12-12-2016

Endogenous Small Interfering RNA: Insights into esiRNA biogenesis and their precursors

Andrew White Harrington
University of Missouri-St. Louis, awhwd2@mail.edu

Endogenous Small Interfering RNA: Insights into esiRNA biogenesis and their precursors

By

Andrew W. Harrington

M.S., Biology, University of Missouri-Saint Louis, 2011
B.S., Biopsychology and Cognitive Science, University of Michigan, 2004

A Dissertation submitted to the Graduate School at the University of Missouri-Saint Louis in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy in Biology
With an Emphasis in Cell and Molecular Biology

December 2016

Advisory Committee

Mindy M. Steiniger, Ph.D.
Chairperson

Bethany K. Zolman, Ph.D.

Wendy M. Olivas, Ph.D.

Michael Hughes, Ph.D.

ABSTRACT

Rarely in research is the path to an answer straightforward. Initial questions lead to more questions, many times doubling back to allow for greater insight into the original question. For example, discovery of interactions between previously unrelated pathways can lead to breakthroughs with regard to understanding of gene regulation. One such novel interaction and the subsequent discoveries this interaction spurred are discussed herein.

Transposons, or “Jumping Genes” are mobile genetic elements found throughout all three major domains of life. Transposons comprise 44% of the human genome and possess the ability to move within the genome. This ability makes them an important driver of evolution, but also requires that they be tightly regulated. In 2008, a number of papers were published outlining a new class of small RNA, endogenous small interfering RNA (esiRNA). These esiRNAs were derived from transposons and structured loci (hairpins.) EsiRNAs are produced from a dsRNA precursor in a Dicer-2 dependent manner. When a novel interaction between the 3’ end processing factor Symplekin and Dicer-2 was discovered, further investigation into this relationship, and the very nature of esiRNA precursors was warranted.

Herein, I uncover the mechanism by which certain classes of transposons create the dsRNA precursor necessary for esiRNA biogenesis and shed light onto their regulation. I further investigate the difference between retrotransposons or hairpin derived esiRNAs with regard to their physical characteristics and subcellular location.

Lastly, I investigate the role 3’ end processing machinery, such as Symplekin, plays in esiRNA biogenesis.

Ph.D. SUMMARY

PUBLICATIONS

Harrington, A. W., McKain, M. R., Michalski, D., Bauer, K., Kidd III, A. R., and Steiniger, M. (2016) Symplekin plays a role in Drosophila endo-siRNA biogenesis. In Preparation.

Harrington, A. W., Steiniger, M. (2016) Bioinformatic Analyses of Sense and Antisense Expression from Terminal Inverted Repeat Transposons in *Drosophila* Somatic cells. *Fly.* March 2016. 10(1):1-10

Russo, J.*, **Harrington, A. W**.*, and Steiniger, M. (2016) Antisense transcription of retrotransposons in *Drosophila*: The origin of endogenous small interfering RNA precursors. *Genetics.* January 2016. 202:107-21. (*co-first authors)

RESEARCH TALKS AND POSTERS

- Guest speaker for the Truman State University Tri-Beta Biological Society, 2016.
- Presenter at 2016, 2015, 2014 UMSL Biology Symposium.
- Poster Presentations at 2016, 2015, 2014 Annual Meeting of the RNA Society.
- Poster Presentation at 2014 Rust Belt RNA Meeting.

AWARDS AND FELLOWSHIPS

- 2016 University of Missouri Saint Louis Dissertation Fellowship.
- 2016 Raju Mehra Outstanding Graduate Student Award.
- 2016 RNA Society Travel Fellowship.
- 2015 Biology Graduate Student Association Travel Award.
- 2015 RNA Society Travel Fellowship.
- 2014 RNA Society Travel Fellowship.

ACKNOWLEDGMENTS

I have always felt that the scientific community is much like a family, and I cannot think of a situation that exemplifies this more than a Ph.D. program. Growing up surrounded by your science family of parents, aunts and uncles, brothers and sisters, and eventually your own children. There are many members of my science family I feel have supported, guided, and in many cases prodded me along this path. First and foremost, I want to thank and acknowledge my Science Mom, Dr. Mindy Steiniger. When I asked her if I could do my Ph.D. in her lab I can only describe the look I got in response as a mix of hopefully optimistic apprehension with just a touch of confusion. I am Mindy's first born Science Baby, and I knew going into a lab under the direction of a new, untenured professor would be very challenging. I knew it would be work, I knew it would be hard; I knew we would both have to grow together. What I didn't know was how rewarding it would be. To be part of a project that started as no more than a seed of an idea, and watch that seed grow into a tree that will be providing fruit for our lab for a very long time is an amazing feeling. Words cannot express how grateful I am for her guidance, support, and friendship.

Outside of Mindy, there are so many more members of my family that I need to thank and acknowledge as well. I would like to thank all my Aunts and Uncles that made up my committee. I would like to especially thank Dr. Zolman and Dr. Olivas. Both Bethany and Wendy were instrumental in guiding and supporting me during the less than easy parts of the journey. I would also like to thank Dr. Hughes for providing

insightful advice and of course all his help with our sequencing experiments. In addition to my committee, there are a few professors that I would like to especially thank, namely, Dr. Trey Kidd and Dr. Lon Chubiz.

