM3, Sb^{1} Ser^{1}	Transcript	Yes	BDSC 13644
U1:82Eb	y^{1} w^{67c23}; P{EPgy2}snRNA:U1:82Eb^{EY04087}	Transcript	No	BDSC 15716
U1:82Eb	y* w*; P{GawB}snRNA:U1:82Eb^{NP1591} / TM6, P{UAS-lacZ.UW23-1}UW23-1	Transcript	Yes	DGRC 112719
U1:95Ca	w^{1118}; P{EP}snRNA:U1:95Ca^{G14596}	Transcript	No	BDSC 30205
U1:95Cb	NA			
U1:95Cc	NA			
U2:14B	NA			
U2:34ABa	y^{1} w^{67c23}; P{EPgy2}snRNA:U2:34ABa^{EY07636}	Transcript	No	BDSC 16413
U2:34ABb	y^{1} w*; P{EP}snRNA:U2:34ABb^{G2309}	Transcript	No	BDSC 26985
U2:34ABc	w^{1118}; P{EP}snRNA:U2:34ABc^{EP850}	Transcript	No	DGRC 122029
U2:34ABc	y^{1} w^{67c23}; P{SUPor-P}snRNA:U2:34ABc^{KG0762}	Transcript	No	BDSC 14545
U2:38ABa	NA			
U2:38ABb	y^{1} w^{67c23}; P{SUPor-P}snRNA:U2:38ABb^{KG05695}	Transcript	No	BDSC 14451
U4:25F	NA			
U4:38AB	NA			
U4:39B	y^{1} w^{67c23}; P{lacW}snRNA:U4:39B^{k09410} CG8678^{k09410}/CyO	Transcript	Yes	BDSC 10879
U4:39B	y^{1} w^{67c23} ; P{lacW}snRNA:U4:39B^{k09414} / CyO	NA	Yes	DGRC 102708
U4:39B	y^{d2} w^{1118} P{ey-FLP.N}2 P{GMR-lacZ.C(38.1)} TPN1; P{lacW}snRNA:U4:39B^{k09410} P{neoFRT} 40A/CyO y^{+}	Transcript	Yes	DGRC 111260
U4:39B	y* w*; P{GawB}snRNA:U4:39B^{NP0252} / TM6, P{UAS-lacZ.UW23-1}UW23-1	Transcript	Yes	DGRC 112110
U4:39B	w^{1118}; P{RS3}snRNA:U4:39B$^{CB-0586-3}$	Transcript	No	DGRC 123246
U5:14B	NA			
U5:23D	NA			
U5:34A	y^{1} w^{67c23}; P{EPgy2}snRNA:U5:34A^{EY04189}/SM6a	Transcript	Yes	BDSC 20070
U5:35D	y^{1} w*; P{EP}snRNA:U5:35D^{G3078}	Transcript	No	BDSC 27468
U5:38ABa	y^{1} w^{67c23}; P{SUPor-P}snRNA:U5:38ABa^{KG09823}	Transcript	No	BDSC 14791
U5:38ABb	NA			
U5:63BC	NA			

Table 6.1: **Available *Drosophila* snRNA mutants**. NA: information not available.

cell types. Our live imaging analysis of the U bodies showed that they are transported from the nurse cells to the oocytes through ring canals during oogenesis. The supply of snRNPs to growing oocytes is normal in the absence of the U bodies, suggesting that transport of snRNPs is not an essential function of the U bodies.

The only other tissue where we reliably observe cytoplasmic aggregates of snRNP assembly components is the *Drosophila* testis. Very surprisingly, these RNP granules do not contain snRNAs. Various anecdotal reports of U body-like structures have been published in the past two decades (Table 5.1). Our examination of these reports lead to a unified theory for explaining the formation of these structures. We believe the formation of these RNP granules is a consequence of the altered flow of components through the assembly machinery. The molecular interactions among the snRNP assembly components, especially the oligomerization property of SMN is the prerequisite of the granule formation.

The framework of this project is already established, including many of the critical experiments, however, more experiments are still needed to support the conclusions. A number of general directions are listed below, with the purpose of finishing this paper and for further studies (more detailed experiments are listed within Chapter 5). (1). The follicle cells on the surface of egg chambers also seem to have U body (or SMN body) like structures. However, it is not clear whether they are real U bodies. Further immunofluorescence studies will be needed to determine the components localized in these granules. It would be interesting to see whether these granules form as a consequence of blocking the snRNP assembly pathway at certain steps. (2). The other tissue that may have U body (or SMN body) like structures is the larval brain, as shown in the paper by Liu and Gall (Liu *et al.*, 2006). Further immunofluorescence studies are needed to show whether these are U bodies or SMN bodies, and what components are there. Similar to the primary spermatocytes and follicle cells, these may also form as a consequence of altered snRNP assembly flow. (3). The formation of U bodies in egg chambers and spermatocytes are stage-specific, it would be interesting to see what kinds of factors really determine the granule formation. This involves analysis of the temporal expression of the snRNP assembly components in these tissues and could be technically challenging.

6.9 Concluding remarks

Structure and function of RNA has always been an important subject of studies since the beginning of the molecular biology era. Recent investigations have benefited a lot from new technologies. In this dissertation, I developed new experimental and computational methods for studying various aspects of Sm protein containing RNP complexes, including their composition, interaction, structure, expression, assembly and localization etc. As any other explorative research, these studies laid a solid foundation for further mechanistic studies, and opened up more questions than providing answers. Hopefully the work presented here has been a tiny bit of contribution to the collection of human knowledge, and provided some inspiration to other scientists.

APPENDIX A: Characterization of SmD3 antibodies

A.1 Rationale

Antibodies are important reagents for research. Many antibodies have been developed or identified for Sm proteins, however, most of them recognize SmB, and few of them specifically recognize other Sm proteins (Fury *et al.*, 1999). In order to facilitate the study of Sm proteins, it is important to characterize the specificity of known Sm protein antibodies and find ones that are good for other Sm proteins.

A.2 Results and Discussion

Previously Dr. T. K. Rajendra suggested that the mouse monoclonal KSm4 antibody specifically recognizes SmD3 in flies, but this was not documented and was not reproduced by Dr. Graydon Gonsalvez. Fury et al. reported that KSm4 mainly recognizes human SmB and SmD1/D2 (Fury *et al.*, 1999). In order to test the specificity of KSm4 and resolve the discrepancy, I performed western blots on lysates from flies expressing VFP-tagged Sm proteins (Figure A.1). KSm4 mainly recognizes SmD3 (the 17kD bands) in flies, and SmB (the 25kD bands), to a lesser extent. The KSm4 antibody was stored in hybridoma cell culture supernatant, presumably not concentrated, at 4C. I tested several different dilutions of the KSm4 antibody, 1:10, 1:50, 1:200 and 1:1000 (the last one not shown). All of the dilutions worked well for the western blot. The 1:1000 dilution may not work well if the amount of protein loaded is low.

The specificity of KSm4 is the opposite of the Y12 antibody identified by Dr. Joan Steitz (Lerner and Steitz, 1979). Y12 antibody mainly recognizes SmB, and to a lesser extent, SmD3. Our result is different from the Fury et al. result, and it could be caused by the difference between fly and human Sm protein sequences. KSm4 is currently the only known monoclonal antibody that specifically recognizes the SmD3 protein.

I also tried immunofluorescence and immunoprecipitation experiments using the KSm4 antibody, however, none of these applications worked. A different rabbit polyclonal antibody targeting SmD3 developed by the Gonsalvez lab works really well for immunofluorescence, but

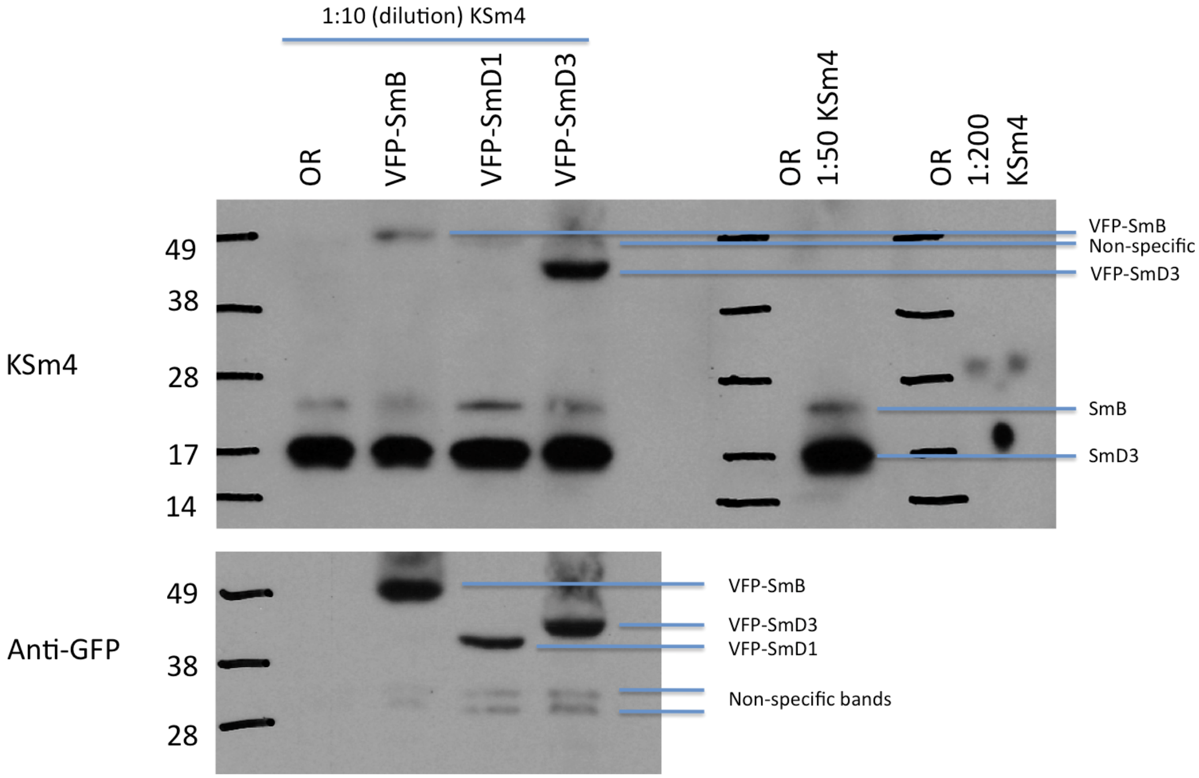


Figure A.1: **Characterization of the KSm4 antibody.** The KSm antibody was used at the dilutions as indicated in the figure. Polyclonal GFP antibody was also used to show the presence of VFP-Sm transgene expression. Molecular weight markers in kD were labeled on the left side of the blots. Lysates were prepared from Oregon R (OR), *nos-Gal4, VFP-SmB* (VFP-SmB), *da-Gal4 VFP-SmD1* (VFP-SmD1), and *nos-Gal4 VFP-SmD3* (VFP-SmD3) flies, and 50 ug total protein was loaded to each lane. Note: The rightmost lane was not run properly.

not for western blot or immunoprecipitation (I tested IP using the polyclonal SmD3 antibody, data not shown).

The original motivation for testing the polyclonal anti-SmD3 antibody was to see if it colocalize with some of the Sm-associated mRNAs as described in Chapter 2. Immunofluorescence using the anti-SmD3 antibody showed that it localizes mostly to the cortex of the oocytes, whereas nuclear staining is very faint (Figures A.2 and A.3). The oocyte staining of anti-SmD3 appears very early, at least in stage 3. The anti-SmD3 positive signal initially looks like granules, then these granules seem to disappear and signal appears around the cortex starting in stage 6 or 7, and persists until very late in oogenesis (maybe beyond stage 10 and 11). This is in contrast to the known primary nuclear localization of Sm proteins. It is possible that a specific subset of SmD3 molecules exist in a unique conformation that is recognized by the poly-clonal anti-SmD3 antibody. However, we do not have any evidence for that. Further more, the localization of anti-SmD3 signal is very close to actin (Figure A.3).

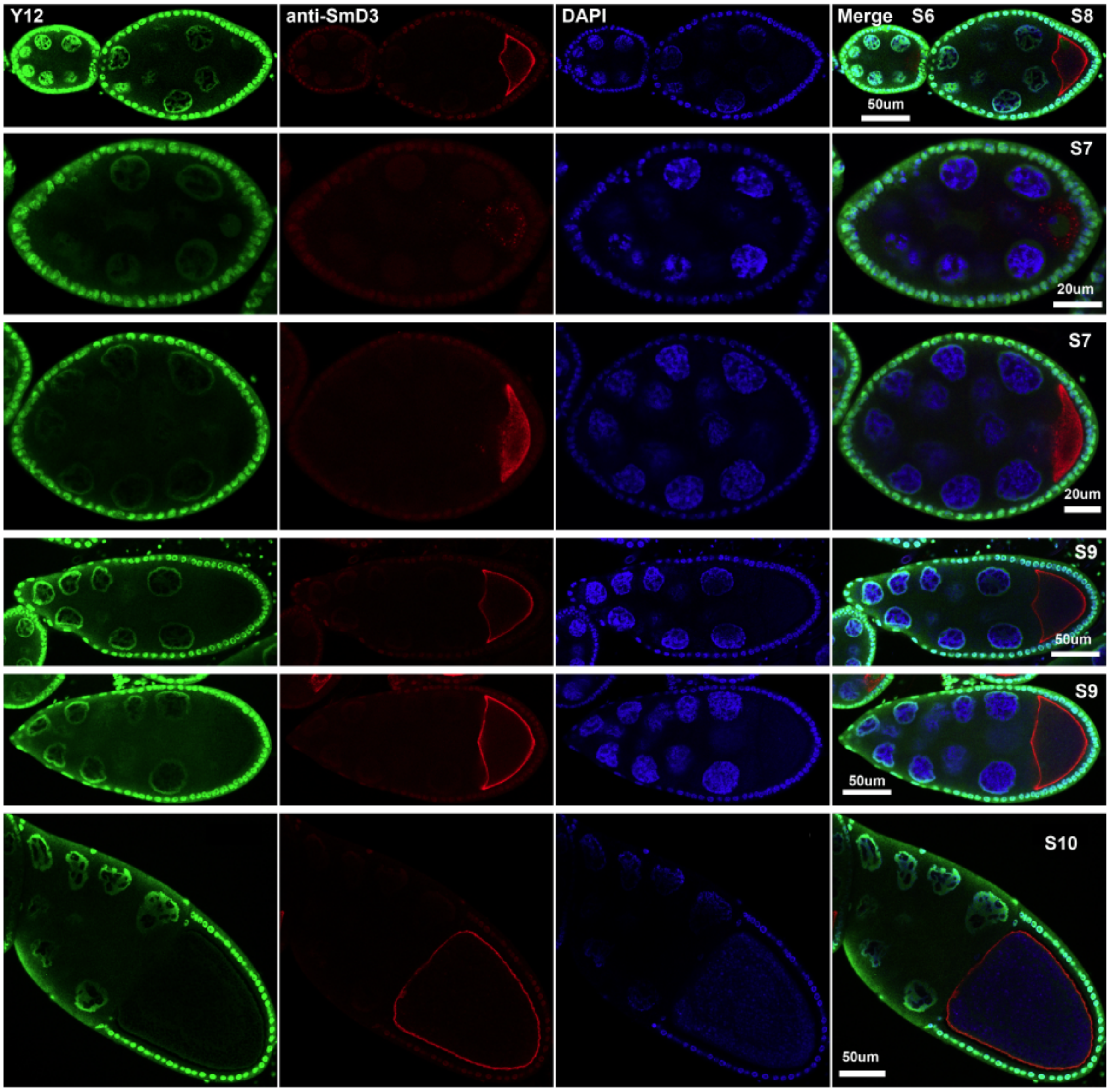


Figure A.2: **Localization of SmD3 in *Drosophila* egg chambers using the polyclonal anti-SmD3 antibody**. Immunofluorescence was performed on Oregon R (wild type) fly ovaries using the Y12 (monoclonal anti-SmB) and polyclonal rabbit anti-SmD3 antibodies. DNA was stained using DAPI. Stages of the egg chambers were determined based on length, relative size of oocytes, position of follicle cells etc.

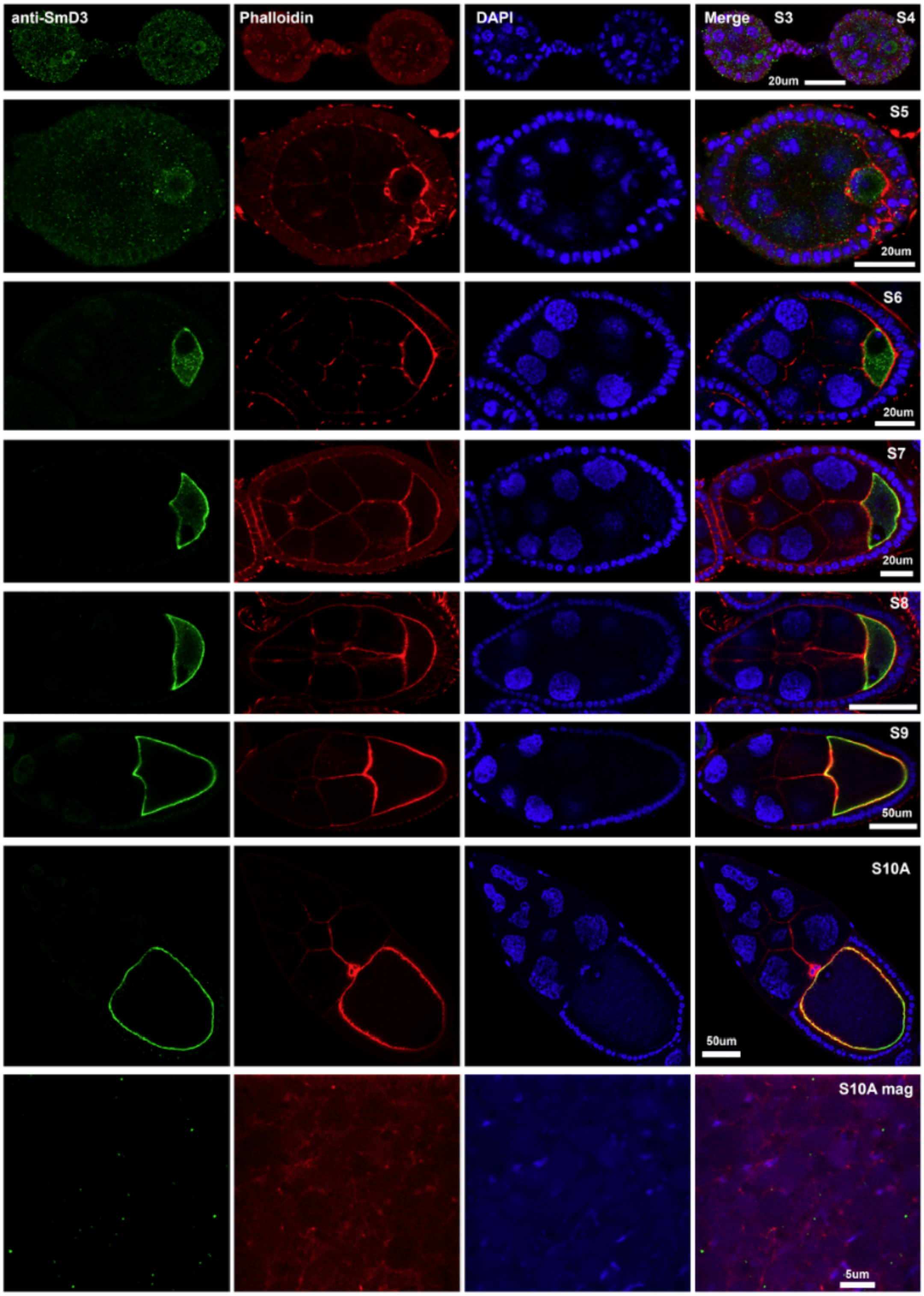


Figure A.3: **Colocalization of anti-SmD3 and actin.**

APPENDIX B: RIP-seq analysis of Lsm11 in *Drosophila*

B.1 Introduction

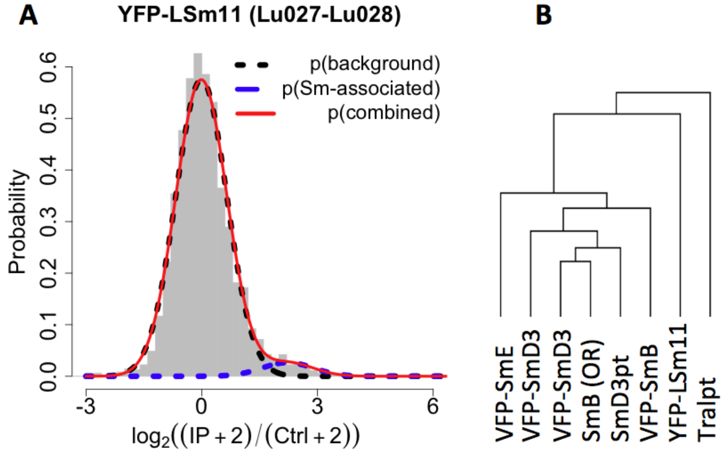
In order to comprehensively identify the targets of Sm-associated RNAs, I performed RIP-seq and RIP-PCR on several Sm class proteins, including SmB, SmD1, SmD3, SmE, Lsm11 and Tral. All of these data have been presented in the RIP-seq paper (Chapter 2), except for Lsm11. Lsm11 is a unique subunit of the U7 snRNP core, where Lsm10 and Lsm11 replace SmD1 and SmD2. The presence of Lsm10/Lsm11 in U7 snRNP confers distinct properties, which are partially responsible for association with FLASH () (papers from Marzluff lab), and participation in histone pre-mRNA processing.

Previously, the Duronio lab showed that *Drosophila* mutants for Lsm10/Lsm11 and U7 snRNA have distinct phenotypes (Godfrey *et al.*, 2009). Whereas U7 null mutant is viable, Lsm10 and Lsm11 mutants do not survive to adulthood. Surprisingly, there is no difference in histone mRNA processing among the mutants. These data suggest that Lsm10/Lsm11 has an essential function that is distinct from histone pre-mRNA processing and independent of U7 snRNA.

We hypothesized that the unique Lsm10-11 ring binds a set of RNAs different from the ones bound by the canonical Sm ring. However, since five Sm proteins are shared between the two rings, the RNA targets of the canonical and Lsm11 proteins identified by RIP-seq should partially overlap. Identifying the RNA targets would provide clues about the potential new functions of Lsm10/Lsm11 dimer.

B.2 Materials and Methods

The YFP-Lsm11 transgene was driven by nos-Gal4, expressed in the ovaries. The transgenic fly line was provided by the Gall lab (Liu and Gall, 2007). RIP-seq experiments and analysis was performed essentially the same as described in Chapter 2. These two libraries were prepared using random hexamers only.



C

Compare with	Common (with ncRNAs)	Common (without ncRNAs)
VFP-SmD3	22/202	13/187
VFP-SmD3	14/130	6/117
VFP-SmE	10/94	6/84
VFP-SmB	28/505	19/489
Tralpt	5/200	5/195
SmB (OR)	21/217	14/204
SmD3pt	15/71	6/56
Consensus	14/86	6/72

Figure B.1: **Analysis of Lsm11 RIP-seq data.** (A). Gaussian mixture modeling of Lsm11 RIP-seq. A total of 293 RNAs are defined as Lsm11-associated based on Gaussian mixture modeling. (B). Hierarchical clustering of all 8 RIP-seq datasets. RIP-seq profile of YFP-Lsm11 clusters outside of the canonical Sm proteins, but closer to canonical Sm proteins than Tralpt. (C). Comparison of targets identified by RIP-seq between YFP-Lsm11 and other Sm proteins. The second column is the number of overlapped RNAs over number of RNAs identified in that particular RIP-seq, including the ncRNAs. The third column is the same as the second column, except that the ncRNAs are excluded. The consensus set of RNAs are defined as enriched in at least 4 out of the 6 canonical Sm protein IPs (see Chapter 2 for details).

B.3 Results and Discussion

I applied Gaussian mixture modeling on the RIP-seq data for Lsm11 IP (Figure B.1A). As expected, two Gaussian functions converged and fit the distribution of ratios nicely. This analysis defined the top 293 RNAs as Lsm11-associated (Table B.1). Note that we were unable to assess the reproducibility of this experiment since only a single pair of Ctrl/IP was performed. When looking at the enrichment ratios of the enriched RNAs, the number of reads should also be considered.

Next, I compared the RIP-seq profile of Lsm11 to those of other Sm proteins presented in Chapter 2, using hierarchical clustering (Figure B.1B). Not surprisingly, Lsm11 clustered outside of the canonical Sm proteins, because Lsm11 associate with a subpopulation of Sm proteins, and the level of Lsm11 is much lower than the canonical Sm proteins. Lsm11 clustered closer than Tralpt, to the Sm proteins, consistent with the fact that Tralpt is now known to participate in Sm ring formation.

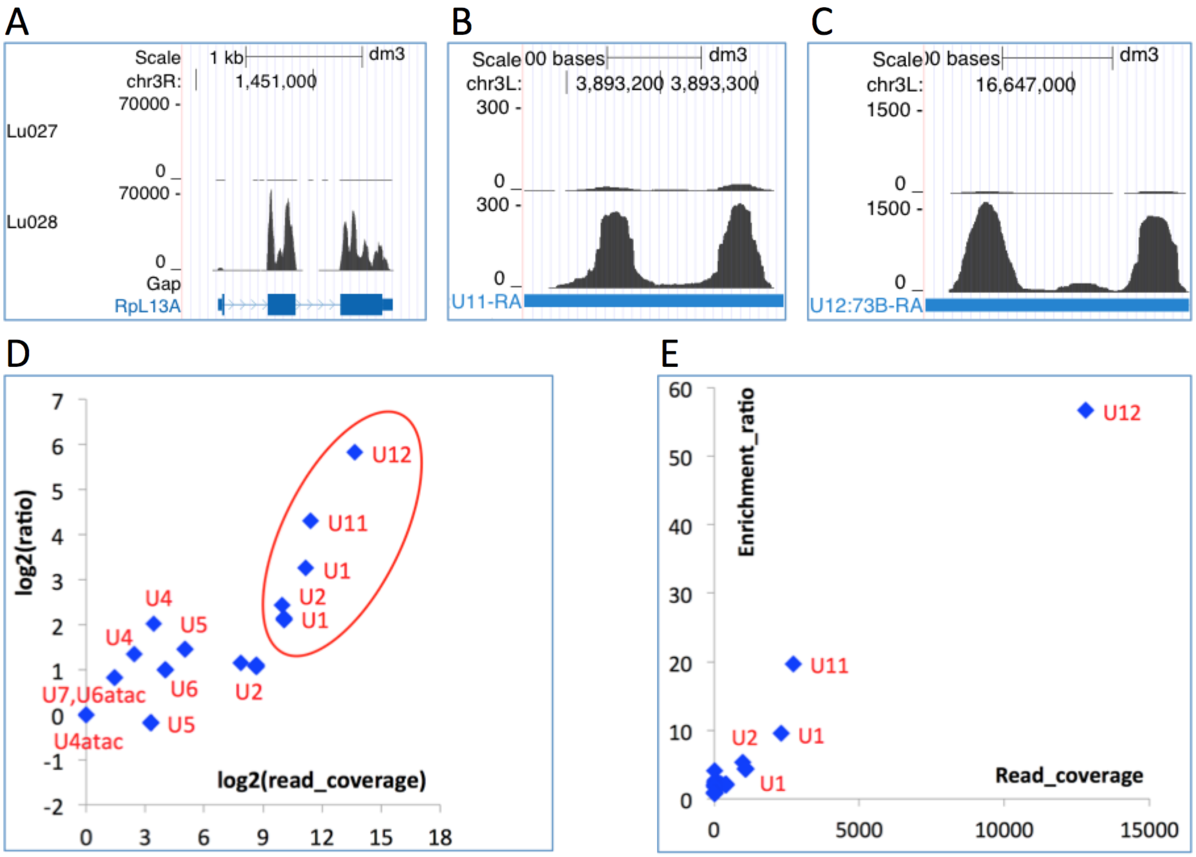


Figure B.2: **Examples of Lsm11-associated RNAs.** (A). The most highly enriched RNA in Lsm11 IP is RpL13A (CG1475). This RNA is also highly enriched in the canonical Sm RIP-seqs. Lu027 is the control; Lu028 is the GFP IP. (B,C). Enrichment of U11 and U12 snRNAs. (D,E). Enrichment of all spliceosomal snRNAs in the Lsm11 IP, plotting enrichment ratios against relative read coverage, in log scale (D) and normal scale (E).

I further examined the overlap between Lsm11 targets and the targets of other Sm proteins (Figure B.1C). Interestingly, a small number of RNAs are immunoprecipitated by the SmB, SmD3, SmE and Lsm11 (14/86 and 6/72 for comparison with the consensus set). The 6 common mRNAs are: CG12173, CG7883, CG1475, CG17531, CG30105, CG3552. This data suggests that these RNAs associate with the Lsm10-11 ring. The genome browser tracks for CG1475, which is enriched around 150 fold, were shown in Figure B.2A. Not surprisingly, the Lsm11-associated mRNAs are also not pre-mRNAs. This interaction could be direct binding of Lsm10-11 ring to the mRNAs, or U7 snRNP to the mRNAs. But considering the low level of U7 snRNP and the high enrichment of these mRNAs in cells, it is more likely that the association is direct.

Another surprising finding is the identification of U11 and U12 snRNAs in the Lsm11 IP (Figure B.2B,C,D and E). U11 and U12 are known to associate with the canonical Sm ring, that contains SmB, D1, D2, D3, E, F and G. Our data, if not an artifact of Lsm11 overexpression, suggests that the Lsm10-11 ring can also bind U11 and U12. This finding could potentially explain the discrepancies between fly mutants for U7 and Lsm10/11.

Further studies would be needed to verify the association between Lsm10/11 and the RNA targets identified in this RIP-seq. Antibodies against endogenously expressed Lsm10 and Lsm11 should be used, to rule out the possibility of artifacts generated by overexpression.

If the interaction between Lsm10-11 ring with mRNAs is confirmed, whether direct or indirect, this would further support the hypothesis that eukaryotic Sm proteins and potentially snRNPs bind distinct subsets of mature mRNAs to regulate their metabolism.

The potentially direct interaction between Lsm10-11 ring with mRNAs, and the potential interaction between Lsm10-11 ring with U11 and U12 snRNAs are two prime candidates for explaining the discrepancies between phenotypes of fly mutants for Lsm10/11 proteins and U7 snRNAs.

#GID	Lu027_N	Lu028_N	ratio_N	#GID	Lu027_N	Lu028_N	ratio_N	#GID	Lu027_N	Lu028_N	ratio_N
CG1475	7819	1166254	149.1	CG9809	3420	25847	7.6	CG2330	0	9	5.6
CG7067	88	7774	86.3	CG12919	2	26	7.4	CG6186	0	9	5.6
CR32162	128	7344	56.7	CG8472	1454	10588	7.3	CG10469	0	9	5.6
CG9200	32	1896	55.2	CG30116	0	12	7.2	CG12814	0	9	5.6
CG12924	8826	455077	51.6	CG42338	3	36	7.2	CG14964	0	9	5.6
CG10655	129	4679	35.6	CG13506	0	12	7.1	CG42253	0	9	5.6
CR31400	410	11143	27.1	CG6719	334	2393	7.1	CG12201	0	9	5.5
CG5826	180	4824	26.6	CG1098	719	5090	7.1	CG12398	0	9	5.5
CG3218	1770	37637	21.2	CG8908	0	12	7.1	CG4859	0	9	5.5
CR34151	77	1563	19.7	CG8161	140	1000	7.1	CG40293	14	88	5.5
CG42231	23	483	19.2	CR34680	2	24	7.0	CG8050	4	29	5.5
CG15394	0	34	17.8	CG7828	124	865	6.9	CG32019	20	119	5.5
CG9250	59	1083	17.7	CG16734	20	146	6.8	CG14463	59	332	5.5
CG33196	0	32	17.0	CG4764	173	1161	6.7	CR32913	103	562	5.4
CG5820	29	488	16.0	CG12075	50	345	6.6	CG14230	435	2350	5.4
CG9958	34	523	14.5	CG4210	13	93	6.5	CR33946	2	18	5.3
CG9960	34	523	14.5	CG7913	785	5068	6.4	CG31868	248	1333	5.3
CG10360	983	13238	13.4	CG17839	0	11	6.4	CG8740	2	18	5.3
CG14285	0	24	13.2	CG42315	0	11	6.4	CG8025	58	314	5.3
CG3710	541	6446	11.9	CG18304	0	11	6.4	CG7762	552	2922	5.3
CG1750	4	64	11.8	CG14655	0	11	6.4	CG13607	4	27	5.2
CG3679	762	9002	11.8	CG12581	4	33	6.4	CG1683	9	55	5.2
CG9249	11	148	11.7	CG7463	0	11	6.4	CG8533	9	55	5.2
CG8260	0	21	11.7	CG1488	0	11	6.4	CG10988	194	981	5.0
CG15105	0	18	10.2	CG14864	0	11	6.4	CG11387	5	35	5.0
CG32744	12467	126140	10.1	CG5612	0	11	6.3	CG9138	0	8	5.0
CG10091	4	53	9.9	CG42492	0	11	6.3	CG32448	2	17	4.9
CG7602	298	2867	9.6	CG16777	0	11	6.3	CG32743	92	455	4.9
CR32862	137	1321	9.5	CG4533	0	11	6.3	CG9885	0	8	4.9
CG12781	0	17	9.5	CG33141	0	10	6.2	CG3830	0	8	4.8
CG8001	216	2048	9.4	CG30001	9	66	6.2	CG6329	0	8	4.8
CG1806	0	17	9.4	CG4633	171	1058	6.1	CG10207	0	8	4.8
CG32577	0	17	9.3	CG14318	2	21	6.1	CG14394	0	8	4.8
CG3552	67	636	9.3	CG9224	2	21	6.0	CG9628	0	8	4.8
CG9240	50	476	9.1	CG17927	25	161	6.0	CG14885	0	8	4.8
CG6178	325	2983	9.1	CG2093	112	671	5.9	CG4738	178	868	4.8
CG3206	4	49	9.1	CG4036	14	95	5.9	CG6908	0	8	4.8
CG14221	2	32	8.9	CG10067	27	169	5.9	CG4696	5	34	4.8
CG3358	50	454	8.7	CG5851	63	374	5.8	CG9416	0	8	4.8
CG33324	0	15	8.6	CG2051	140	804	5.7	CG30409	0	8	4.8
CG12296	0	15	8.6	CG8184	323	1841	5.7	CG10160	0	8	4.8
CG6453	120	1029	8.4	CG32592	0	9	5.7	CG33556	0	8	4.8
CG8884	180	1507	8.3	CG33519	2	20	5.6	CG8317	0	8	4.8
CG6176	22	185	7.9	CG14401	0	9	5.6	CG8646	0	8	4.8
CG10374	4	43	7.9	CG1066	0	9	5.6	CG13521	0	8	4.8
CG14222	201	1609	7.9	CG12324	144	814	5.6	CG5481	0	8	4.8
CR32905	0	14	7.9	CG8937	0	9	5.6	CG1522	0	7	4.7
CG5996	0	14	7.8	CG5927	0	9	5.6	CG31127	111	535	4.7
CG3420	124	952	7.6	CG11173	144	810	5.6	CG9261	0	7	4.7

Table B.1: **List of RNAs associated with Lsm11.** (Table legend on the next page.)