I would also like to thank all my science brothers and sisters, those students past and present that fueled discussions, helped me trouble shoot experiments, or just loaned me some 10X SDS Run Buffer once in a while. Foremost, I would like to thank Dan Michalski for his support, scientific discussions, and friendship. I would also like to thank Joe Russo, Tony Fischer, Vanessica Jawahir, and Silpi Thota. I would also like to thank my younger brothers and sisters – those students I was lucky enough to work with and mentor over the years. I would especially like to thank my younger sister, Tabitha McCullers. Being able to guide you as you began this journey has taught me more about science and mentorship than I would have thought possible.

Lastly, I would be remiss if I didn't thank and acknowledge my non-science family as well. I want to thank my mother and father for always supporting me in whatever I wanted to do. I want to thank my sister for always being there and keeping an eye on mom and dad while I was away. I would like to thank Paul and Kelcye for keeping me sane and well fed for the first part of this journey. I want to thank Eric Juntenun for showing me that you can leave a well paying job, get massively in debt, get a Ph.D., get married and live happily ever after. I hope to follow in your footsteps. I want to thank Karl Barton for telling me when I was 16 years old that all the good ideas have already been had, laying the angry foundation that has partly fueled me through this process. Lastly, I want to thank my mostest, Karla Susanne Terhark. I have no doubt that without

your love, support, and friendship over the past two years that I would not be writing these acknowledgments in the first place. With you by my side, I can do anything.

DEDICATION

I would like to dedicate this Ph.D. dissertation to my grandfather,

Dr. Frederic Randolph White.

TABLE OF CONTENTS

FIGURES AND TABLES

CHAPTER I: INTRODUCTION

Introduction

Regulation of gene expression is critical to proper cellular function and viability of an organism. This regulation can occur in myriad ways from transcriptional repression and post-transcriptional modification, to targeted degradation of the mRNA message. Induction of heterochromatin formation by small interfering RNA can control whether a message is made in the first place. Post transcriptional modification such as 3' end processing influences export and stability of a message once it has been transcribed. Finally, the canonical RNAi pathway is able to selectively target a message for degradation prior to translation. These mechanisms are by no means the only way gene expression is regulated, but for the purpose of the following discussion, they will be the main areas of focus. It is the goal of this work to provide greater understanding into the post-transcriptional regulation of transposable elements via the endogenous small interfering RNA (esiRNA) pathway.

Transposable Elements.

Since their discovery in 1950 by Barbara McClintock, transposable elements (TEs) have been the topic of extensive study (McClintock, 1950). They make up approximately 50% of the maize genome and 30% of the *Drosophila* genome. After completion of the human genome project, it was found that approximately 44% of the human genome is estimated to be comprised of TEs, intensifying the interest in their study (Goodier, 2016). TEs, or "Jumping Genes" are mobile genetic elements that possess the ability to move within the genome. Their unchecked proliferation would have drastic consequences for the cell. Because of this, they are not only potent drivers of evolution, but also some of

the most tightly regulated examples of gene expression. TEs are classified broadly into two classes: DNA transposons and retrotransposons.

DNA transposons comprise approximately 3% of the human genome. They move throughout the genome by means of a "cut and paste" mechanism. The majority of DNA transposons have short terminal inverted repeats (TIRs). TIR transposons encode their own transposes, allowing for the excision and movement of the transposon itself (Kaminker et al., 2002). Analysis of the TIR transposon *pogo and* 1360 (ProtoP) is discussed within this work. The *pogo* TIR is a member of the Tc1/*mariner* family of transposons while 1360 is thought to be derived from ancient P-like elements (Kaminker et al., 2002). There are currently no known active DNA transposons in mammals, however *Drosophila* S2 cells do transcribe their TIRs, allowing for the study of their regulation. Our data suggest that despite the transcription of TIR transposons, their movement is restricted by the lack of a functional transposase.

In contrast to TIR transposons, retrotransposons utilize a "copy and paste" mechanism for their movement. In short, the retrotransposon encodes a reverse transcriptase, allowing for the retrotransposon mRNA message to be reverse transcribed back into DNA. This DNA element is then reinserted into the genome. Using this RNA intermediate allows for the genomic amplification of the retrotransposon and can explain why retrotransposons are so much more abundant than TIR transposons. Retrotransposons are regulated in the germline by the piwi-interacting small RNA pathway (piRNA). However, in somatic cells, retrotransposons are regulated by the esiRNA pathway. The different small interfering RNA pathways will be further discussed

later in this work. Retrotransposons are further subdivided into two categories: those that contain Long Terminal Repeats (LTR retrotransposons) and those that do not (non-LTR retrotransposons.)

LTR retrotransposons are so named based on the presence of long terminal repeats on either end of TE. There are many families of LTR retrotransposons. In this work, I have chosen to study LTR transposons belonging to the gypsy family: specifically Dm297, Blood, and Mdg1. These three LTR retrotransposons were chosen based on their relatively high expression level and low ratio of sense to antisense transcription. Both LTR and Non-LTR retrotransposons contain a reverse transcriptase, however the LTR retrotransposon usually contains a pol and a gag gene, whereas the non-LTR retrotransposon may not.

Non-LTR retrotransposons are similar to their LTR counterparts in that they encode a reverse transcriptase, but do not contain the characteristic long terminal repeats at the 3'and 5' ends of the gene. Two non-LTR transposons, Jockey and Juan, are focused on within this study as they were among the few transposons that showed both sense and antisense transcription, as well as relatively high expression. Most non-LTR retrotransposons are classified as LINES (Long INterspersed Elements) and SINES (Short INterspersed Elements). In humans, the LINE-1 retrotransposons are an active autonomous mobile element and are believed to be regulated by small interfering RNAs (Yang & Kazazian, 2006). The LINE-1 retrotransposon is closely related to *Drosophila* non-LTR retrotransposon Jockey and Juan, making the study of these two retrotransposons even more intriguing (Mizrokhi *et al.* 1988; Speek 2001). LTR

retrotransposons are thought to be as old as the some of the first unicellular organisms (Malik, Burke, & Eickbush, 1999). It is thought that non-LTR retrotransposons predate LTR transposons (Cordaux & Batzer, 2009). Both classes of retrotransposons are similar to modern day retroviruses. In fact, it is believed that the gypsy family of LTR retrotransposons acquired an envelop gene which allowed for the retrotransposon to escape the cell and become what we consider to be a retrovirus (Malik, Henikoff, & Eickbush, 2000).

Small silencing RNA pathways

As previously discussed, regulation of gene expression is essential to proper cellular function. Small silencing RNAs are one method the cell uses to regulate both heterochromatin formation and post-transcriptional gene silencing (Czech et al., 2008; Fagegaltier et al., 2009). In *Drosophila,* these small silencing RNAs are divided into three categories: Piwi-RNA (piRNA), micro-RNA (miRNA) and small interfering RNA (siRNA) (Figure 1.1 (Ghildiyal & Zamore, 2009)). The last category, siRNA, can be further subdivided into endogenously derived siRNA (esiRNA) and exogenously derived siRNA (exo-siRNA). These categories are defined by the biosynthetic pathway of the mature small RNA, specifically which member(s) of the Argonaute family of proteins they interact with (Czech et al., 2008). The Argonaute family is subdivided into two groups. The first group is comprised of Argonaute 1 and Argonaute 2 (Ago1 and Ago2) and associate with miRNA and siRNA, respectively. The second group is comprised of the Piwi family of proteins and interacts with piRNA.