#GID	Lu027_N	Lu028_N	ratio_N
CR32361	4	24	4.7
CG1449	0	7	4.6
CG32352	4	25	4.6
CG10512	5	32	4.6
CG3173	124	576	4.6
CG17764	2	15	4.5
CG6318	4	23	4.5
CG8383	234	1049	4.5
CG3871	11	55	4.5
CG18321	4	23	4.5
CG17531	11	55	4.4
CR31185	135	603	4.4
CG14471	7	38	4.4
CR32866	139	609	4.3
CR31656	139	608	4.3
CR31341	139	608	4.3
CG9533	4	23	4.3
CG18572	250	1072	4.3
CG4250	11	52	4.2
CG42254	9	44	4.2
CG8448	216	912	4.2
CG7860	4	21	4.2
CG33950	30	134	4.2
CG6303	290	1211	4.2
CG3926	2	14	4.1
CG32154	2	14	4.1
CG42458	0	6	4.1
CG12323	505	2072	4.1
CG8777	40	168	4.1
CG10246	2	14	4.1
CG1674	0	6	4.1
CG33473	0	6	4.1
CG31467	0	6	4.1
CG14808	0	6	4.1
CG8012	0	6	4.1
CG40196	0	6	4.1
CG1964	0	6	4.1
CG6696	0	6	4.1
CG5646	0	6	4.1
CR32879	0	6	4.1
CG16857	0	6	4.1
CG7397	0	6	4.1
CG33985	0	6	4.1
CG12590	0	6	4.0
CG14696	0	6	4.0
CG7296	0	6	4.0
CG7702	0	6	4.0
CG13796	0	6	4.0
CG17777	0	6	4.0
CG32773	0	6	4.0
CG8877	624	2527	4.0
CG31721	0	6	4.0
CG18268	0	6	4.0
CG32364	0	6	4.0
CG32179	90	369	4.0
CG32252	0	6	4.0
CG10830	0	6	4.0
CR40621	0	6	4.0
CG7759	0	6	4.0
CG14731	0	6	4.0
CG17876	0	6	4.0
CG33464	0	6	4.0
CG11983	0	6	4.0
CG5126	57	236	4.0
CG5604	1404	5627	4.0
CR33662	128	518	4.0
CG31369	0	6	4.0
CG31317	0	6	4.0
CG33715	36	148	4.0
CG30105	11	49	4.0
CG6907	111	448	4.0
CG31221	0	6	4.0
CG8557	275	1095	4.0
CG6726	29	119	3.9
CG34359	5	27	3.9
CG3891	370	1458	3.9
CG33054	4	20	3.9
CG4405	4	20	3.8
CG8355	2	12	3.8
CG13761	156	601	3.8
CG12116	9	40	3.8
CG2525	5	26	3.8
CG18290	5	26	3.8
CG8021	2	12	3.7
CG18275	2	12	3.7
CG10810	2	12	3.7
CG6980	2	12	3.7
CR31433	2	12	3.7
CG8854	2	12	3.7
CG5023	2	12	3.7
CG10275	7	32	3.7
CG9886	2	12	3.7
CG5715	7	32	3.7
CG6535	54	204	3.7
CG15321	32	124	3.7
CG13852	196	718	3.6
CG14135	4	18	3.6
CG7883	75	279	3.6
CG3585	232	845	3.6
CG10602	243	883	3.6
CG33087	57	212	3.6
CG7487	523	1891	3.6
CG11110	27	102	3.6
CG11734	97	355	3.6
CG42309	4	18	3.6
CG32165	218	785	3.6
CG4931	424	1520	3.6
CG11814	43	160	3.6
CG4715	5	24	3.6
CG8964	5	24	3.6
CG14472	391	1395	3.6
CG15536	18	69	3.5
CG5931	282	1000	3.5
CG17530	108	383	3.5
CG9771	66	238	3.5
CG1915	52	186	3.5
CG8327	158	544	3.4
CG5114	207	706	3.4
CG8318	54	186	3.4
CG6796	9	35	3.4
CG11275	7	29	3.4
CG12499	241	816	3.4
CG1648	2	11	3.4
CG1774	45	156	3.4
CG4178	4	17	3.4
CG3772	4	17	3.3
CG34180	2	11	3.3
CG18410	4	17	3.3
CG18627	84	287	3.3
CG13067	2	11	3.3
CR31540	2	11	3.3
CG2052	4	17	3.3
CG9716	2	11	3.3
CG11678	95	322	3.3
CG12173	108	363	3.3
CG33466	4	17	3.3
CG5029	41	142	3.3
CG9383	201	672	3.3
CG11043	2	11	3.3
CG7144	2	11	3.3
CG9617	131	439	3.3
CG2331	2117	6981	3.3
CG30115	6	23	3.3
CG33103	29	99	3.3
CG13902	56	188	3.3
CG31373	14	52	3.3

Table B.1: **List of RNAs associated with Lsm11.** #GID: gene ID. Lu027_N: normalized raw read numbers for Lu027 (control). Lu028_N: normalized raw read numbers for Lu028 (Lsm11 IP). Ratio_N: ratio between the raw read numbers using (IP+2)/(Ctrl+2).

APPENDIX C: RNA-seq analysis *Drosophila LU* snRNA mutants

C.1 Rationale

In our RIP-seq analysis of Sm proteins in *Drosophila*, we identified a newly evolved snRNA gene, *LU* (described in Chapter 2 in detail). This gene is only present in the melanogaster groups of *Drosophilids*, yet it is highly conserved, suggesting that it is functional. In order to study the function of this gene, I characterized several P element insertion mutants and performed rescue experiments using a genomic fragment containing the gene and the PhiC31 system. Original analysis of the mutants showed that some of these mutants are lethal. I crossed the two transgene lines (LU.tg51C and LU.tg58A, transgenes at the 51C and 58A loci) with the *LU[k05816]* (also known as *LU[10580]*) mutant line. I obtained three recombinants: *LU.10580.tg51C1, LU.10580.tg51C2* and *LU.10580.tg58A*. These recombinants seems to have rescued the pharate lethality of *LU[k05816]*. However, complementation of the mutants with two deficiency lines suggest that the lethality is caused by a second site mutation, which is removed in the rescued lines (See Materials and Methods in Chapter 2).

When I performed RNA-seq analysis on the mutants and the rescued lines, I did not have the results from the complementation experiments. This mistake makes the RNA-seq results uninterpretable. However, it is important to document the experiments performed, so that future experiments can be planned accordingly. Comparison of the rescued lines with mutants that do not have the second site mutations would be helpful in determining the molecular defects in the *LU* gene mutants. In addition, the ribo-minus RNA-seq data from these pharate adults have been very useful for the analysis of the chimeric reads presented in Chapter 3 and circular RNAs.

C.2 Rescue of the *LU* mutation line *k05816(10580)*

The genomic region cloned to rescue the *LU* gene mutations is as follows. The three upper case regions are the PSEA, PSEB and transcribed regions, respectively. Primer binding sites are underlined. The cloned region is 605bp long, including promoter, transcribed region and the 3' end until the intron-exon boundary, and it is expected to be sufficient for its ectopic expression. The whole region is cloned into the pAttB vector, and sequence is verified using conventional Sanger sequencing, and then the construct is injected into flies carrying PhiC31 integration sites at 51C and 58A.

```
cccagatcgcacagctaaatgatggattgtttcgtaaatctaataaacagagctaattatacttaatcccacgacaat
atttgataatagattaatagtgggctgatttgtacagtttttcaatagaactagttatcctaatgcctataagaaatt
gaaagagaacaatatattttcgatcgactgctcaattggccaagcaaggaataacactatcgcttgggtctcccccc
cccctcgtttgaaagcaagccgttcgagcggcgcaagtggctgactcatagcaaacatccacccctcttggttcgctt
tcgcgatcatctttcagagcgccctcgttttccttcaggcgcacaagTAATTCTCAACTGGTTATGGCaagccaacGT
AGAATCcccatcgagtgtcggggatcatcATGTCTCGATCGCCGCTTCAGTTGTGGAGCGAGAGCTTACGCAATGGAG
CGGAGTGATGAGCACATTATCCGAGGCAATTTTTTTAGTGCCTGGCCGCGAAATGCCGCCGGGCCGTtagaaatgaat
atgaaaccatctactttaaatatgattgtaatgtaaaaacttgcatcaacactaaaagg
```

The cloning primers are as follows. The underlined regions are the BamHI and NotI restriction sites. Note that the *LU* gene was initially named *Srv2k*, short for small RNA v(2)k05816.

Srv2kBamH1: ATGGGATCCCCCAGATCGCACAGCTAAATGAT

Srv2kNot1: ATGGCGGCCGCCCTTTTAGTGTTGATGCAAGTTTTTAC

C.3 *LU* mutants and rescue RNA-seq results and discussion

Total RNA was prepared from approximately 15-20 pharates per sample. LU RNA level in these samples has been tested by qRT-PCR (See Chapter 2). Sent 6ug RNA for each sample for sequencing. The first batch were poly(A) selected (Lu031-Lu038), while the second batch were depleted of rRNA (Lu041-Lu044). The samples were sequenced 2X50 paired-ends. A total of 8 samples were multiplexed in one lane. The ribo-minus RNA-seq data were uploaded to GEO, with the accession number: GSE50711. The poly(A)-selected RNA-seq data were not uploaded.

Samples sequenced on 2012-08-25 (poly(A) selected)

Lu031 LU.10580/CyO Actin::GFG

Lu032 LU.10580/CyO Actin::GFG

Lu033 LU.10580/LU.10580

Lu034 LU.10580/LU.10580

Lu035 LU.10580 LU.tg.51C[2] hom (LU.1.3)

Lu036 LU.10580 LU.tg.51C[2] hom (LU.1.3) (sample prepared but degraded, and not sequenced)

Lu037 LU.10580 LU.tg.58A hom (LU.2.10)

Lu038 LU.10580 LU.tg.58A hom (LU.2.10)

Samples sequenced on 2013-01-04 (rRNA-depleted, ribo-minus). Note that Lu045-Lu048 were samples prepared for the RIP-seq analysis of human SmB (see Chapter 2), but sequenced together with the LU mutant and rescue libraries.

Lu041 (Lu031) LU.10580/CyO Actin::GFG

Lu042 (Lu033) LU.10580/LU.10580

Lu043 (-----) LU.10580 LU.tg.51C[1] hom (LU.1.2)

Lu044 (Lu037) LU.10580 LU.tg.58A hom (LU.2.10)

Lu045 HeLa Ctrl

Lu046 HeLa Ctrl

Lu047 HeLa Y12-IP

Lu048 HeLa Y12-IP

Comparison among the *LU* snRNA heterozygotes, homozygotes and rescued pharate adults (Lu031-Lu038) showed that expression of some RNAs are significantly down- or up-regulated in homozygotes compared to heterozygotes and rescued pharate adults. However, it is difficult to determine whether it is because of loss of the LU snRNA. A few examples are shown in the following figure (Figure C.1).

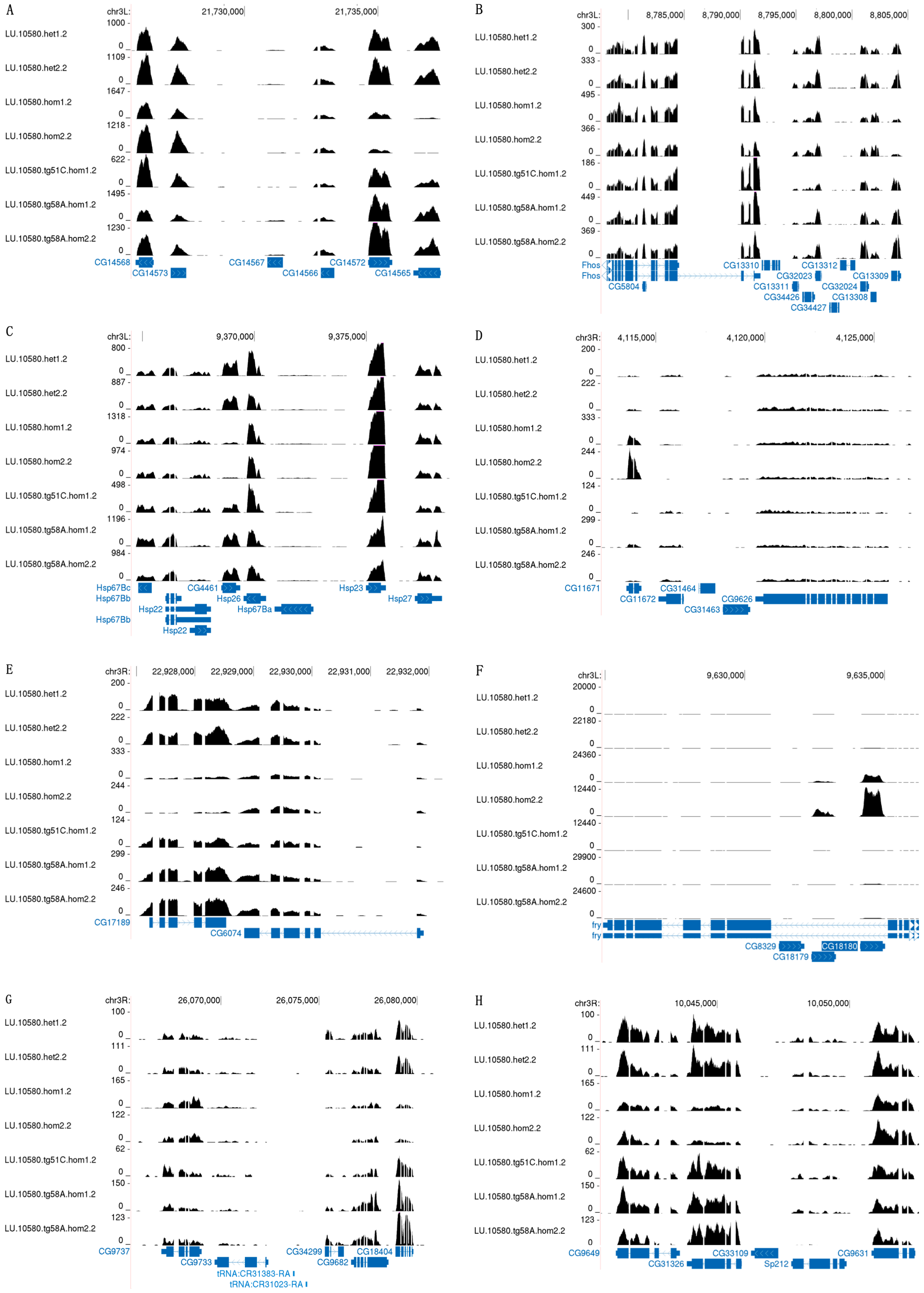


Figure C.1: **Example mRNAs with altered expression in the *LU* mutant.** Continued on the next page.

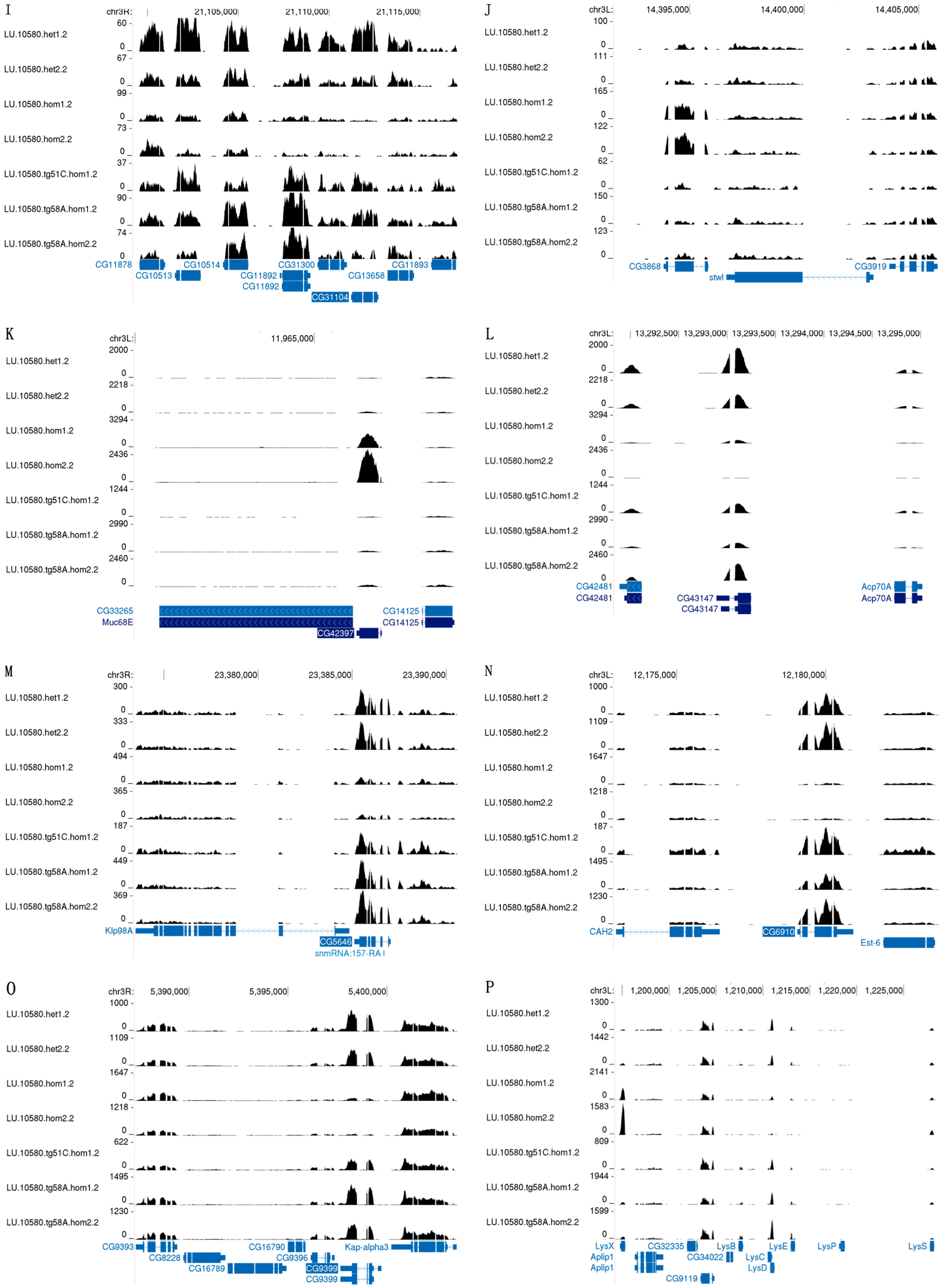


Figure C.1: **Example mRNAs with altered expression in the *LU* mutant.** Continued on the next page.

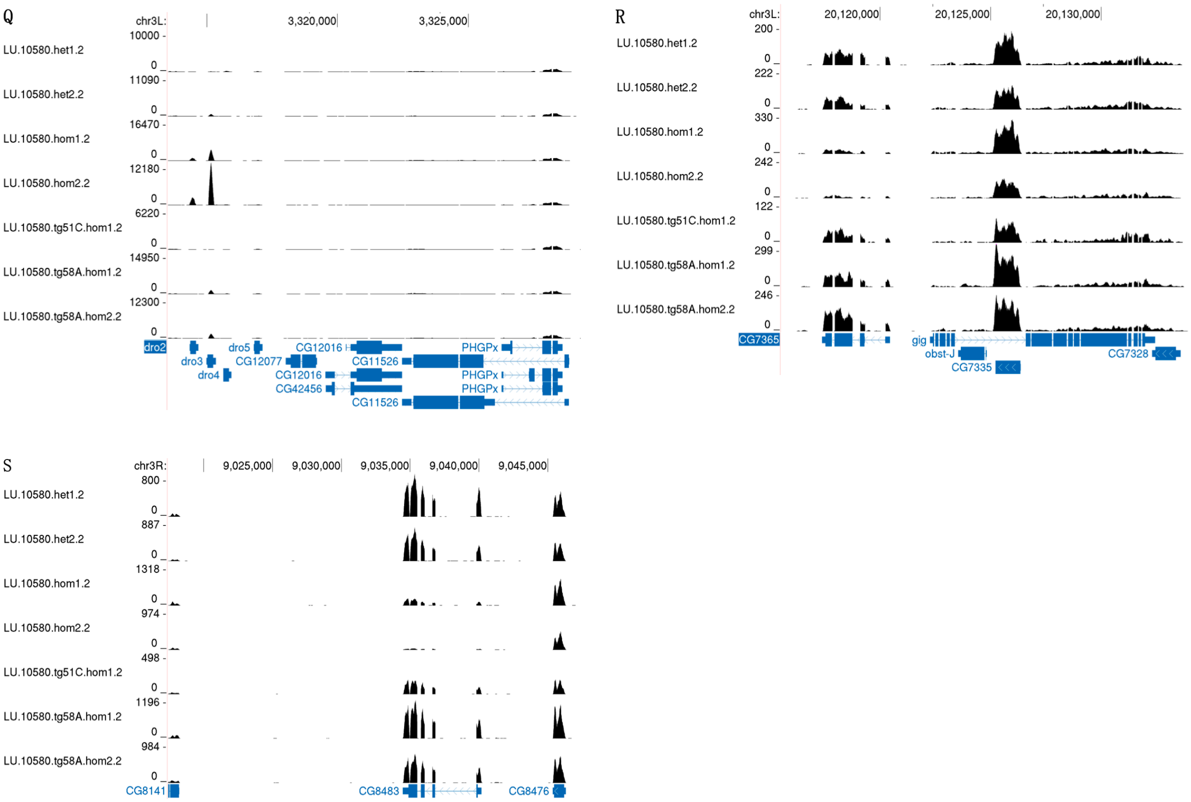


Figure C.1: **Example mRNAs with altered expression in the *LU* mutant.** Note that the scales are read numbers after normalization of the tracks.

APPENDIX D: List of publications and manuscripts

Lu, Z. and Matera, A. G. (2014). Vicinal: a method for the determination of ncRNA ends using chimeric reads from RNA-seq experiments. ***Nucleic Acids Research***, March 12, 2014

Lu, Z., Guan, X., Schmidt, C. A., and Matera, A. G. (2014). RIP-seq analysis of eukaryotic Sm proteins identifies three major categories of Sm-containing ribonucleoproteins. ***Genome Biology***, 15(1), R7.

Lu, Z. and Matera, A. G. (2014). Cytoplasmic SMN body formation caused by a block in snRNP assembly. *Manuscript in preparation.*

Lu, Z. and Matera, A. G. (2014). Developmental switching of major snRNA isoforms is conserved in evolution. *Manuscript in preparation.*

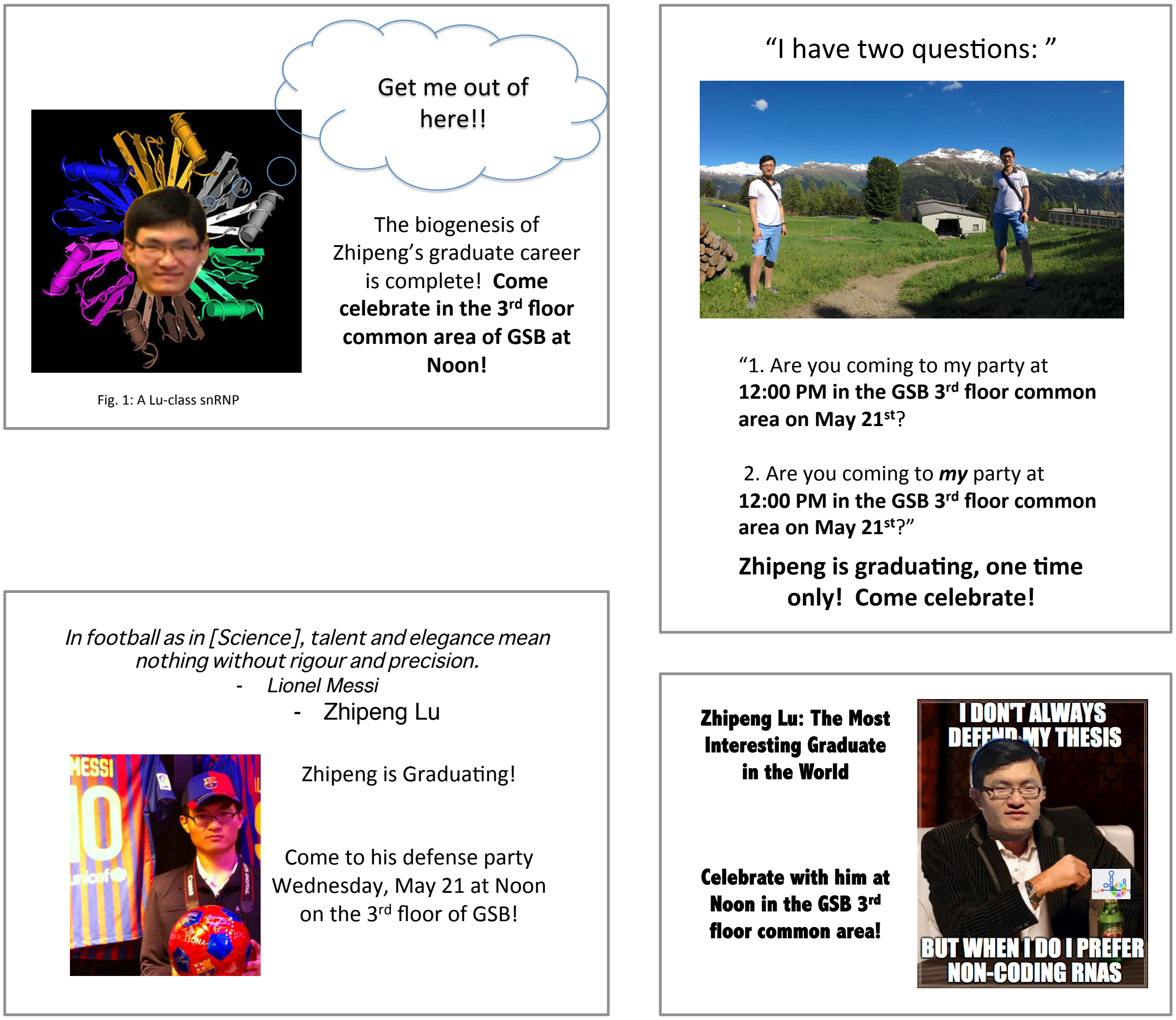
Garcia, E. L., **Lu, Z.**, Meers, M. P., Praveen, K., and Matera, A. G. (2013). Developmental arrest of *Drosophila Survival motor neuron* (*Smn*) mutants accounts for differences in expression of minor intron-containing genes. ***RNA***, 19(11), 1510–6.

Wang H., Zhang S., Wang S., Lu J., Wu W., Weng L., Chen D., Zhang Y., **Lu Z.**, Yang J., Chen Y., Zhang X., Chen X., Xi C., Lu D. and Zhao S. (2009) *REV3L* confers chemoresistance to cisplatin in human gliomas: The potential of its RNAi for synergistic therapy. ***Neuro-Oncology*** 11: 790–802.

Jiang M., Fei J., Lan M.S., **Lu Z.**, Liu M., Fan W., Gao X. and Lu D., (2008). Hypermethylation of hepatic *Gck* promoter in ageing rats contributes to diabetogenic potential. ***Diabetologia*** 51: 1525–1533

APPENDIX E: Defense flyers

At the end of my dissertation, I would like to include as an appendix these four lovely flyers created by my labmates Kelsey Gray and Mike Meers for my defense (Figure E.1). I don't always write dissertations, but when I do, I prefer to make them fun.



Get me out of here!!

The biogenesis of Zhipeng's graduate career is complete! **Come celebrate in the 3rd floor common area of GSB at Noon!**

Fig. 1: A Lu-class snRNP

In football as in [Science], talent and elegance mean nothing without rigour and precision.
- *Lionel Messi*
- Zhipeng Lu

Zhipeng is Graduating!

Come to his defense party Wednesday, May 21 at Noon on the 3rd floor of GSB!

"I have two questions: "

"1. Are you coming to my party at **12:00 PM in the GSB 3rd floor common area on May 21st**?

2. Are you coming to ***my*** party at **12:00 PM in the GSB 3rd floor common area on May 21st**?"

Zhipeng is graduating, one time only! Come celebrate!

Zhipeng Lu: The Most Interesting Graduate in the World

Celebrate with him at Noon in the GSB 3rd floor common area!

MESSI
I DON'T ALWAYS DEFEND MY THESIS
BUT WHEN I DO I PREFER NON-CODING RNAS

Figure E.1: **Defense party flyers made by Kelsey and Mike**

BIBLOGRAPHY

Achsel, T., Brahms, H., Kastner, B., Bachi, A., Wilm, M., and Luhrmann, R. (1999). A doughnut-shaped heteromer of human sm-like proteins binds to the 3'-end of u6 snrna, thereby facilitating u4/u6 duplex formation in vitro. *EMBO J*, **18**(20), 5789–802.

Achsel, T., Stark, H., and Luhrmann, R. (2001). The sm domain is an ancient rna-binding motif with oligo(u) specificity. *Proc Natl Acad Sci U S A*, **98**(7), 3685–9.

Afonyushkin, T., Vecerek, B., Moll, I., Blasi, U., and Kaberdin, V. R. (2005). Both rnase e and rnase iii control the stability of sodb mrna upon translational inhibition by the small regulatory rna ryhb. *Nucleic Acids Res*, **33**(5), 1678–89.

Albrecht, J. C. and Fleckenstein, B. (1992). Nucleotide sequence of hsur 6 and hsur 7, two small rnas of herpesvirus saimiri. *Nucleic Acids Res*, **20**(7), 1810.

Albrecht, M. and Lengauer, T. (2004). Novel sm-like proteins with long c-terminal tails and associated methyltransferases. *FEBS Lett*, **569**(1-3), 18–26.

Alioto, T. S. (2007). U12db: a database of orthologous u12-type spliceosomal introns. *Nucleic Acids Res*, **35**(Database issue), D110–5.

Almada, A. E., Wu, X., Kriz, A. J., Burge, C. B., and Sharp, P. A. (2013). Promoter directionality is controlled by u1 snrnp and polyadenylation signals. *Nature*, **499**(7458), 360–3.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**(3), 403–10.

Anantharaman, V., Koonin, E. V., and Aravind, L. (2002). Comparative genomics and evolution of proteins involved in rna metabolism. *Nucleic Acids Res*, **30**(7), 1427–64.

Andrade, L. E., Tan, E. M., and Chan, E. K. (1993). Immunocytochemical analysis of the coiled body in the cell cycle and during cell proliferation. *Proc Natl Acad Sci U S A*, **90**(5), 1947–51.

Anko, M. L. and Neugebauer, K. M. (2012). Rna-protein interactions in vivo: global gets specific. *Trends Biochem Sci*, **37**(7), 255–62.

Anko, M. L., Muller-McNicoll, M., Brandl, H., Curk, T., Gorup, C., Henry, I., Ule, J., and Neugebauer, K. M. (2012). The rna-binding landscapes of two sr proteins reveal unique functions and binding to diverse rna classes. *Genome Biol*, **13**(3), R17.

Anne, J., Ollo, R., Ephrussi, A., and Mechler, B. M. (2007). Arginine methyltransferase capsuleen is essential for methylation of spliceosomal sm proteins and germ cell formation in drosophila. *Development*, **134**(1), 137–46.

Argaman, L. and Altuvia, S. (2000). fhla repression by oxys rna: kissing complex formation at two sites results in a stable antisense-target rna complex. *J Mol Biol*, **300**(5), 1101–12.

Ashe, M. P., Pearson, L. H., and Proudfoot, N. J. (1997). The hiv-1 5' ltr poly(a) site is inactivated by u1 snrnp interaction with the downstream major splice donor site. *EMBO J*, **16**(18), 5752–63.

Ashe, M. P., Furger, A., and Proudfoot, N. J. (2000). Stem-loop 1 of the u1 snrnp plays a critical role in the suppression of hiv-1 polyadenylation. *RNA*, **6**(2), 170–7.

Bach, M., Krol, A., and Luhrmann, R. (1990). Structure-probing of u1 snrnps gradually depleted of the u1-specific proteins a, c and 70k. evidence that a interacts differentially with developmentally regulated mouse u1 snrna variants. *Nucleic Acids Res*, **18**(3), 449–57.

Bahn, J. H., Lee, J. H., Li, G., Greer, C., Peng, G., and Xiao, X. (2012). Accurate identification of a-to-i rna editing in human by transcriptome sequencing. *Genome Res*, **22**(1), 142–50.

Baillat, D., Hakimi, M. A., Naar, A. M., Shilatifard, A., Cooch, N., and Shiekhattar, R. (2005). Integrator, a multiprotein mediator of small nuclear rna processing, associates with the c-terminal repeat of rna polymerase ii. *Cell*, **123**(2), 265–76.

Baltz, A. G., Munschauer, M., Schwanhausser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., and Landthaler, M. (2012). The mrna-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol Cell*, **46**(5), 674–90.

Barbee, S. A. and Evans, T. C. (2006). The sm proteins regulate germ cell specification during early c. elegans embryogenesis. *Dev Biol*, **291**(1), 132–43.

Barbee, S. A., Lublin, A. L., and Evans, T. C. (2002). A novel function for the sm proteins in germ granule localization during c. elegans embryogenesis. *Curr Biol*, **12**(17), 1502–6.

Bardill, J. P., Zhao, X., and Hammer, B. K. (2011). The vibrio cholerae quorum sensing response is mediated by hfq-dependent srna/mrna base pairing interactions. *Mol Microbiol*, **80**(5), 1381–94.

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-c: a comprehensive technique to capture the conformation of genomes. *Methods*, **58**(3), 268–76.

Benaglia, T., Chauveau, D., Hunter, D., and Young, D. (2009). mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software*, **32**(6), 29.

Benjamin, J. A., Desnoyers, G., Morissette, A., Salvail, H., and Masse, E. (2010). Dealing with oxidative stress and iron starvation in microorganisms: an overview. *Can J Physiol Pharmacol*, **88**(3), 264–72.

Benoit, B., Nemeth, A., Aulner, N., Kuhn, U., Simonelig, M., Wahle, E., and Bourbon, H. M. (1999). The drosophila poly(a)-binding protein ii is ubiquitous throughout drosophila development and has the same function in mrna polyadenylation as its bovine homolog in vitro. *Nucleic Acids Res*, **27**(19), 3771–8.

Bentley, D. L. (2014). Coupling mrna processing with transcription in time and space. *Nat Rev Genet*, **15**(3), 163–75.

Berg, M. G., Singh, L. N., Younis, I., Liu, Q., Pinto, A. M., Kaida, D., Zhang, Z., Cho, S., Sherrill-Mix, S., Wan, L., and Dreyfuss, G. (2012). U1 snrnp determines mrna length and regulates isoform expression. *Cell*, **150**(1), 53–64.

Bernstein, L. B., Mount, S. M., and Weiner, A. M. (1983). Pseudogenes for human small nuclear rna u3 appear to arise by integration of self-primed reverse transcripts of the rna into new chromosomal sites. *Cell*, **32**(2), 461–72.

Bilinski, S. M., Jaglarz, M. K., Szymanska, B., Etkin, L. D., and Kloc, M. (2004). Sm proteins, the constituents of the spliceosome, are components of nuage and mitochondrial cement in xenopus oocytes. *Exp Cell Res*, **299**(1), 171–8.

Blackwell, E., Zhang, X., and Ceman, S. (2010). Arginines of the rgg box regulate fmrp association with polyribosomes and mrna. *Hum Mol Genet*, **19**(7), 1314–23.

Boelens, W. C., Palacios, I., and Mattaj, I. W. (1995). Nuclear retention of rna as a mechanism for localization. *RNA*, **1**(3), 273–83.

Bordonne, R. and Tarassov, I. (1996). The yeast sme1 gene encodes the homologue of the human e core protein. *Gene*, **176**(1-2), 111–7.

Bourguignon, G. J., Tattersall, P. J., and Ward, D. C. (1976). Dna of minute virus of mice: self-priming, nonpermuted, single-stranded genome with a 5'-terminal hairpin duplex. *J Virol*, **20**(1), 290–306.

Bouvier, M., Sharma, C. M., Mika, F., Nierhaus, K. H., and Vogel, J. (2008). Small rna binding to 5' mrna coding region inhibits translational initiation. *Mol Cell*, **32**(6), 827–37.

Box, J. A., Bunch, J. T., Tang, W., and Baumann, P. (2008). Spliceosomal cleavage generates the 3' end of telomerase rna. *Nature*, **456**(7224), 910–4.

Brahms, H., Meheus, L., de Brabandere, V., Fischer, U., and Luhrmann, R. (2001). Symmetrical dimethylation of arginine residues in spliceosomal sm protein b/b' and the sm-like protein lsm4, and their interaction with the smn protein. *RNA*, **7**(11), 1531–42.

Brand, A. H. and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, **118**(2), 401–15.

Brescia, C. C., Mikulecky, P. J., Feig, A. L., and Sledjeski, D. D. (2003). Identification of the hfq-binding site on dsra rna: Hfq binds without altering dsra secondary structure. *RNA*, **9**(1), 33–43.

Brooks, A. N., Yang, L., Duff, M. O., Hansen, K. D., Park, J. W., Dudoit, S., Brenner, S. E., and Graveley, B. R. (2011). Conservation of an rna regulatory map between drosophila and mammals. *Genome Res*, **21**(2), 193–202.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., and Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat Methods*, **10**(12), 1213–8.

Bullard, J. H., Purdom, E., Hansen, K. D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**, 94.

Bullock, S. L. and Ish-Horowicz, D. (2001). Conserved signals and machinery for rna transport in drosophila oogenesis and embryogenesis. *Nature*, **414**(6864), 611–6.

Burch, B. D., Godfrey, A. C., Gasdaska, P. Y., Salzler, H. R., Duronio, R. J., Marzluff, W. F., and Dominski, Z. (2011). Interaction between flash and lsm11 is essential for histone pre-mrna processing in vivo in drosophila. *RNA*, **17**(6), 1132–47.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P., and Bateman, A. (2013). Rfam 11.0: 10 years of rna families. *Nucleic Acids Res*, **41**(Database issue), D226–32.

Burlet, P., Huber, C., Bertrandy, S., Ludosky, M. A., Zwaenepoel, I., Clermont, O., Roume, J., Delezoide, A. L., Cartaud, J., Munnich, A., and Lefebvre, S. (1998). The distribution of smn protein complex in human fetal tissues and its alteration in spinal muscular atrophy. *Hum Mol Genet*, **7**(12), 1927–33.