FIGURE 1.1 Schematic of Small RNA Silencing Pathway

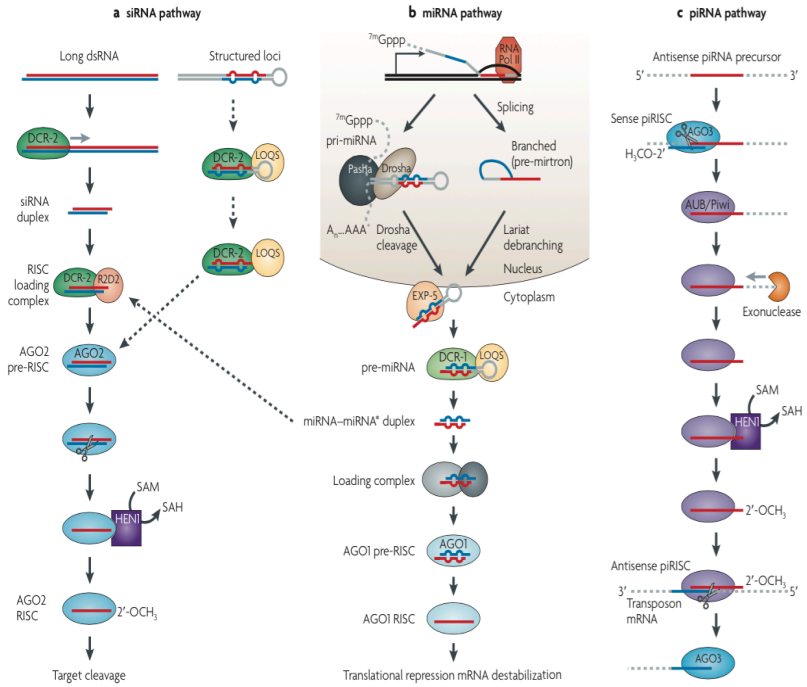


Figure 1.1 Small RNA Pathways The three small RNA silencing pathways in flies are the small interfering RNA (siRNA), microRNA (miRNA) and Piwi-interacting RNA (piRNA) pathways. These pathways differ in their substrates, biogenesis, effector proteins and modes of target regulation. **a** | dsRNA precursors are processed by Dicer-2 (DCR-2) to generate siRNA duplexes containing guide and passenger strands. DCR-2 and the dsRNA-binding protein R2D2 (which together form the RISC-loading complex, RLC) load the duplex into Argonaute2 (AGO2). A subset of endogenous siRNAs (endo-siRNAs) exhibits dependence on dsRNA-binding protein Loquacious (LOQS), rather than on R2D2. The passenger strand is later destroyed and the guide strand directs AGO2 to the target RNA. **b** | miRNAs are encoded in the genome and are transcribed to yield a primary miRNA (pri-miRNA) transcript, which is cleaved by Drosha to yield a short precursor miRNA (pre-miRNA). Alternatively, miRNAs can be present in introns (termed mirtrons) that are liberated following splicing to yield authentic pre-miRNAs. pre-miRNAs are exported from the nucleus to the cytoplasm, where they are further processed by DCR-1 to generate a duplex containing two strands, miRNA and miRNA*. Once loaded into AGO1, the miRNA strand guides translational repression of target RNAs. **c** | piRNAs are thought to derive from ssRNA precursors and are made without a dicing step. piRNAs are mostly antisense, but a small fraction is in the sense orientation. Antisense piRNAs are preferentially loaded into Piwi or Aubergine (AUB), whereas sense piRNAs associate with AGO3. The methyltransferase HEN1 adds the 2'-*O*-methyl modification at the 3' end. Piwi and AUB collaborate with AGO3 to mediate an interdependent amplification cycle that generates additional piRNAs, preserving the bias towards antisense. The antisense piRNAs probably direct cleavage of transposon mRNA or chromatin modification at transposon loci. SAH, *S*-adenosyl homocysteine; SAM, *S*-adenosyl methionine.

PiRNA exist in the germ line and is necessary for proper development and suppression of transposable elements (Ghildiyal & Zamore, 2009). In contrast to siRNA and miRNA, piRNA precursors are not cleaved by a Dicer protein. *Drosophila* possesses two Dicer proteins, Dicer-1 (Dcr-1) and Dicer-2 (Dcr-2). Though structurally similar, these proteins have distinct roles within the cell. Dcr-2 is required for production of siRNA, while Dcr-1 is required for miRNA production. MiRNA are endogenously produced small RNAs found throughout a number of species, including humans and *Drosophila.* MiRNAs inhibit protein translation by binding to complementary sequences of mRNA and targeting them for degradation. MiRNA are formed from primary miRNA transcripts (pri-miRNAs) that are processed in the nucleus by the proteins Drosha and Pasha. This processing results in pre-miRNA, a 70-nucleotide sequence containing a stem loop structure, 2-nucleotide 3' overhang, and the mature miRNA sequence. These pre-miRNA are exported to the cytoplasm by the nuclear export protein Exportin 5, where they are further processed by Dcr-1 and loaded into the RNA induced silencing complex (RISC) by Ago1 (Lucas & Raikhel, 2013).

The third class of small silencing RNA, siRNA, is divided into two categories based on the origin of the dsRNA precursor. The exo-siRNA pathway has historically been viewed as a means of defense against viral RNA (Sabin et al., 2013). Because of this, the pathway can be exploited in Dmel-2 cells to induce transient protein knockdown. Exo-siRNA precursors are exogenous in nature and processed by Dcr-2 in the cytoplasm. The cleaved Dcr-2/RNA complex associates with R2D2 and is loaded into Ago2 forming the RISC. This complex binds mRNA complementary to the dsRNA precursor and targets it

for degradation (Tomari & Zamore, 2005). In contrast, esiRNAs are derived from endogenous sources, most often from transposons and structured loci (hairpins). Unlike miRNA precursors, the secondary structure and processing of esiRNA precursors is poorly understood (Cenik et al., 2011). EsiRNAs have been found in a number of species including *C. elegans, Drosophila,* and in mouse oocytes (Duchaine et al., 2006; Tam et al., 2008). Recently, the LINE-1 retrotransposon in humans appears to be regulated by "natural siRNA" that may be similar to canonical esiRNA (Yang & Kazazian, 2006). Their main role in *Drosophila* is to silence transposable elements in somatic cells, such as the ones discussed in the first part of this introduction (Czech et al., 2008). Like their exogenous counterparts, they are cleaved by Dcr-2 and the mature esiRNA associates with Ago2. However, unlike exo-siRNAs, they are R2D2 independent and require a specific isoform of Loquacious, Loquacious PD (Loqs-PD), for correct cleavage (Zhou et al., 2009).

Besides the source of the dsRNA precursor, one of the most prominent differences between the exo-siRNA and esiRNA pathways is their choice of accessory proteins. Biogenesis of exo-siRNAs requires R2D2, while the esiRNA pathway requires Loqs-PD (Hartig & Förstemann, 2011; X. Liu, Jiang, Kalidas, Smith, & Liu, 2006) (figure 1.1). Recent studies have indicated that Loqs-PD and R2D2 may be antagonistic to one another as they appear to bind the same site on Dcr-2 (Hartig & Förstemann, 2011). The miRNA pathway requires a different isoform of the same protein, Loquacious PB (Loqs-PB). Loqs-PB interacts with Dcr-1 in the cytoplasm to facilitate processing of miRNA, suggesting cytoplasmic localization. Loqs-PD has also been shown to be cytoplasmic (K.

Miyoshi, Miyoshi, Hartig, Siomi, & Siomi, 2010). The disparity between exo-siRNA and esiRNA protein preference highlights an even greater difference between the two pathways. While exo-siRNA cleavage and processing is known to occur in the cytoplasm, the same cannot be claimed for esiRNAs. Investigation into the localization of esiRNA precursor molecules is a main focus of this work.

Interestingly, many of the proteins involved in siRNA biogenesis have been found in the nucleus (Tomari & Zamore, 2005). Ago2, whose location of action had historically been thought of as cytoplasmic due to its role in RNAi, has been implicated in pre-mRNA splicing and transcriptional repression, indicating that it has a nuclear localization and activity (Taliaferro et al., 2013). Additionally, ChIP experiments have indicated that both Dcr-2 and Ago2 associate with chromatin and transcription machinery such as RNA Polymerase II (RNA Pol II) (Cernilogar et al., 2011). Since the raw materials (precursor RNA) and necessary siRNA machinery are at least transiently nuclear, it is possible that processing of esiRNA precursors is a nuclear event.

Multiple groups reported identification of esiRNAs concurrently in 2008 (Czech et al., 2008; Kawamura et al., 2008; Okamura, Balla, Martin, Liu, & Lai, 2008). Many esiRNAs map back to structured loci, the most well studied being the Esi1 and Esi2 locus. Esi1 (CG18854) is the precursor message for the esi1.2 esiRNA and Esi2 (AY119020) is the precursor for Esi2.1. These two individual esiRNAs are the most abundant of hairpin derived esiRNAs (Czech et al., 2008). The majority of esiRNAs however map back to retrotransposons (Ghildiyal et al., 2008). As such, investigation into the biogenesis of both types of precursors comprises the majority of this work.

3’ end processing and the Core Cleavage Complex

In addition to the mechanisms of protein regulation mentioned above, processing of RNA, such as addition of a 5’ cap or 3’ cleavage and polyadenylation can have drastic effects on gene expression. Many RNA Pol II transcripts are polyadenylated, a modification required for stability and export (O'Hare, 1995). First, pre-mRNAs are cleaved between a conserved AAUAAA sequence and a downstream G/U rich element. Following cleavage, poly(A) polymerase (PAP) adds a number of adenosine monophosphate nucleotides to the 3’ end, creating a poly(A) tail. The cleavage prior to polyadenylation is accomplished by the Core Cleavage Complex (Sullivan, Steiniger, & Marzluff, 2009). The Core Cleavage Complex (CCC) is composed of a trio of proteins: Symplekin, Cpsf73, and Cpsf100. CPSF100 forms a heterodimer with CPSF73, with Symplekin acting as a scaffold (Takagaki & Manley, 2000). These proteins interact to form a tight nuclear complex that regulates 3’ end processing in a co-transcriptional manner (Sullivan et al., 2009). Additionally, the core cleavage complex has been found to interact with the CTD of RNA Pol II (Sullivan et al., 2009). Recent experiments by our lab indicate there is a biological connection between Symplekin and Dcr-2. This is exciting as Symplekin has not previously been associated neither with the siRNA pathway nor its machinery. The interplay between the siRNA machinery and 3’ end processing factors such as Symplekin have yet to be investigated and comprise a large part of this work.

Preliminary Data

Previous research performed by Dr. Steiniger indicates a biological interaction between the 3’ end processing factor Symplekin and the siRNA pathway associated

protein Dcr-2. Symplekin was immunoprecipitated and the associated proteins were identified by mass spectroscopy (Figure 1.2). Among the identified proteins were known binding partners of Symplekin, such as CPSF73 and CPSF100. Unexpectedly, Dcr-2 was also identified in this screen. Given the role of Dcr-2 in siRNA biogenesis, a northern blot to the mature esiRNA esi2.1 was performed in Dcr-2 and Symplekin knockdowns (Figure 1.3). This experiment indicates that depletion of Symplekin results in a decreased level of esi2.1.

Experiments by Josh Daughtry, a previous member of the Steiniger lab, identified the precursor to esi2.1, AY119020, as a Pol II polyadenylated transcript (Figure 1.4). Given the role of Symplekin in cleavage of polyadenylated transcripts and the effect of Symplekin RNAi depletion on esi2.1 levels, it is plausible that the processing of esiRNAs is occurring in the nucleus, potentially in a co-transcriptional fashion. These experiments suggest a previously uncharacterized role for Symplekin in the production of esiRNAs and were the main impetus in pursuing this project.

Figure 1.2 Mass Spectroscopy of Symplekin associated proteins

Sample	Name	FlybaseID	MW (Da)	# of Peptides	Function
a	CPSF160	FBgn0024698	164.7	33	3' end processing
b	Dicer-2	FBgn0034246	197.8	23	siRNA
c	Symp	FBgn0037371	132.1	47	3' end processing
d	CPSF100	FBgn0027873	85.4	31	3' end processing
e	WDR33	FBgn0046222	90.5	29	3' end processing
f	CPSF6	FBgn0035872	71.1	29	3' end processing/ poly(A) site selection
g	CstF77	FBgn0003559	84.5	32	3' end processing
h	CPSF6	FBgn0035872	71.1	22	3' end processing/ poly(A) site selection
i	CPSF73	FBgn0261065	76.8	20	3' end processing
	Hsc70-4	FBgn0266599	71.1	11	RISC loading

Figure 1.2 Symplekin was immunoprecipitated and the associated proteins were identified by mass spectroscopy. The table lists a selection of the proteins identified including 3' end processing factors as well as Dicer-2.