Callis, J., Fromm, M., and Walbot, V. (1987). Introns increase gene expression in cultured maize cells. *Genes Dev*, **1**(10), 1183–200.

Calvet, J. P. and Pederson, T. (1979). Photochemical cross-linking of secondary structure in hela cell heterogeneous nuclear rna in situ 1. *Nucleic Acids Res*, **6**(5), 1993–2001.

Calvet, J. P. and Pederson, T. (1981). Base-pairing interactions between small nuclear rnas and nuclear rna precursors as revealed by psoralen cross-linking in vivo. *Cell*, **26**(3 Pt 1), 363–70.

Carmo-Fonseca, M. and Hurt, E. C. (1991). Across the nuclear pores with the help of nucleoporins. *Chromosoma*, **101**(4), 199–205.

Carvalho, T., Almeida, F., Calapez, A., Lafarga, M., Berciano, M. T., and Carmo-Fonseca, M. (1999). The spinal muscular atrophy disease gene product, smn: A link between snrnp biogenesis and the cajal (coiled) body. *J Cell Biol*, **147**(4), 715–28.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., and Hentze, M. W. (2012). Insights into rna biology from an atlas of mammalian mrna-binding proteins. *Cell*, **149**(6), 1393–406.

Castello, A., Fischer, B., Hentze, M. W., and Preiss, T. (2013). Rna-binding proteins in mendelian disease. *Trends Genet*, **29**(5), 318–27.

Cauchi, R. J. (2011). Gem formation upon constitutive gemin3 overexpression in drosophila. *Cell Biol Int*, **35**(12), 1233–8.

Cauchi, R. J., Sanchez-Pulido, L., and Liu, J. L. (2010). Drosophila smn complex proteins gemin2, gemin3, and gemin5 are components of u bodies. *Exp Cell Res*, **316**(14), 2354–64.

Cazalla, D., Yario, T., and Steitz, J. A. (2010). Down-regulation of a host microrna by a herpesvirus saimiri noncoding rna. *Science*, **328**(5985), 1563–6.

Chari, A., Golas, M. M., Klingenhager, M., Neuenkirchen, N., Sander, B., Englbrecht, C., Sickmann, A., Stark, H., and Fischer, U. (2008). An assembly chaperone collaborates with the smn complex to generate spliceosomal snrnps. *Cell*, **135**(3), 497–509.

Chen, C., Ridzon, D. A., Broomer, A. J., Zhou, Z., Lee, D. H., Nguyen, J. T., Barbisin, M., Xu, N. L., Mahuvakar, V. R., Andersen, M. R., Lao, K. Q., Livak, K. J., and Guegler, K. J. (2005a). Real-time quantification of micrornas by stem-loop rt-pcr. *Nucleic Acids Res*, **33**(20), e179.

Chen, L., Lullo, D. J., Ma, E., Celniker, S. E., Rio, D. C., and Doudna, J. A. (2005b). Identification and analysis of u5 snrna variants in drosophila. *RNA*, **11**(10), 1473–7.

Chen, M. and Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol*, **10**(11), 741–54.

Chen, S., Zhang, A., Blyn, L. B., and Storz, G. (2004). Micc, a second small-rna regulator of omp protein expression in escherichia coli. *J Bacteriol*, **186**(20), 6689–97.

Cheng, Y., Lund, E., Kahan, B. W., and Dahlberg, J. E. (1997). Control of mouse u1 snrna gene expression during in vitro differentiation of mouse embryonic stem cells. *Nucleic Acids Res*, **25**(11), 2197–204.

Chowdhury, A., Mukhopadhyay, J., and Tharun, S. (2007). The decapping activator lsm1p-7p-pat1p complex has the intrinsic ability to distinguish between oligoadenylated and polyadenylated rnas. *RNA*, **13**(7), 998–1016.

Chu, C., Quinn, J., and Chang, H. Y. (2012). Chromatin isolation by rna purification (chirp). *J Vis Exp*, (61).

Clark, A., Meignin, C., and Davis, I. (2007). A dynein-dependent shortcut rapidly delivers axis determination transcripts into the drosophila oocyte. *Development*, **134**(10), 1955–65.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**(5258), 561–3.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). Weblogo: a sequence logo generator. *Genome Res*, **14**(6), 1188–90.

Croucher, N. J. and Thomson, N. R. (2010). Studying bacterial transcriptomes using rna-seq. *Curr Opin Microbiol*, **13**(5), 619–24.

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., Hu, S., and Yu, J. (2010). A comparison between ribo-minus rna-sequencing and polya-selected rna-sequencing. *Genomics*, **96**(5), 259–65.

Cziko, A. M., McCann, C. T., Howlett, I. C., Barbee, S. A., Duncan, R. P., Luedemann, R., Zarnescu, D., Zinsmaier, K. E., Parker, R. R., and Ramaswami, M. (2009). Genetic modifiers of dfmr1 encode rna granule components in drosophila. *Genetics*, **182**(4), 1051–60.

Danan, M., Schwartz, S., Edelheit, S., and Sorek, R. (2012). Transcriptome-wide discovery of circular rnas in archaea. *Nucleic Acids Res*, **40**(7), 3131–42.

Darty, K., Denise, A., and Ponty, Y. (2009). Varna: Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, **25**(15), 1974–5.

Darzacq, X., Jady, B. E., Verheggen, C., Kiss, A. M., Bertrand, E., and Kiss, T. (2002). Cajal body-specific small nuclear rnas: a novel class of 2'-o-methylation and pseudouridylation guide rnas. *EMBO J*, **21**(11), 2746–56.

de Hoog, C. L., Foster, L. J., and Mann, M. (2004). Rna and rna binding proteins participate in early stages of cell spreading through spreading initiation centers. *Cell*, **117**(5), 649–62.

Dej, K. J. and Spradling, A. C. (1999). The endocycle controls nurse cell polytene chromosome structure during drosophila oogenesis. *Development*, **126**(2), 293–303.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. *Science*, **295**(5558), 1306–11.

Delihas, N. and Forst, S. (2001). Micf: an antisense rna gene involved in response of escherichia coli to global stress factors. *J Mol Biol*, **313**(1), 1–12.

Denison, R. A., Van Arsdell, S. W., Bernstein, L. B., and Weiner, A. M. (1981). Abundant pseudogenes for small nuclear rnas are dispersed in the human genome. *Proc Natl Acad Sci U S A*, **78**(2), 810–4.

Derelle, R., Momose, T., Manuel, M., Da Silva, C., Wincker, P., and Houliston, E. (2010). Convergent origins and rapid evolution of spliced leader trans-splicing in metazoa: insights from the ctenophora and hydrozoa. *RNA*, **16**(4), 696–707.

Deryusheva, S. and Gall, J. G. (2013). Novel small cajal-body-specific rnas identified in drosophila: probing guide rna function. *RNA*, **19**(12), 1802–14.

Dominski, Z., Yang, X. C., Purdy, M., and Marzluff, W. F. (2003). Cloning and characterization of the drosophila u7 small nuclear rna. *Proc Natl Acad Sci U S A*, **100**(16), 9422–7.

Domitrovich, A. M. and Kunkel, G. R. (2003). Multiple, dispersed human u6 small nuclear rna genes with varied transcriptional efficiencies. *Nucleic Acids Res*, **31**(9), 2344–52.

Dostie, J. and Dekker, J. (2007). Mapping networks of physical interactions between genomic elements using 5c technology. *Nat Protoc*, **2**(4), 988–1002.

Doudna, J. A. and Cech, T. R. (2002). The chemical repertoire of natural ribozymes. *Nature*, **418**(6894), 222–8.

Edgren, H., Murumagi, A., Kangaspeska, S., Nicorici, D., Hongisto, V., Kleivi, K., Rye, I. H., Nyberg, S., Wolf, M., Borresen-Dale, A. L., and Kallioniemi, O. (2011). Identification of fusion genes in breast cancer by paired-end rna-sequencing. *Genome Biol*, **12**(1), R6.

Eliceiri, G. L. and Sayavedra, M. S. (1976). Small rnas in the nucleus and cytoplasm of hela cells. *Biochem Biophys Res Commun*, **72**(2), 507–12.

Eulalio, A., Behm-Ansmant, I., Schweizer, D., and Izaurralde, E. (2007). P-body formation is a consequence, not the cause, of rna-mediated gene silencing. *Mol Cell Biol*, **27**(11), 3970–81.

Fenger-Gron, M., Fillman, C., Norrild, B., and Lykke-Andersen, J. (2005). Multiple processing body factors and the are binding protein ttp activate mrna decapping. *Mol Cell*, **20**(6), 905–15.

Fernandez, C. F., Pannone, B. K., Chen, X., Fuchs, G., and Wolin, S. L. (2004). An lsm2-lsm7 complex in saccharomyces cerevisiae associates with the small nucleolar rna snr5. *Mol Biol Cell*, **15**(6), 2842–52.

Fica, S. M., Tuttle, N., Novak, T., Li, N. S., Lu, J., Koodathingal, P., Dai, Q., Staley, J. P., and Piccirilli, J. A. (2013). Rna catalyses nuclear pre-mrna splicing. *Nature*, **503**(7475), 229–34.

Fischer, S., Benz, J., Spath, B., Maier, L. K., Straub, J., Granzow, M., Raabe, M., Urlaub, H., Hoffmann, J., Brutschy, B., Allers, T., Soppa, J., and Marchfelder, A. (2010). The archaeal lsm protein binds to small rnas. *J Biol Chem*, **285**(45), 34429–38.

Fischer, U., Liu, Q., and Dreyfuss, G. (1997). The smn-sip1 complex has an essential role in spliceosomal snrnp biogenesis. *Cell*, **90**(6), 1023–9.

Flickinger, T. W. and Salz, H. K. (1994). The drosophila sex determination gene snf encodes a nuclear protein with sequence and functional similarity to the mammalian u1a snrnp protein. *Genes Dev*, **8**(8), 914–25.

Foley, K. and Cooley, L. (1998). Apoptosis in late stage drosophila nurse cells does not require genes within the h99 deficiency. *Development*, **125**(6), 1075–82.

Folichon, M., Arluison, V., Pellegrini, O., Huntzinger, E., Regnier, P., and Hajnsdorf, E. (2003). The poly(a) binding protein hfq protects rna from rnase e and exoribonucleolytic degradation. *Nucleic Acids Res*, **31**(24), 7302–10.

Forbes, D. J., Kirschner, M. W., Caput, D., Dahlberg, J. E., and Lund, E. (1984). Differential expression of multiple u1 small nuclear rnas in oocytes and embryos of xenopus laevis. *Cell*, **38**(3), 681–9.

Franze de Fernandez, M. T., Eoyang, L., and August, J. T. (1968). Factor fraction required for the synthesis of bacteriophage qbeta-rna. *Nature*, **219**(5154), 588–90.

Freund, M., Hicks, M. J., Konermann, C., Otte, M., Hertel, K. J., and Schaal, H. (2005). Extended base pair complementarity between u1 snrna and the 5' splice site does not inhibit splicing in higher eukaryotes, but rather increases 5' splice site recognition. *Nucleic Acids Res*, **33**(16), 5112–9.

Frey, M. R. and Matera, A. G. (1995). Coiled bodies contain u7 small nuclear rna and associate with specific dna sequences in interphase human cells. *Proc Natl Acad Sci U S A*, **92**(13), 5915–9.

Friedersdorf, M. B. and Keene, J. D. (2014). Advancing the functional utility of par-clip by quantifying background binding to mrnas and lncrnas. *Genome Biol*, **15**(1), R2.

Friend, K., Lovejoy, A. F., and Steitz, J. A. (2007). U2 snrnp binds intronless histone pre-mrnas to facilitate u7-snrnp-dependent 3' end formation. *Mol Cell*, **28**(2), 240–52.

Fu, D. and Collins, K. (2006). Human telomerase and cajal body ribonucleoproteins share a unique specificity of sm protein association. *Genes Dev*, **20**(5), 531–6.

Fujita, T., Asano, Y., Ohtsuka, J., Takada, Y., Saito, K., Ohki, R., and Fujii, H. (2013). Identification of telomere-associated molecules by engineered dna-binding molecule-mediated chromatin immunoprecipitation (enchip). *Sci Rep*, **3**, 3171.

Fury, M., Andersen, J., Ponda, P., Aimes, R., and Zieve, G. W. (1999). Thirteen anti-sm monoclonal antibodies immunoprecipitate the three cytoplasmic snrnp core protein precursors in six distinct subsets. *J Autoimmun*, **12**(2), 91–100.

Gall, J. G. (2000). Cajal bodies: the first 100 years. *Annu Rev Cell Dev Biol*, **16**, 273–300.

Garcia, E. L., Lu, Z., Meers, M. P., Praveen, K., and Matera, A. G. (2013). Developmental arrest of drosophila survival motor neuron (smn) mutants accounts for differences in expression of minor intron-containing genes. *RNA*, **19**(11), 1510–6.

Gerber, A. P., Herschlag, D., and Brown, P. O. (2004). Extensive association of functionally and cytotopically related mrnas with puf family rna-binding proteins in yeast. *PLoS Biol*, **2**(3), E79.

Gerber, A. P., Luschnig, S., Krasnow, M. A., Brown, P. O., and Herschlag, D. (2006). 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.

Gilmour, D. S. and Lis, J. T. (1984). Detecting protein-dna interactions in vivo: distribution of rna polymerase on specific bacterial genes. *Proc Natl Acad Sci U S A*, **81**(14), 4275–9.

Godfrey, A. C., White, A. E., Tatomer, D. C., Marzluff, W. F., and Duronio, R. J. (2009). The drosophila u7 snrnp proteins lsm10 and lsm11 are required for histone pre-mrna processing and play an essential role in development. *RNA*, **15**(9), 1661–72.

Gonsalvez, G. B., Rajendra, T. K., Tian, L., and Matera, A. G. (2006). The sm-protein methyltransferase, dart5, is essential for germ-cell specification and maintenance. *Curr Biol*, **16**(11), 1077–89.

Gonsalvez, G. B., Tian, L., Ospina, J. K., Boisvert, F. M., Lamond, A. I., and Matera, A. G. (2007). Two distinct arginine methyltransferases are required for biogenesis of sm-class ribonucleoproteins. *J Cell Biol*, **178**(5), 733–40.

Gonsalvez, G. B., Rajendra, T. K., Wen, Y., Praveen, K., and Matera, A. G. (2010). Sm proteins specify germ cell fate by facilitating oskar mrna localization. *Development*, **137**(14), 2341–51.

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on u3 snorna and pre-rrna by uv cross-linking and high-throughput analysis of cdnas. *Proc Natl Acad Sci U S A*, **106**(24), 9613–8.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., Artieri, C. G., van Baren, M. J., Boley, N., Booth, B. W., Brown, J. B., Cherbas, L., Davis, C. A., Dobin, A., Li, R., Lin, W., Malone, J. H., Mattiuzzo, N. R., Miller, D., Sturgill, D., Tuch, B. B., Zaleski, C., Zhang, D., Blanchette, M., Dudoit, S., Eads, B., Green, R. E., Hammonds, A., Jiang, L., Kapranov, P., Langton, L., Perrimon, N., Sandler, J. E., Wan, K. H., Willingham, A., Zhang, Y., Zou, Y., Andrews, J., Bickel, P. J., Brenner, S. E., Brent, M. R., Cherbas, P., Gingeras, T. R., Hoskins, R. A., Kaufman, T. C., Oliver, B., and Celniker, S. E. (2011). The developmental transcriptome of drosophila melanogaster. *Nature*, **471**(7339), 473–9.

Grima, D. P., Sullivan, M., Zabolotskaya, M. V., Browne, C., Seago, J., Wan, K. C., Okada, Y., and Newbury, S. F. (2008). The 5'-3' exoribonuclease pacman is required for epithelial sheet sealing in drosophila and genetically interacts with the phosphatase puckered. *Biol Cell*, **100**(12), 687–701.

Grimm, C., Chari, A., Pelz, J. P., Kuper, J., Kisker, C., Diederichs, K., Stark, H., Schindelin, H., and Fischer, U. (2013). Structural basis of assembly chaperone- mediated snrnp formation. *Mol Cell*, **49**(4), 692–703.

Gruss, P., Lai, C. J., Dhar, R., and Khoury, G. (1979). Splicing as a requirement for biogenesis of functional 16s mrna of simian virus 40. *Proc Natl Acad Sci U S A*, **76**(9), 4317–21.

Gubler, U. and Hoffman, B. J. (1983). A simple and very efficient method for generating cdna libraries. *Gene*, **25**(2-3), 263–9.

Guild, G. M., Connelly, P. S., Shaw, M. K., and Tilney, L. G. (1997). Actin filament cables in drosophila nurse cells are composed of modules that slide passively past one another during dumping. *J Cell Biol*, **138**(4), 783–97.

Gunderson, S. I., Beyer, K., Martin, G., Keller, W., Boelens, W. C., and Mattaj, L. W. (1994). The human u1a snrnp protein regulates polyadenylation via a direct interaction with poly(a) polymerase. *Cell*, **76**(3), 531–41.

Gunderson, S. I., Polycarpou-Schwarz, M., and Mattaj, I. W. (1998). U1 snrnp inhibits pre-mrna polyadenylation through a direct interaction between u1 70k and poly(a) polymerase. *Mol Cell*, **1**(2), 255–64.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., J., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010). Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, **141**(1), 129–41.

Hajnsdorf, E. and Regnier, P. (2000). Host factor hfq of escherichia coli stimulates elongation of poly(a) tails by poly(a) polymerase i. *Proc Natl Acad Sci U S A*, **97**(4), 1501–5.

Hallais, M., Pontvianne, F., Andersen, P. R., Clerici, M., Lener, D., Benbahouche Nel, H., Gostan, T., Vandermoere, F., Robert, M. C., Cusack, S., Verheggen, C., Jensen, T. H., and Bertrand, E. (2013). Cbc-ars2 stimulates 3'-end maturation of multiple rna families and favors cap-proximal processing. *Nat Struct Mol Biol*, **20**(12), 1358–66.

Hamm, J. and Mattaj, I. W. (1989). An abundant u6 snrnp found in germ cells and embryos of xenopus laevis. *EMBO J*, **8**(13), 4179–87.

Hanley, B. A. and Schuler, M. A. (1991). Developmental expression of plant snrnas. *Nucleic Acids Res*, **19**(22), 6319–25.

Hansen, T. B., Kjems, J., and Damgaard, C. K. (2013). Circular rna and mir-7 in cancer. *Cancer Res*, **73**(18), 5609–12.