Figure 1.3 Northern blot of Esi2.1 in Symplekin knockdown

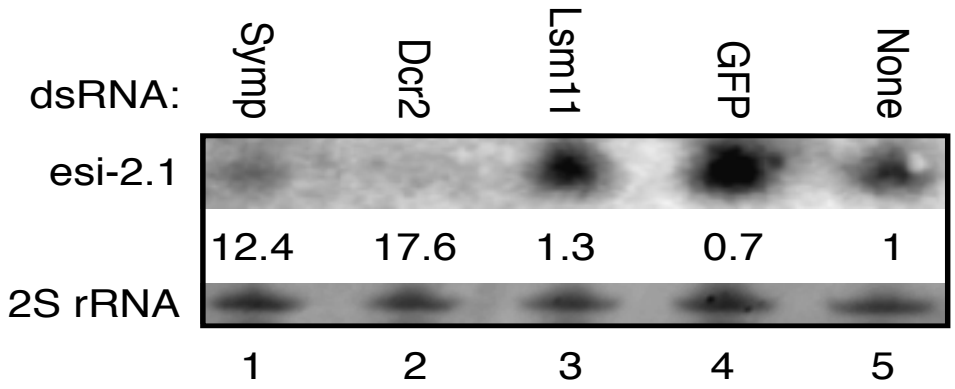


Figure 1.3 Northern blot showing that after knockdown of Symplekin via RNAi, the level of esi-2.1 decreases.

Figure 1.4 Hairpin derived esiRNA precursors are polyadenylated transcripts

AY	Cq
Total	21.17
PolyA(+)	17.2
PolyA(-)	21.04

GAP	Cq
Total	20.69
PolyA(+)	16.35
PolyA(-)	20.54

54	Cq
Total	25.02
PolyA(+)	21.23
PolyA(-)	25.25

18S	Cq
Total	17.66
PolyA(+)	19.3
PolyA(-)	16.08

Actin	Cq
Total	17.6
PolyA(+)	13.91
PolyA(-)	17.04

Figure 1.5 Cq values for total, PolyA+, and Poly A- fractions for esi2.1 and esi1.2 precursors.

Overall Significance and Conclusions

Research presented herein provides new and exciting insights into the world of endogenous small RNA biogenesis and the regulation of transposable elements. In the first project, I investigate the mechanism by which retrotransposons form the double stranded RNA substrate necessary for them to be cleaved by Dcr-2. I show that the majority of retrotransposon families are generated from convergent transcription of sense and antisense transcripts. I further show that these precursors are processed by Dcr-2 into endogenous small interfering RNAs. Finally, I show canonical transcriptional start sites that not only drive the transcription of the retrotransposon itself, but have the potential to form fusion transcripts and drive downstream gene expression. This analysis was further expanded in a second project in which TIR DNA transposons were investigated in a similar fashion to retrotransposons. Interestingly, bioinformatic analysis revealed that unlike retrotransposons, TIR transposons do not undergo convergent transcription, and very few small RNAs map back to these loci. This suggests that while the retrotransposons are regulated by Dcr-2 dependent generation of small RNAs, TIR transposon movement is regulated by another mechanism, most likely the lack of a functional transposase. As transposable elements make up approximately 44% of the human genome, studying the mechanism of their repression in other species is critical to our understanding of their regulation as a whole.

The second project presented in this dissertation further examines the esiRNA pathway and its relationship to 3' end processing factors Symplekin, Cpsf73, Cpsf100. Additionally, differences between retrotransposon and hairpin precursors are also investigated. In this work, I further establish the association between Dcr-2 and

Symplekin to be a bona fide relationship via a direct interaction with the N-terminus of Symplekin. Subsequent bioinformatics analysis revealed indirect effects of Symplekin on the processing of both retrotransposon and hairpin precursor molecules. Most interestingly, differences in the esiRNAs were observed in the knockdown conditions depending on whether the small RNAs were derived from retrotransposons or from hairpins. Differences in 3' or 5' base preference, size, and subcellular location of the small RNAs were also observed. Interestingly, knockdown of 3' end processing factors had differing effects on levels of retrotransposon or hairpin precursors and small RNAs, decreasing overall levels of retrotransposon precursors while increasing levels of hairpin precursors.

The work discussed in this dissertation furthers the general understanding of not only retrotransposon regulation but also of small RNA biogenesis as a whole. As discussed, regulation of transposable elements is crucial to proper cellular function and development. Though there have not been any reports of bona fide esiRNAs in humans, there have been reports of esiRNAs in mouse oocytes (Tam et al., 2008). There have also been reports of natural double stranded RNAs (ndsRNA) in human cells that could potentially have regulatory functions (Portal, Pavet, Erb, & Gronemeyer, 2015). Additionally, though most of the LINE-1 retrotransposons in humans are inactive, there have been recent reports of certain LINE-1 retrotransposons being regulated by natural siRNAs in human culture cells (Yang & Kazazian, 2006). Furthering our understanding of how transposable elements are regulated by esiRNAs in *Drosophila* may have important

implications for understanding small RNA biogenesis and regulation of retrotransposons in mammals.