Hebert, M. D., Shpargel, K. B., Ospina, J. K., Tucker, K. E., and Matera, A. G. (2002). Coilin methylation regulates nuclear body formation. *Dev Cell*, **3**(3), 329–37.

Helwak, A. and Tollervey, D. (2014). Mapping the mirna interactome by cross-linking ligation and sequencing of hybrids (clash). *Nat Protoc*, **9**(3), 711–28.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, **153**(3), 654–65.

Hentze, M. W. and Preiss, T. (2010). The rem phase of gene regulation. *Trends Biochem Sci*, **35**(8), 423–6.

Hentze, M. W. and Preiss, T. (2013). Circular rnas: splicing's enigma variations. *EMBO J*, **32**(7), 923–5.

Hernandez, N. (2001). Small nuclear rna genes: a model system to study fundamental mechanisms of transcription. *J Biol Chem*, **276**(29), 26733–6.

Hinas, A., Larsson, P., Avesson, L., Kirsebom, L. A., Virtanen, A., and Soderbom, F. (2006). Identification of the major spliceosomal rnas in dictyostelium discoideum reveals developmentally regulated u2 variants and polyadenylated snrnas. *Eukaryot Cell*, **5**(6), 924–34.

Hofacker, I. L. (2003). Vienna rna secondary structure server. *Nucleic Acids Res*, **31**(13), 3429–31.

Houlden, H. and Singleton, A. B. (2012). The genetics and neuropathology of parkinson's disease. *Acta Neuropathol*, **124**(3), 325–38.

Hu, D., Smith, E. R., Garruss, A. S., Mohaghegh, N., Varberg, J. M., Lin, C., Jackson, J., Gao, X., Saraf, A., Florens, L., Washburn, M. P., Eissenberg, J. C., and Shilatifard, A. (2013a). The little elongation complex functions at initiation and elongation phases of snrna gene transcription. *Mol Cell*, **51**(4), 493–505.

Hu, D., Garruss, A. S., Gao, X., Morgan, M. A., Cook, M., Smith, E. R., and Shilatifard, A. (2013b). The mll2 branch of the compass family regulates bivalent promoters in mouse embryonic stem cells. *Nat Struct Mol Biol*, **20**(9), 1093–7.

Hua, Y. and Zhou, J. (2004a). Rpp20 interacts with smn and is re-distributed into smn granules in response to stress. *Biochem Biophys Res Commun*, **314**(1), 268–76.

Hua, Y. and Zhou, J. (2004b). Survival motor neuron protein facilitates assembly of stress granules. *FEBS Lett*, **572**(1-3), 69–74.

Huang, R., Jaritz, M., Guenzl, P., Vlatkovic, I., Sommer, A., Tamir, I. M., Marks, H., Klampfl, T., Kralovics, R., Stunnenberg, H. G., Barlow, D. P., and Pauler, F. M. (2011). An rna-seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncrnas. *PLoS One*, **6**(11), e27288.

Huntzinger, E., Boisset, S., Saveanu, C., Benito, Y., Geissmann, T., Namane, A., Lina, G., Etienne, J., Ehresmann, B., Ehresmann, C., Jacquier, A., Vandenesch, F., and Romby, P. (2005). Staphylococcus aureus rnaiii and the endoribonuclease iii coordinately regulate spa gene expression. *EMBO J*, **24**(4), 824–35.

Ionescu, D., Voss, B., Oren, A., Hess, W. R., and Muro-Pastor, A. M. (2010). Heterocyst-specific transcription of nsir1, a non-coding rna encoded in a tandem array of direct repeats in cyanobacteria. *J Mol Biol*, **398**(2), 177–88.

Jady, B. E., Darzacq, X., Tucker, K. E., Matera, A. G., Bertrand, E., and Kiss, T. (2003). Modification of sm small nuclear rnas occurs in the nucleoplasmic cajal body following import from the cytoplasm. *EMBO J*, **22**(8), 1878–88.

Jady, B. E., Bertrand, E., and Kiss, T. (2004). Human telomerase rna and box h/aca scarnas share a common cajal body-specific localization signal. *J Cell Biol*, **164**(5), 647–52.

Jae, N., Wang, P., Gu, T., Huhn, M., Palfi, Z., Urlaub, H., and Bindereif, A. (2010). Essential role of a trypanosome u4-specific sm core protein in small nuclear ribonucleoprotein assembly and splicing. *Eukaryot Cell*, **9**(3), 379–86.

Jeck, W. R., Sorrentino, J. A., Wang, K., Slevin, M. K., Burd, C. E., Liu, J., Marzluff, W. F., and Sharpless, N. E. (2013). Circular rnas are abundant, conserved, and associated with alu repeats. *RNA*, **19**(2), 141–57.

Jensen, R. C., Wang, Y., Hardin, S. B., and Stumph, W. E. (1998). The proximal sequence element (pse) plays a major role in establishing the rna polymerase specificity of drosophila u-snrna genes. *Nucleic Acids Res*, **26**(2), 616–22.

Jia, Y., Mu, J. C., and Ackerman, S. L. (2012). Mutation of a u2 snrna gene causes global disruption of alternative splicing and neurodegeneration. *Cell*, **148**(1-2), 296–308.

Jung, C. H., Hansen, M. A., Makunin, I. V., Korbie, D. J., and Mattick, J. S. (2010). Identification of novel non-coding rnas using profiles of short sequence reads from next generation sequencing data. *BMC Genomics*, **11**, 77.

Kaida, D., Berg, M. G., Younis, I., Kasim, M., Singh, L. N., Wan, L., and Dreyfuss, G. (2010). U1 snrnp protects pre-mrnas from premature cleavage and polyadenylation. *Nature*, **468**(7324), 664–8.

Kambach, C. and Mattaj, I. W. (1994). Nuclear transport of the u2 snrnp-specific u2b" protein is mediated by both direct and indirect signalling mechanisms. *J Cell Sci*, **107 (Pt 7)**, 1807–16.

Kambach, C., Walke, S., Young, R., Avis, J. M., de la Fortelle, E., Raker, V. A., Luhrmann, R., Li, J., and Nagai, K. (1999). Crystal structures of two sm protein complexes and their implications for the assembly of the spliceosomal snrnps. *Cell*, **96**(3), 375–87.

Karijolich, J. and Yu, Y. T. (2010). Spliceosomal snrna modifications and their function. *RNA Biol*, **7**(2), 192–204.

Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). Rip-chip: the isolation and identification of mrnas, micrornas and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc*, **1**(1), 302–7.

Kiebler, M. A. and Bassell, G. J. (2006). Neuronal rna granules: movers and makers. *Neuron*, **51**(6), 685–90.

Kiss, T. (2001). Small nucleolar rna-guided post-transcriptional modification of cellular rnas. *EMBO J*, **20**(14), 3617–22.

Kondrashov, A. V., Kiefmann, M., Ebnet, K., Khanam, T., Muddashetty, R. S., and Brosius, J. (2005). Inhibitory effect of naked neural bc1 rna or bc200 rna on eukaryotic in vitro translation systems is reversed by poly(a)-binding protein (pabp). *J Mol Biol*, **353**(1), 88–103.

Koonin, E. V. (2003). Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, **1**(2), 127–36.

Korf, G. M., Botros, I. W., and Stumph, W. E. (1988). Developmental and tissue-specific expression of u4 small nuclear rna genes. *Mol Cell Biol*, **8**(12), 5566–9.

Kshirsagar, M. and Parker, R. (2004). Identification of edc3p as an enhancer of mrna decapping in saccharomyces cerevisiae. *Genetics*, **166**(2), 729–39.

Kudla, G., Granneman, S., Hahn, D., Beggs, J. D., and Tollervey, D. (2011). Cross-linking, ligation, and sequencing of hybrids reveals rna-rna interactions in yeast. *Proc Natl Acad Sci U S A*, **108**(24), 10010–5.

Kyburz, A., Friedlein, A., Langen, H., and Keller, W. (2006). Direct interactions between subunits of cpsf and the u2 snrnp contribute to the coupling of pre-mrna 3' end processing and splicing. *Mol Cell*, **23**(2), 195–205.

Lall, S., Friedman, C. C., Jankowska-Anyszka, M., Stepinski, J., Darzynkiewicz, E., and Davis, R. E. (2004). Contribution of trans-splicing, 5' -leader length, cap-poly(a) synergism, and initiation factors to nematode translation in an ascaris suum embryo cell-free system. *J Biol Chem*, **279**(44), 45573–85.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Methods*, **9**(4), 357–9.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, **10**(3), R25.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007). Clustal w and clustal x version 2.0. *Bioinformatics*, **23**(21), 2947–8.

Lee, K. M. and Tarn, W. Y. (2013). Coupling pre-mrna processing to transcription on the rna factory assembly line. *RNA Biol*, **10**(3), 380–90.

Lee, L., Davies, S. E., and Liu, J. L. (2009). The spinal muscular atrophy protein smn affects drosophila germline nuclear organization through the u body-p body pathway. *Dev Biol*, **332**(1), 142–55.

Lee, S. I., Murthy, S. C., Trimble, J. J., Desrosiers, R. C., and Steitz, J. A. (1988). Four novel u rnas are encoded by a herpesvirus. *Cell*, **54**(5), 599–607.

Lefebvre, S., Burglen, L., Reboullet, S., Clermont, O., Burlet, P., Viollet, L., Benichou, B., Cruaud, C., Millasseau, P., Zeviani, M., and et al. (1995). Identification and characterization of a spinal muscular atrophy-determining gene. *Cell*, **80**(1), 155–65.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S., and Bassler, B. L. (2004). The small rna chaperone hfq and multiple small rnas control quorum sensing in vibrio harveyi and vibrio cholerae. *Cell*, **118**(1), 69–82.

Leonardi, J., Box, J. A., Bunch, J. T., and Baumann, P. (2008). Ter1, the rna subunit of fission yeast telomerase. *Nat Struct Mol Biol*, **15**(1), 26–33.

Lerner, M. R. and Steitz, J. A. (1979). Antibodies to small nuclear rnas complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc Natl Acad Sci U S A*, **76**(11), 5495–9.

Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. (1980). Are snrnps involved in splicing? *Nature*, **283**(5743), 220–4.

Lestrade, L. and Weber, M. J. (2006). snorna-lbme-db, a comprehensive database of human h/aca and c/d box snornas. *Nucleic Acids Res*, **34**(Database issue), D158–62.

Leung, A. K., Nagai, K., and Li, J. (2011). Structure of the spliceosomal u4 snrnp core domain and its implication for snrnp biogenesis. *Nature*, **473**(7348), 536–9.

Li, H., Ruan, J., and Durbin, R. (2008). Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res*, **18**(11), 1851–8.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The sequence alignment/map format and samtools. *Bioinformatics*, **25**(16), 2078–9.

Lin, P. C. and Xu, R. M. (2012). Structure and assembly of the sf3a splicing factor complex of u2 snrnp. *EMBO J*, **31**(6), 1579–90.

Lipson, S. E. and Hearst, J. E. (1988). Psoralen cross-linking of ribosomal rna. *Methods Enzymol*, **164**, 330–41.

Liu, J. L. and Gall, J. G. (2007). U bodies are cytoplasmic structures that contain uridine-rich small nuclear ribonucleoproteins and associate with p bodies. *Proc Natl Acad Sci U S A*, **104**(28), 11655–9.

Liu, J. L., Murphy, C., Buszczak, M., Clatterbuck, S., Goodman, R., and Gall, J. G. (2006). The drosophila melanogaster cajal body. *J Cell Biol*, **172**(6), 875–84.

Liu, P., Gucwa, A., Stover, M. L., Buck, E., Lichtler, A., and Rowe, D. (2002). Analysis of inhibitory action of modified u1 snrnas on target gene expression: discrimination of two rna targets differing by a 1 bp mismatch. *Nucleic Acids Res*, **30**(11), 2329–39.

Liu, Q., Fischer, U., Wang, F., and Dreyfuss, G. (1997). The spinal muscular atrophy disease gene product, smn, and its associated protein sip1 are in a complex with spliceosomal snrnp proteins. *Cell*, **90**(6), 1013–21.

Liu, W., Zhao, Y., Cui, P., Lin, Q., Ding, F., Xin, C., Tan, X., Song, S., Yu, J., and Hu, S. (2011). Thousands of novel transcripts identified in mouse cerebrum, testis, and es cells based on ribo-minus rna sequencing. *Front Genet*, **2**, 93.

Lo, P. C. and Mount, S. M. (1990). Drosophila melanogaster genes for u1 snrna variants and their expression during development. *Nucleic Acids Res*, **18**(23), 6971–9.

Lobo, S. M., Marzluff, W. F., Seufert, A. C., Dean, W. L., Schultz, G. A., Simerly, C., and Schatten, G. (1988). Localization and expression of u1 rna in early mouse embryo development. *Dev Biol*, **127**(2), 349–61.

Lossky, M., Anderson, G. J., Jackson, S. P., and Beggs, J. (1987). Identification of a yeast snrnp protein and detection of snrnp-snrnp interactions. *Cell*, **51**(6), 1019–26.

Lu, Z. and Matera, A. G. (2014). Vicinal: a method for the determination of ncrna ends using chimeric reads from rna-seq experiments. *Nucleic Acids Res*.

Lu, Z., Guan, X., Schmidt, C. A., and Matera, A. G. (2014). Rip-seq analysis of eukaryotic sm proteins identifies three major categories of sm-containing ribonucleoproteins. *Genome Biol*, **15**(1), R7.

Lukong, K. E., Chang, K. W., Khandjian, E. W., and Richard, S. (2008). Rna-binding proteins in human genetic disease. *Trends Genet*, **24**(8), 416–25.

Lund, E. (1988). Heterogeneity of human u1 snrnas. *Nucleic Acids Res*, **16**(13), 5813–26.

Lund, E. and Dahlberg, J. E. (1987). Differential accumulation of u1 and u4 small nuclear rnas during xenopus development. *Genes Dev*, **1**(1), 39–46.

Lund, E., Kahan, B., and Dahlberg, J. E. (1985). Differential control of u1 small nuclear rna expression during mouse development. *Science*, **229**(4719), 1271–4.

Lund, E., Bostock, C. J., and Dahlberg, J. E. (1987). The transcription of xenopus laevis embryonic u1 snrna genes changes when oocytes mature into eggs. *Genes Dev*, **1**(1), 47–56.

Lunde, B. M., Moore, C., and Varani, G. (2007). Rna-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, **8**(6), 479–90.

Lyons, S. M., Ricciardi, A. S., Guo, A. Y., Kambach, C., and Marzluff, W. F. (2014). The c-terminal extension of lsm4 interacts directly with the 3' end of the histone mrnp and is required for efficient histone mrna degradation. *RNA*, **20**(1), 88–102.

Madhani, H. D. and Guthrie, C. (1992). A novel base-pairing interaction between u2 and u6 snrnas suggests a mechanism for the catalytic activation of the spliceosome. *Cell*, **71**(5), 803–17.

Maki, K., Uno, K., Morita, T., and Aiba, H. (2008). Rna, but not protein partners, is directly responsible for translational silencing by a bacterial hfq-binding small rna. *Proc Natl Acad Sci U S A*, **105**(30), 10332–7.

Malca, H., Shomron, N., and Ast, G. (2003). The u1 snrnp base pairs with the 5' splice site within a penta-snrnp complex. *Mol Cell Biol*, **23**(10), 3442–55.

Maniatis, T. and Reed, R. (2002). An extensive network of coupling among gene expression machines. *Nature*, **416**(6880), 499–506.

Markham, N. R. and Zuker, M. (2008). Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, **453**, 3–31.

Marquez, S. M., Harris, J. K., Kelley, S. T., Brown, J. W., Dawson, S. C., Roberts, E. C., and Pace, N. R. (2005). Structural implications of novel diversity in eucaryal rnase p rna. *RNA*, **11**(5), 739–51.

Marz, M., Kirsten, T., and Stadler, P. F. (2008). Evolution of spliceosomal snrna genes in metazoan animals. *J Mol Evol*, **67**(6), 594–607.

Masse, E., Escorcia, F. E., and Gottesman, S. (2003). Coupled degradation of a small regulatory rna and its mrna targets in escherichia coli. *Genes Dev*, **17**(19), 2374–83.

Massenet, S., Pellizzoni, L., Paushkin, S., Mattaj, I. W., and Dreyfuss, G. (2002). The smn complex is associated with snrnps throughout their cytoplasmic assembly pathway. *Mol Cell Biol*, **22**(18), 6533–41.

Matera, A. G. (1999). Nuclear bodies: multifaceted subdomains of the interchromatin space. *Trends Cell Biol*, **9**(8), 302–9.

Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nat Rev Mol Cell Biol*, **15**(2), 108–21.

Matera, A. G. and Ward, D. C. (1993). Nucleoplasmic organization of small nuclear ribonucleoproteins in cultured human cells. *J Cell Biol*, **121**(4), 715–27.

Matera, A. G., Terns, R. M., and Terns, M. P. (2007). Non-coding rnas: lessons from the small nuclear and small nucleolar rnas. *Nat Rev Mol Cell Biol*, **8**(3), 209–20.

Matera, A. G., Izaguire-Sierra, M., Praveen, K., and Rajendra, T. K. (2009). Nuclear bodies: random aggregates of sticky proteins or crucibles of macromolecular assembly? *Dev Cell*, **17**(5), 639–47.

Mattaj, I. W. (1986). Cap trimethylation of u snrna is cytoplasmic and dependent on u snrnp protein binding. *Cell*, **46**(6), 905–11.

Mattaj, I. W. and Hamm, J. (1989). Regulated splicing in early development and stage-specific u snrnps. *Development*, **105**(2), 183–9.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding rna. *Hum Mol Genet*, **15 Spec No 1**, R17–29.

Mayes, A. E., Verdone, L., Legrain, P., and Beggs, J. D. (1999). Characterization of sm-like proteins in yeast and their association with u6 snrna. *EMBO J*, **18**(15), 4321–31.

McManus, C. J., Duff, M. O., Eipper-Mains, J., and Graveley, B. R. (2010). Global analysis of trans-splicing in drosophila. *Proc Natl Acad Sci U S A*, **107**(29), 12975–9.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S. D., Gregersen, L. H., Munschauer, M., Loewer, A., Ziebold, U., Landthaler, M., Kocks, C., le Noble, F., and Rajewsky, N. (2013). Circular rnas are a large class of animal rnas with regulatory potency. *Nature*, **495**(7441), 333–8.

Mili, S. and Steitz, J. A. (2004). Evidence for reassociation of rna-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA*, **10**(11), 1692–4.

Miller, O. L., J., Hamkalo, B. A., and Thomas, C. A., J. (1970). Visualization of bacterial genes in action. *Science*, **169**(3943), 392–5.

Mizuno, T., Chou, M. Y., and Inouye, M. (1984). A unique mechanism regulating gene expression: translational inhibition by a complementary rna transcript (micrna). *Proc Natl Acad Sci U S A*, **81**(7), 1966–70.

Mohanty, B. K., Maples, V. F., and Kushner, S. R. (2004). The sm-like protein hfq regulates polyadenylation dependent mrna decay in escherichia coli. *Mol Microbiol*, **54**(4), 905–20.

Moll, I., Afonyushkin, T., Vytvytska, O., Kaberdin, V. R., and Blasi, U. (2003). Coincident hfq binding and rnase e cleavage sites on mrna and small regulatory rnas. *RNA*, **9**(11), 1308–14.

Moller, T., Franch, T., Hojrup, P., Keene, D. R., Bachinger, H. P., Brennan, R. G., and Valentin-Hansen, P. (2002a). Hfq: a bacterial sm-like protein that mediates rna-rna interaction. *Mol Cell*, **9**(1), 23–30.

Moller, T., Franch, T., Udesen, C., Gerdes, K., and Valentin-Hansen, P. (2002b). Spot 42 rna mediates discoordinate expression of the e. coli galactose operon. *Genes Dev*, **16**(13), 1696–706.

Morris, A. R., Mukherjee, N., and Keene, J. D. (2008). Ribonomic analysis of human pum1 reveals cis-trans conservation across species despite evolution of diverse mrna target sets. *Mol Cell Biol*, **28**(12), 4093–103.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, **5**(7), 621–8.

Mouaikel, J., Narayanan, U., Verheggen, C., Matera, A. G., Bertrand, E., Tazi, J., and Bordonne, R. (2003). Interaction between the small-nuclear-rna cap hypermethylase and the spinal muscular atrophy protein, survival of motor neuron. *EMBO Rep*, **4**(6), 616–22.

Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K., and Makalowski, W. (2007). Spliceosomal small nuclear rna genes in 11 insect genomes. *RNA*, **13**(1), 5–14.

Mousavi, K., Zare, H., Dell'orso, S., Grontved, L., Gutierrez-Cruz, G., Derfoul, A., Hager, G. L., and Sartorelli, V. (2013). ernas promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell*, **51**(5), 606–17.

Mowry, K. L. and Steitz, J. A. (1987). Identification of the human u7 snrnp as one of several factors involved in the 3' end maturation of histone premessenger rna's. *Science*, **238**(4834), 1682–7.

Mura, C., Phillips, M., Kozhukhovsky, A., and Eisenberg, D. (2003). Structure and assembly of an augmented sm-like archaeal protein 14-mer. *Proc Natl Acad Sci U S A*, **100**(8), 4539–44.

Mura, C., Randolph, P. S., Patterson, J., and Cozen, A. E. (2013). Archaeal and eukaryotic homologs of hfq: A structural and evolutionary perspective on sm function. *RNA Biol*, **10**(4), 636–51.

Nagai, K., Muto, Y., Pomeranz Krummel, D. A., Kambach, C., Ignjatovic, T., Walke, S., and Kuglstatter, A. (2001). Structure and assembly of the spliceosomal snrnps. novartis medal lecture. *Biochem Soc Trans*, **29**(Pt 2), 15–26.

Nakamura, R., Takeuchi, R., Takata, K., Shimanouchi, K., Abe, Y., Kanai, Y., Ruike, T., Ihara, A., and Sakaguchi, K. (2008). Trf4 is involved in polyadenylation of snrnas in drosophila melanogaster. *Mol Cell Biol*, **28**(21), 6620–31.

Narayanan, U., Ospina, J. K., Frey, M. R., Hebert, M. D., and Matera, A. G. (2002). Smn, the spinal muscular atrophy protein, forms a pre-import snrnp complex with snurportin1 and importin beta. *Hum Mol Genet*, **11**(15), 1785–95.

Nash, M. A., Sakallah, S., Santiago, C., Yu, J. C., and Marzluff, W. F. (1989). A developmental switch in sea urchin u1 rna. *Dev Biol*, **134**(2), 289–96.

Natalizio, A. H. and Matera, A. G. (2013). Identification and characterization of drosophila snurportin reveals a role for the import receptor moleskin/importin-7 in snrnp biogenesis. *Mol Biol Cell*, **24**(18), 2932–42.

Neugebauer, K. M. (2002). On the importance of being co-transcriptional. *J Cell Sci*, **115**(Pt 20), 3865–71.

Neuman de Vegvar, H. E. and Dahlberg, J. E. (1990). Nucleocytoplasmic transport and processing of small nuclear rna precursors. *Mol Cell Biol*, **10**(7), 3365–75.

Nizami, Z., Deryusheva, S., and Gall, J. G. (2010). The cajal body and histone locus body. *Cold Spring Harb Perspect Biol*, **2**(7), a000653.

Noble, S. M. and Guthrie, C. (1996). Transcriptional pulse-chase analysis reveals a role for a novel snrnp-associated protein in the manufacture of spliceosomal snrnps. *EMBO J*, **15**(16), 4368–79.

Nott, A., Le Hir, H., and Moore, M. J. (2004). Splicing enhances translation in mammalian cells: an additional function of the exon junction complex. *Genes Dev*, **18**(2), 210–22.

Ntini, E., Jarvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jorgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P. R., Andersen, P. K., Preker, P., Valen, E., Zhao, X., Pelechano, V., Steinmetz, L. M., Sandelin, A., and Jensen, T. H. (2013). Polyadenylation site-induced decay of upstream transcripts enforces promoter directionality. *Nat Struct Mol Biol*, **20**(8), 923–8.

Ohno, M., Segref, A., Bachi, A., Wilm, M., and Mattaj, I. W. (2000). Phax, a mediator of u snrna nuclear export whose activity is regulated by phosphorylation. *Cell*, **101**(2), 187–98.

O'Reilly, D., Dienstbier, M., Cowley, S. A., Vazquez, P., Drozdz, M., Taylor, S., James, W. S., and Murphy, S. (2013). Differentially expressed, variant u1 snrnas regulate gene expression in human cells. *Genome Res*, **23**(2), 281–91.

Palacios, I., Weis, K., Klebe, C., Mattaj, I. W., and Dingwall, C. (1996). Ran/tc4 mutants identify a common requirement for snrnp and protein import into the nucleus. *J Cell Biol*, **133**(3), 485–94.

Palacios, I., Hetzer, M., Adam, S. A., and Mattaj, I. W. (1997). Nuclear import of u snrnps requires importin beta. *EMBO J*, **16**(22), 6783–92.

Pante, N., Jarmolowski, A., Izaurralde, E., Sauder, U., Baschong, W., and Mattaj, I. W. (1997). Visualizing nuclear export of different classes of rna by electron microscopy. *RNA*, **3**(5), 498–513.

Park, E., Williams, B., Wold, B. J., and Mortazavi, A. (2012). Rna editing in the human encode rna-seq data. *Genome Res*, **22**(9), 1626–33.

Parker, R. and Sheth, U. (2007). P bodies and the control of mrna translation and degradation. *Mol Cell*, **25**(5), 635–46.

Patel, A. A. and Steitz, J. A. (2003). Splicing double: insights from the second spliceosome. *Nat Rev Mol Cell Biol*, **4**(12), 960–70.

Pavelitz, T., Rusche, L., Matera, A. G., Scharf, J. M., and Weiner, A. M. (1995). Concerted evolution of the tandem array encoding primate u2 snrna occurs in situ, without changing the cytological context of the rnu2 locus. *EMBO J*, **14**(1), 169–77.

Pavelitz, T., Liao, D., and Weiner, A. M. (1999). Concerted evolution of the tandem array encoding primate u2 snrna (the rnu2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences. *EMBO J*, **18**(13), 3783–92.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, **185**, 40.

Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S., Kent, J., Miller, W., and Haussler, D. (2006). Identification and classification of conserved rna secondary structures in the human genome. *PLoS Comput Biol*, **2**(4), e33.

Pellizzoni, L., Kataoka, N., Charroux, B., and Dreyfuss, G. (1998). A novel function for smn, the spinal muscular atrophy disease gene product, in pre-mrna splicing. *Cell*, **95**(5), 615–24.

Pellizzoni, L., Yong, J., and Dreyfuss, G. (2002). Essential role for the smn complex in the specificity of snrnp assembly. *Science*, **298**(5599), 1775–9.

Pereira-Simon, S., Sierra-Montes, J. M., Ayesh, K., Martinez, L., Socorro, A., and Herrera, R. J. (2004). Variants of u1 small nuclear rna assemble into spliceosomal complexes. *Insect Mol Biol*, **13**(2), 189–94.

Pillai, R. S., Grimmler, M., Meister, G., Will, C. L., Luhrmann, R., Fischer, U., and Schumperli, D. (2003). Unique sm core structure of u7 snrnps: assembly by a specialized smn complex and the role of a new component, lsm11, in histone rna processing. *Genes Dev*, **17**(18), 2321–33.

Pinto, A. L. and Steitz, J. A. (1989). The mammalian analogue of the yeast prp8 splicing protein is present in the u4/5/6 small nuclear ribonucleoprotein particle and the spliceosome. *Proc Natl Acad Sci U S A*, **86**(22), 8742–6.

Pomeranz Krummel, D. A., Oubridge, C., Leung, A. K., Li, J., and Nagai, K. (2009). Crystal structure of human spliceosomal u1 snrnp at 5.5 a resolution. *Nature*, **458**(7237), 475–80.

Prasad, M., Jang, A. C., Starz-Gaiano, M., Melani, M., and Montell, D. J. (2007). A protocol for culturing drosophila melanogaster stage 9 egg chambers for live imaging. *Nat Protoc*, **2**(10), 2467–73.

Praveen, K., Wen, Y., and Matera, A. G. (2012). A drosophila model of spinal muscular atrophy uncouples snrnp biogenesis functions of survival motor neuron from locomotion and viability defects. *Cell Rep*, **1**(6), 624–31.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M. S., Mapendano, C. K., Schierup, M. H., and Jensen, T. H. (2008). Rna exosome depletion reveals transcription upstream of active human promoters. *Science*, **322**(5909), 1851–4.

Proudfoot, N. J., Furger, A., and Dye, M. J. (2002). Integrating mrna processing with transcription. *Cell*, **108**(4), 501–12.

Qi, H., Watanabe, T., Ku, H. Y., Liu, N., Zhong, M., and Lin, H. (2011). The yb body, a major site for piwi-associated rna biogenesis and a gateway for piwi expression and transport to the nucleus in somatic cells. *J Biol Chem*, **286**(5), 3789–97.

Quinones-Coello, A. T., Petrella, L. N., Ayers, K., Melillo, A., Mazzalupo, S., Hudson, A. M., Wang, S., Castiblanco, C., Buszczak, M., Hoskins, R. A., and Cooley, L. (2007a). Exploring strategies for protein trapping in drosophila. *Genetics*, **175**(3), 1089–104.

Quinones-Coello, A. T., Petrella, L. N., Ayers, K., Melillo, A., Mazzalupo, S., Hudson, A. M., Wang, S., Castiblanco, C., Buszczak, M., Hoskins, R. A., and Cooley, L. (2007b). Exploring strategies for protein trapping in drosophila. *Genetics*, **175**(3), 1089–104.

Rajendra, T. K., Gonsalvez, G. B., Walker, M. P., Shpargel, K. B., Salz, H. K., and Matera, A. G. (2007). A drosophila melanogaster model of spinal muscular atrophy reveals a function for smn in striated muscle. *J Cell Biol*, **176**(6), 831–41.

Rajendra, T. K., Praveen, K., and Matera, A. G. (2010). Genetic analysis of nuclear bodies: from nondeterministic chaos to deterministic order. *Cold Spring Harb Symp Quant Biol*, **75**, 365–74.

Ramaswami, G., Zhang, R., Piskol, R., Keegan, L. P., Deng, P., O'Connell, M. A., and Li, J. B. (2013). Identifying rna editing sites using rna sequencing data alone. *Nat Methods*, **10**(2), 128–32.

Regnier, P. and Hajnsdorf, E. (2013). The interplay of hfq, poly(a) polymerase i and exoribonucleases at the 3' ends of rnas resulting from rho-independent termination: A tentative model. *RNA Biol*, **10**(4), 602–9.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microrna/target duplexes. *RNA*, **10**(10), 1507–17.

Repoila, F., Majdalani, N., and Gottesman, S. (2003). Small non-coding rnas, co-ordinators of adaptation processes in escherichia coli: the rpos paradigm. *Mol Microbiol*, **48**(4), 855–61.

Rhee, H. S. and Pugh, B. F. (2011). Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, **147**(6), 1408–19.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends Genet*, **16**(6), 276–7.

Richard, P., Darzacq, X., Bertrand, E., Jady, B. E., Verheggen, C., and Kiss, T. (2003). A common sequence motif determines the cajal body-specific localization of box h/aca scarnas. *EMBO J*, **22**(16), 4283–93.

Riley, K. J. and Steitz, J. A. (2013). The "observer effect" in genome-wide surveys of protein-rna interactions. *Mol Cell*, **49**(4), 601–4.

Roy, G., Miron, M., Khaleghpour, K., Lasko, P., and Sonenberg, N. (2004). The drosophila poly(a) binding protein-interacting protein, dpaip2, is a novel effector of cell growth. *Mol Cell Biol*, **24**(3), 1143–54.

Roy, J., Zheng, B., Rymond, B. C., and Woolford, J. L., J. (1995). Structurally related but functionally distinct yeast sm d core small nuclear ribonucleoprotein particle proteins. *Mol Cell Biol*, **15**(1), 445–55.

Roy, S. W. and Gilbert, W. (2006). The evolution of spliceosomal introns: patterns, puzzles and progress. *Nat Rev Genet*, **7**(3), 211–21.

Ruan, X. and Ruan, Y. (2012). Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (rna-pet). *Methods Mol Biol*, **809**, 535–62.

Ruiz, N. and Silhavy, T. J. (2003). Constitutive activation of the escherichia coli pho regulon upregulates rpos translation in an hfq-dependent fashion. *J Bacteriol*, **185**(20), 5984–92.

Rymond, B. C. (1993). Convergent transcripts of the yeast prp38-smd1 locus encode two essential splicing factors, including the d1 core polypeptide of small nuclear ribonucleoprotein particles. *Proc Natl Acad Sci U S A*, **90**(3), 848–52.

Sajic, R., Lee, K., Asai, K., Sakac, D., Branch, D. R., Upton, C., and Cochrane, A. (2007). Use of modified u1 snrnas to inhibit hiv-1 replication. *Nucleic Acids Res*, **35**(1), 247–55.

Sakarya, O., Breu, H., Radovich, M., Chen, Y., Wang, Y. N., Barbacioru, C., Utiramerur, S., Whitley, P. P., Brockman, J. P., Vatta, P., Zhang, Z., Popescu, L., Muller, M. W., Kudlingar, V., Garg, N., Li, C. Y., Kong, B. S., Bodeau, J. P., Nutter, R. C., Gu, J., Bramlett, K. S., Ichikawa, J. K., Hyland, F. C., and Siddiqui, A. S. (2012). Rna-seq mapping and detection of gene fusions with a suffix array algorithm. *PLoS Comput Biol*, **8**(4), e1002464.

Saldanha, A. J. (2004). Java treeview–extensible visualization of microarray data. *Bioinformatics*, **20**(17), 3246–8.

Salgado-Garrido, J., Bragado-Nilsson, E., Kandels-Lewis, S., and Seraphin, B. (1999). Sm and sm-like proteins assemble in two related complexes of deep evolutionary origin. *EMBO J*, **18**(12), 3451–62.

Salles, F. J. and Strickland, S. (1995). Rapid and sensitive analysis of mrna polyadenylation states by pcr. *PCR Methods Appl*, **4**(6), 317–21.

Salzman, J., Gawad, C., Wang, P. L., Lacayo, N., and Brown, P. O. (2012). Circular rnas are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One*, **7**(2), e30733.

Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L., and Brown, P. O. (2013). Cell-type specific features of circular rna expression. *PLoS Genet*, **9**(9), e1003777.

Salzman, L. A. and Fabisch, P. (1979). Nucleotide sequence of the self-priming 3' terminus of the single-stranded dna extracted from the parvovirus kilham rat virus. *J Virol*, **30**(3), 946–50.

Santiago, C. and Marzluff, W. F. (1989). Expression of the u1 rna gene repeat during early sea urchin development: evidence for a switch in u1 rna genes during development. *Proc Natl Acad Sci U S A*, **86**(8), 2572–6.

Sastry, S. S., Ross, B. M., and P'Arraga, A. (1997). Cross-linking of dna-binding proteins to dna with psoralen and psoralen furan-side monoadducts. comparison of action spectra with dna-dna cross-linking. *J Biol Chem*, **272**(6), 3715–23.

Sauter, C., Basquin, J., and Suck, D. (2003). Sm-like proteins in eubacteria: the crystal structure of the hfq protein from escherichia coli. *Nucleic Acids Res*, **31**(14), 4091–8.

Sauterer, R. A., Feeney, R. J., and Zieve, G. W. (1988). Cytoplasmic assembly of snrnp particles from stored proteins and newly transcribed snrna's in l929 mouse fibroblasts. *Exp Cell Res*, **176**(2), 344–59.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, **270**(5235), 467–70.

Schisa, J. A. (2012). New insights into the regulation of rnp granule assembly in oocytes. *Int Rev Cell Mol Biol*, **295**, 233–89.

Schneider, I. (1972). Cell lines derived from late embryonic stages of drosophila melanogaster. *J Embryol Exp Morphol*, **27**(2), 353–65.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, **18**(20), 6097–100.

Scotto-Lavino, E., Du, G., and Frohman, M. A. (2006a). 3' end cdna amplification using classic race. *Nat Protoc*, **1**(6), 2742–5.

Scotto-Lavino, E., Du, G., and Frohman, M. A. (2006b). 5' end cdna amplification using classic race. *Nat Protoc*, **1**(6), 2555–62.

Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., and Sharp, P. A. (2008). Divergent transcription from active promoters. *Science*, **322**(5909), 1849–51.

Seto, A. G., Zaug, A. J., Sobel, S. G., Wolin, S. L., and Cech, T. R. (1999). Saccharomyces cerevisiae telomerase is an sm small nuclear ribonucleoprotein particle. *Nature*, **401**(6749), 177–80.

Shen, C. K., Hsieh, T. S., Wang, J. C., and Hearst, J. E. (1977). Photochemical cross-linking of dna-rna helices by psoralen derivatives. *J Mol Biol*, **116**(4), 661–79.