References

Cenik, E. S., Fukunaga, R., Lu, G., Dutcher, R., Wang, Y., Tanaka Hall, T. M., & Zamore, P. D. (2011). Phosphate and R2D2 restrict the substrate specificity of Dicer-2, an ATP-driven ribonuclease. *Molecular Cell*, *42*(2), 172–184. http://doi.org/10.1016/j.molcel.2011.03.002

Cernilogar, F. M., Onorati, M. C., Kothe, G. O., Burroughs, A. M., Parsi, K. M., Breiling, A., et al. (2011). Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. *Nature*, *480*(7377), 391–395. http://doi.org/10.1038/nature10492

Cordaux, R., & Batzer, M. A. (2009). The impact of retrotransposons on human genome evolution. *Nature Reviews. Genetics*, *10*(10), 691–703. http://doi.org/10.1038/nrg2640

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., et al. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature*, *453*(7196), 798–802. http://doi.org/10.1038/nature07007

Duchaine, T. F., Wohlschlegel, J. A., Kennedy, S., Bei, Y., Conte, D., Pang, K., et al. (2006). Functional proteomics reveals the biochemical niche of C. elegans DCR-1 in multiple small-RNA-mediated pathways. *Cell*, *124*(2), 343–354. http://doi.org/10.1016/j.cell.2005.11.036

Fagegaltier, D., Bougé, A.-L., Berry, B., Poisot, E., Sismeiro, O., Coppée, J.-Y., et al. (2009). The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21258–21263. http://doi.org/10.1073/pnas.0809208105

Ghildiyal, M., & Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews. Genetics*, *10*(2), 94–108. http://doi.org/10.1038/nrg2504

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., et al. (2008). Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science (New York, N.Y.)*, *320*(5879), 1077–1081. http://doi.org/10.1126/science.1157396

Hartig, J. V., & Förstemann, K. (2011). Loqs-PD and R2D2 define independent pathways for RISC generation in Drosophila. *Nucleic Acids Research*, *39*(9), 3836–3851. http://doi.org/10.1093/nar/gkq1324

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., et al. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biology*, *3*(12), RESEARCH0084. http://doi.org/10.1186/gb-2002-3-12-research0084

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., et al. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, *453*(7196), 793–797. http://doi.org/10.1038/nature06938

Liu, X., Jiang, F., Kalidas, S., Smith, D., & Liu, Q. (2006). Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes. *RNA (New York, N.Y.)*, *12*(8), 1514–1520. http://doi.org/10.1261/rna.101606

Lucas, K., & Raikhel, A. S. (2013). Insect microRNAs: biogenesis, expression profiling and biological functions. *Insect Biochemistry and Molecular Biology*, *43*(1), 24–38. http://doi.org/10.1016/j.ibmb.2012.10.009

Malik, H. S., Burke, W. D., & Eickbush, T. H. (1999). The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, *16*(6), 793–805.

Malik, H. S., Henikoff, S., & Eickbush, T. H. (2000). Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research*, *10*(9), 1307–1318.

McCLINTOCK, B. (1950). The origin and behavior of mutable loci in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344–355.

Miyoshi, K., Miyoshi, T., Hartig, J. V., Siomi, H., & Siomi, M. C. (2010). Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. *RNA (New York, N.Y.)*, *16*(3), 506–515. http://doi.org/10.1261/rna.1952110

O'Hare, K. (1995). mRNA 3' ends in focus. *Trends in Genetics : TIG*, *11*(7), 255–257.

Okamura, K., Balla, S., Martin, R., Liu, N., & Lai, E. C. (2008). Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nature Structural & Molecular Biology*, *15*(9), 998–998. http://doi.org/10.1038/nsmb0908-998c

Portal, M. M., Pavet, V., Erb, C., & Gronemeyer, H. (2015). Human cells contain natural double-stranded RNAs with potential regulatory functions. *Nature Structural & Molecular Biology*, *22*(1), 89–97. http://doi.org/10.1038/nsmb.2934

Sabin, L. R., Zheng, Q., Thekkat, P., Yang, J., Hannon, G. J., Gregory, B. D., et al. (2013). Dicer-2 processes diverse viral RNA species. *PloS One*, *8*(2), e55458. http://doi.org/10.1371/journal.pone.0055458

Sullivan, K. D., Steiniger, M., & Marzluff, W. F. (2009). A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Molecular Cell*, *34*(3), 322–332. http://doi.org/10.1016/j.molcel.2009.04.024

Takagaki, Y., & Manley, J. L. (2000). Complex protein interactions within the human polyadenylation machinery identify a novel component. *Molecular and Cellular Biology*, *20*(5), 1515–1525.