Sierra-Montes, J. M., Freund, A. V., Ruiz, L. M., Szmulewicz, M. N., Rowold, D. J., and Herrera, R. J. (2002). Multiple forms of u2 snrna coexist in the silk moth bombyx mori. *Insect Mol Biol*, **11**(1), 105–14.

Sierra-Montes, J. M., Pereira-Simon, S., Freund, A. V., Ruiz, L. M., Szmulewicz, M. N., and Herrera, R. J. (2003). A diversity of u1 small nuclear rnas in the silk moth bombyx mori. *Insect Biochem Mol Biol*, **33**(1), 29–39.

Sierra-Montes, J. M., Pereira-Simon, S., Smail, S. S., and Herrera, R. J. (2005). The silk moth bombyx mori u1 and u2 snrna variants are differentially expressed. *Gene*, **352**, 127–36.

Simoes-Barbosa, A., Chakrabarti, K., Pearson, M., Benarroch, D., Shuman, S., and Johnson, P. J. (2012). Box h/aca snornas are preferred substrates for the trimethylguanosine synthase in the divergent unicellular eukaryote trichomonas vaginalis. *RNA*, **18**(9), 1656–65.

Simon, M. D., Wang, C. I., Kharchenko, P. V., West, J. A., Chapman, B. A., Alekseyenko, A. A., Borowsky, M. L., Kuroda, M. I., and Kingston, R. E. (2011). The genomic binding sites of a noncoding rna. *Proc Natl Acad Sci U S A*, **108**(51), 20497–502.

Sittka, A., Lucchini, S., Papenfort, K., Sharma, C. M., Rolle, K., Binnewies, T. T., Hinton, J. C., and Vogel, J. (2008). Deep sequencing analysis of small noncoding rna and mrna targets of the global post-transcriptional regulator, hfq. *PLoS Genet*, **4**(8), e1000163.

Skripkin, E., Isel, C., Marquet, R., Ehresmann, B., and Ehresmann, C. (1996). Psoralen crosslinking between human immunodeficiency virus type 1 rna and primer trna3(lys). *Nucleic Acids Res*, **24**(3), 509–14.

Sledjeski, D. D., Whitman, C., and Zhang, A. (2001). Hfq is necessary for regulation by the untranslated rna dsra. *J Bacteriol*, **183**(6), 1997–2005.

Smail, S. S., Ayesh, K., Sierra-Montes, J. M., and Herrera, R. J. (2006). U6 snrna variants isolated from the posterior silk gland of the silk moth bombyx mori. *Insect Biochem Mol Biol*, **36**(6), 454–65.

Smith, E. R., Lin, C., Garrett, A. S., Thornton, J., Mohaghegh, N., Hu, D., Jackson, J., Saraf, A., Swanson, S. K., Seidel, C., Florens, L., Washburn, M. P., Eissenberg, J. C., and Shilatifard, A. (2011). The little elongation complex regulates small nuclear rna transcription. *Mol Cell*, **44**(6), 954–65.

Smolinski, D. J., Wrobel, B., Noble, A., Zienkiewicz, A., and Gorska-Brylass, A. (2011). Periodic expression of sm proteins parallels formation of nuclear cajal bodies and cytoplasmic snrnp-rich bodies. *Histochem Cell Biol*, **136**(5), 527–41.

Sobrero, P. and Valverde, C. (2012). The bacterial protein hfq: much more than a mere rna-binding factor. *Crit Rev Microbiol*, **38**(4), 276–99.

Song, T. and Wai, S. N. (2009). A novel srna that modulates virulence and environmental fitness of vibrio cholerae. *RNA Biol*, **6**(3), 254–8.

Sontheimer, E. J. and Steitz, J. A. (1992). Three novel functional variants of human u5 small nuclear rna. *Mol Cell Biol*, **12**(2), 734–46.

Spector, D. L., Lark, G., and Huang, S. (1992). Differences in snrnp localization between transformed and nontransformed cells. *Mol Biol Cell*, **3**(5), 555–69.

Spiller, M. P., Reijns, M. A., and Beggs, J. D. (2007). Requirements for nuclear localization of the lsm2-8p complex and competition between nuclear and cytoplasmic lsm complexes. *J Cell Sci*, **120**(Pt 24), 4310–20.

Spycher, C., Streit, A., Stefanovic, B., Albrecht, D., Koning, T. H., and Schumperli, D. (1994). 3' end processing of mouse histone pre-mrna: evidence for additional base-pairing between u7 snrna and pre-mrna. *Nucleic Acids Res*, **22**(20), 4023–30.

Stark, H. and Luhrmann, R. (2006). Cryo-electron microscopy of spliceosomal components. *Annu Rev Biophys Biomol Struct*, **35**, 435–57.

Stefanovic, B., Li, J. M., Sakallah, S., and Marzluff, W. F. (1991). Isolation and characterization of developmentally regulated sea urchin u2 snrna genes. *Dev Biol*, **148**(1), 284–94.

Stevens, S. W., Ryan, D. E., Ge, H. Y., Moore, R. E., Young, M. K., Lee, T. D., and Abelson, J. (2002). Composition and functional characterization of the yeast spliceosomal penta-snrnp. *Mol Cell*, **9**(1), 31–44.

Sudhakaran, I. P., Hillebrand, J., Dervan, A., Das, S., Holohan, E. E., Hulsmeier, J., Sarov, M., Parker, R., VijayRaghavan, K., and Ramaswami, M. (2014). Fmrp and ataxin-2 function together in long-term olfactory habituation and neuronal translational control. *Proc Natl Acad Sci U S A*, **111**(1), E99–E108.

Suzuki, T., Izumi, H., and Ohno, M. (2010). Cajal body surveillance of u snrna export complex assembly. *J Cell Biol*, **190**(4), 603–12.

Szakmary, A., Reedy, M., Qi, H., and Lin, H. (2009). The yb protein defines a novel organelle and regulates male germline stem cell self-renewal in drosophila melanogaster. *J Cell Biol*, **185**(4), 613–27.

Takahashi, H., Lassmann, T., Murata, M., and Carninci, P. (2012). 5' end-centered expression profiling using cap-analysis gene expression and next-generation sequencing. *Nat Protoc*, **7**(3), 542–61.

Takata, H., Nishijima, H., Maeshima, K., and Shibahara, K. (2012). The integrator complex is required for integrity of cajal bodies. *J Cell Sci*, **125**(Pt 1), 166–75.

Tan, E. M. and Kunkel, H. G. (1966). Characteristics of a soluble nuclear antigen precipitating with sera of patients with systemic lupus erythematosus. *J Immunol*, **96**(3), 464–71.

Tang, W., Kannan, R., Blanchette, M., and Baumann, P. (2012). Telomerase rna biogenesis involves sequential binding by sm and lsm complexes. *Nature*, **484**(7393), 260–4.

Terns, M. P., Lund, E., and Dahlberg, J. E. (1993). A pre-export u1 snrnp in xenopus laevis oocyte nuclei. *Nucleic Acids Res*, **21**(19), 4569–73.

Tharun, S. (2009). Lsm1-7-pat1 complex: a link between 3' and 5'-ends in mrna decay? *RNA Biol*, **6**(3), 228–32.

Tharun, S. and Parker, R. (2001). Targeting an mrna for decapping: displacement of translation factors and association of the lsm1p-7p complex on deadenylated yeast mrnas. *Mol Cell*, **8**(5), 1075–83.

Tharun, S., He, W., Mayes, A. E., Lennertz, P., Beggs, J. D., and Parker, R. (2000). Yeast sm-like proteins function in mrna decapping and decay. *Nature*, **404**(6777), 515–8.

Thompson, J. F. and Hearst, J. E. (1983). Structure of e. coli 16s rna elucidated by psoralen crosslinking. *Cell*, **32**(4), 1355–65.

Tilgner, H., Raha, D., Habegger, L., Mohiuddin, M., Gerstein, M., and Snyder, M. (2013). Accurate identification and analysis of human mrna isoforms using deep long read sequencing. *G3 (Bethesda)*, **3**(3), 387–97.

Tomasevic, N. and Peculis, B. A. (2002). Xenopus lsm proteins bind u8 snorna via an internal evolutionarily conserved octamer sequence. *Mol Cell Biol*, **22**(12), 4101–12.

Toro, I., Thore, S., Mayer, C., Basquin, J., Seraphin, B., and Suck, D. (2001). Rna binding in an sm core domain: X-ray structure and functional analysis of an archaeal sm protein complex. *EMBO J*, **20**(9), 2293–303.

Toro, I., Basquin, J., Teo-Dreher, H., and Suck, D. (2002). Archaeal sm proteins form heptameric and hexameric complexes: crystal structures of the sm1 and sm2 proteins from the hyperthermophile archaeoglobus fulgidus. *J Mol Biol*, **320**(1), 129–42.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., and Pachter, L. (2012). Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat Protoc*, **7**(3), 562–78.

Tritschler, F., Eulalio, A., Truffault, V., Hartmann, M. D., Helms, S., Schmidt, S., Coles, M., Izaurralde, E., and Weichenrieder, O. (2007). A divergent sm fold in edc3 proteins mediates dcp1 binding and p-body targeting. *Mol Cell Biol*, **27**(24), 8600–11.

Tuiskunen, A., Leparc-Goffart, I., Boubis, L., Monteil, V., Klingstrom, J., Tolou, H. J., Lundkvist, A., and Plumet, S. (2010). Self-priming of reverse transcriptase impairs strand-specific detection of dengue virus rna. *J Gen Virol*, **91**(Pt 4), 1019–27.

Tycowski, K. T., Shu, M. D., Kukoyi, A., and Steitz, J. A. (2009). A conserved wd40 protein binds the cajal body localization signal of scarnp particles. *Mol Cell*, **34**(1), 47–57.

Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegard, J., Holmqvist, E., and Wagner, E. G. (2005). Hfq-dependent regulation of ompa synthesis is mediated by an antisense rna. *Genes Dev*, **19**(19), 2355–66.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., and Darnell, R. B. (2003). Clip identifies nova-regulated rna networks in the brain. *Science*, **302**(5648), 1212–5.

Updike, D. and Strome, S. (2010). P granule assembly and function in caenorhabditis elegans germ cells. *J Androl*, **31**(1), 53–60.

Urlaub, H., Hartmuth, K., Kostka, S., Grelle, G., and Luhrmann, R. (2000). A general approach for identification of rna-protein cross-linking sites within native human spliceosomal small nuclear ribonucleoproteins (snrnps). analysis of rna-protein contacts in native u1 and u4/u6.u5 snrnps. *J Biol Chem*, **275**(52), 41458–68.

Urlaub, H., Raker, V. A., Kostka, S., and Luhrmann, R. (2001). Sm protein-sm site rna interactions within the inner ring of the spliceosomal snrnp core structure. *EMBO J*, **20**(1-2), 187–96.

Valentin-Hansen, P., Eriksen, M., and Udesen, C. (2004). The bacterial sm-like protein hfq: a key player in rna transactions. *Mol Microbiol*, **51**(6), 1525–33.

van Eeden, F. J., Palacios, I. M., Petronczki, M., Weston, M. J., and St Johnston, D. (2001). Barentsz is essential for the posterior localization of oskar mrna and colocalizes with it to the posterior pole. *J Cell Biol*, **154**(3), 511–23.

Vankan, P., McGuigan, C., and Mattaj, I. W. (1990). Domains of u4 and u6 snrnas required for snrnp assembly and splicing complementation in xenopus oocytes. *EMBO J*, **9**(10), 3397–404.

Vidal, V. P., Verdone, L., Mayes, A. E., and Beggs, J. D. (1999). Characterization of u6 snrna-protein interactions. *RNA*, **5**(11), 1470–81.

Vogel, J. (2009). A rough guide to the non-coding rna world of salmonella. *Mol Microbiol*, **71**(1), 1–11.

Vogel, J. and Luisi, B. F. (2011). Hfq and its constellation of rna. *Nat Rev Microbiol*, **9**(8), 578–89.

Vogel, J. and Papenfort, K. (2006). Small non-coding rnas and the bacterial outer membrane. *Curr Opin Microbiol*, **9**(6), 605–11.

Vogel, J., Argaman, L., Wagner, E. G., and Altuvia, S. (2004). The small rna istr inhibits synthesis of an sos-induced toxic peptide. *Curr Biol*, **14**(24), 2271–6.

Walter, P. and Blobel, G. (1982). Signal recognition particle contains a 7s rna essential for protein translocation across the endoplasmic reticulum. *Nature*, **299**(5885), 691–8.

Wang, W., Wang, L., Wu, J., Gong, Q., and Shi, Y. (2013). Hfq-bridged ternary complex is important for translation activation of rpos by dsra. *Nucleic Acids Res*, **41**(11), 5938–48.

Wang, X. H., Aliyari, R., Li, W. X., Li, H. W., Kim, K., Carthew, R., Atkinson, P., and Ding, S. W. (2006). Rna interference directs innate immunity against viruses in adult drosophila. *Science*, **312**(5772), 452–4.

Wang, Z., Gerstein, M., and Snyder, M. (2009a). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1), 57–63.

Wang, Z., Gerstein, M., and Snyder, M. (2009b). Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**(1), 57–63.

Washietl, S., Hofacker, I. L., and Stadler, P. F. (2005). Fast and reliable prediction of noncoding rnas. *Proc Natl Acad Sci U S A*, **102**(7), 2454–9.

Wassarman, D. A., Lee, S. I., and Steitz, J. A. (1989). Nucleotide sequence of hsur 5 rna from herpesvirus saimiri. *Nucleic Acids Res*, **17**(3), 1258.

Wilhelm, J. E., Buszczak, M., and Sayles, S. (2005). Efficient protein trafficking requires trailer hitch, a component of a ribonucleoprotein complex localized to the er in drosophila. *Dev Cell*, **9**(5), 675–85.

Will, C. L. and Luhrmann, R. (2011). Spliceosome structure and function. *Cold Spring Harb Perspect Biol*, **3**(7).

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2007). Inferring noncoding rna families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, **3**(4), e65.

Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F., and Backofen, R. (2012). Locarna-p: accurate boundary prediction and improved detection of structural rnas. *RNA*, **18**(5), 900–14.

Williams, A. S. and Marzluff, W. F. (1995). The sequence of the stem and flanking sequences at the 3' end of histone mrna are critical determinants for the binding of the stem-loop binding protein. *Nucleic Acids Res*, **23**(4), 654–62.

Wilusz, C. J. and Wilusz, J. (2005). Eukaryotic lsm proteins: lessons from bacteria. *Nat Struct Mol Biol*, **12**(12), 1031–6.

Wu, X. and Sharp, P. A. (2013). Divergent transcription: a driving force for new gene origination? *Cell*, **155**(5), 990–6.

Xu, K., Bogert, B. A., Li, W., Su, K., Lee, A., and Gao, F. B. (2004). The fragile x-related gene affects the crawling behavior of drosophila larvae by regulating the mrna level of the deg/enac protein pickpocket1. *Curr Biol*, **14**(12), 1025–34.

Yang, L., Duff, M. O., Graveley, B. R., Carmichael, G. G., and Chen, L. L. (2011). Genomewide characterization of non-polyadenylated rnas. *Genome Biol*, **12**(2), R16.

Yang, X. C., Burch, B. D., Yan, Y., Marzluff, W. F., and Dominski, Z. (2009). Flash, a proapoptotic protein involved in activation of caspase-8, is essential for 3' end processing of histone pre-mrnas. *Mol Cell*, **36**(2), 267–78.

Yang, Z., Zhu, Q., Luo, K., and Zhou, Q. (2001). The 7sk small nuclear rna inhibits the cdk9/cyclin t1 kinase to control transcription. *Nature*, **414**(6861), 317–22.

Yong, J., Pellizzoni, L., and Dreyfuss, G. (2002). Sequence-specific interaction of u1 snrna with the smn complex. *EMBO J*, **21**(5), 1188–96.

Young, D. (1998). Package 'mixtools', http://cran.r-project.org/web/packages/mixtools/mixtools.pdf.

Zhang, A., Altuvia, S., Tiwari, A., Argaman, L., Hengge-Aronis, R., and Storz, G. (1998). The oxys regulatory rna represses rpos translation and binds the hfq (hf-i) protein. *EMBO J*, **17**(20), 6061–8.

Zhang, A., Wassarman, K. M., Ortega, J., Steven, A. C., and Storz, G. (2002). The sm-like hfq protein increases oxys rna interaction with target mrnas. *Mol Cell*, **9**(1), 11–22.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G., and Gottesman, S. (2003). Global analysis of small rna and mrna targets of hfq. *Mol Microbiol*, **50**(4), 1111–24.

Zhang, A., Schu, D. J., Tjaden, B. C., Storz, G., and Gottesman, S. (2013). Mutations in interaction surfaces differentially impact e. coli hfq association with small rnas and their mrna targets. *J Mol Biol*, **425**(19), 3678–97.

Zhang, H., Xing, L., Rossoll, W., Wichterle, H., Singer, R. H., and Bassell, G. J. (2006). Multiprotein complexes of the survival of motor neuron protein smn with gemins traffic to neuronal processes and growth cones of motor neurons. *J Neurosci*, **26**(33), 8622–32.

Zhang, L., Shimoji, M., Thomas, B., Moore, D. J., Yu, S. W., Marupudi, N. I., Torp, R., Torgner, I. A., Ottersen, O. P., Dawson, T. M., and Dawson, V. L. (2005). Mitochondrial localization of the parkinson's disease related protein dj-1: implications for pathogenesis. *Hum Mol Genet*, **14**(14), 2063–73.

Zhang, R., So, B. R., Li, P., Yong, J., Glisovic, T., Wan, L., and Dreyfuss, G. (2011). Structure of a key intermediate of the smn complex reveals gemin2's crucial function in snrnp assembly. *Cell*, **146**(3), 384–95.

Zhao, J., Ohsumi, T. K., Kung, J. T., Ogawa, Y., Grau, D. J., Sarma, K., Song, J. J., Kingston, R. E., Borowsky, M., and Lee, J. T. (2010). Genome-wide identification of polycomb-associated rnas by rip-seq. *Mol Cell*, **40**(6), 939–53.

Zhao, Z., Tavoosidana, G., Sjolinder, M., Gondor, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K. S., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., and Ohlsson, R. (2006). Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, **38**(11), 1341–7.

Zhou, L., Hang, J., Zhou, Y., Wan, R., Lu, G., Yin, P., Yan, C., and Shi, Y. (2014). Crystal structures of the lsm complex bound to the 3' end sequence of u6 small nuclear rna. *Nature*, **506**(7486), 116–20.