Taliaferro, J. M., Aspden, J. L., Bradley, T., Marwha, D., Blanchette, M., & Rio, D. C. (2013). Two new and distinct roles for Drosophila Argonaute-2 in the nucleus: alternative pre-mRNA splicing and transcriptional repression. *Genes & Development*, *27*(4), 378–389. http://doi.org/10.1101/gad.210708.112

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, *453*(7194), 534–538. http://doi.org/10.1038/nature06904

Tomari, Y., & Zamore, P. D. (2005). Perspective: machines for RNAi., *19*(5), 517–529. http://doi.org/10.1101/gad.1284105

Yang, N., & Kazazian, H. H. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature Structural & Molecular Biology*, *13*(9), 763–771. http://doi.org/10.1038/nsmb1141

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J. A., Perrimon, N., & Hannon, G. J. (2009). Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. *RNA (New York, N.Y.)*, *15*(10), 1886–1895.

http://doi.org/10.1261/rna.1611309

CHAPTER 2: INSIGHTS INTO ESIRNA PRECURSORS

ANTISENSE TRANSCRIPTION OF RETROTRANSPOSONS IN *DROSOPHILA:* THE ORIGIN OF ENDOGENOUS SMALL INTERFERING RNA PRECURSORS

CONTRIBUTIONS

The work described below has been published in the *Genetics*, January 2016 where I share Co-First Authorship with Dr. Joseph Russo. My contributions include all sequencing experiments and northern blots. Dr. Russo and myself both contributed to strand specific RT-qPCR and primer design. Dr. Russo performed the PolyA+/- fractionation and bioinformatics analysis. Dr. Steiniger, Dr. Russo, and myself prepared the manuscript.

SUMMARY

Movement of transposons causes insertions, deletions, and chromosomal rearrangements potentially leading to premature lethality in *Drosophila melanogaster*. To repress these elements and combat genomic instability, eukaryotes have evolved several small RNA-mediated defense mechanisms. Specifically, in *Drosophila* somatic cells, endogenous small interfering (esi)RNAs suppress retrotransposon mobility. EsiRNAs are produced by Dicer-2 processing of double-stranded RNA precursors, yet the origins of these precursors are unknown. We show that most transposon families are transcribed in both the sense (S) and antisense (AS) direction in Dmel-2 cells. LTR retrotransposons Dm297, mdg1, and blood, and non-LTR retrotransposons juan and jockey transcripts, are generated from intraelement transcription start sites with canonical RNA polymerase II promoters. We also determined that retrotransposon

antisense transcripts are less polyadenylated than sense. RNA-seq and small RNA-seq revealed that Dicer-2 RNA interference (RNAi) depletion causes a decrease in the number of esiRNAs mapping to retrotransposons and an increase in expression of both S and AS retrotransposon transcripts. These data support a model in which double-stranded RNA precursors are derived from convergent transcription and processed by Dicer-2 into esiRNAs that silence both sense and antisense retrotransposon transcripts. Reduction of sense retrotransposon transcripts potentially lowers element-specific protein levels to prevent transposition. This mechanism preserves genomic integrity and is especially important for *Drosophila* fitness because mobile genetic elements are highly active.

INTRODUCTION

Mobile genetic elements are one source of genetic alterations that drive evolution, but can also lead to catastrophic genomic instability. Thus, maintaining an appropriate balance between the potential harm and benefit of transposons (Tns) is vital. If active Tns are not adequately controlled by their hosts, mutations produced by their movement can be detrimental (Lee and Marx 2013). Specifically in *Drosophila*, genetic rearrangements that cause hybrid digenesis syndrome (Kidwell *et al.* 1977; Picard *et al.* 1978) are linked to transposon movement (Bingham *et al.* 1982; Rubin *et al.* 1982).

Since the discovery of Tns by Barbara McClintock more than 60 years ago (McClintock 1950), researchers have elucidated key mechanisms describing how Tns incorporate into genomes and how hosts combat these potentially toxic genomic

perturbations. However, many aspects of Tn biology remain elusive. While ~44% of the human genome is composed of Tns (Cordaux and Batzer 2009), there is little diversity in active transposons (Mills *et al.* 2007); only autonomous LINE-1 and nonautonomous Alu and SVA retrotransposons are currently mobile (Brouha *et al.* 2003; Cordaux and Batzer 2009; Deininger 2011). While the *Drosophila* genome is only ~22% transposons, many (~30%) of these elements are full length and thought to be active (Kaminker *et al.* 2002; Lerat *et al.* 2003; Kofler *et al.* 2015). Having active transposons from all three major classes of mobile elements to investigate offers a unique opportunity to understand silencing mechanisms in eukaryotic organisms.

Tns are defined by their approach to mobility. Terminal inverted repeat (TIR) Tns encode a Transposase that binds Tn inverted repeats (in most cases), creates double-strand breaks at the ends of the Tn, and integrates the Tn into a new genomic location. This mechanism can create genomic rearrangements such as insertions, deletions, and inversions. Unlike TIR Tns, retrotransposons (retroTns) include an RNA intermediate in their movement mechanism and therefore encode a reverse transcriptase (RT). RetroTns are divided into long terminal repeat (LTR) and non-LTR retroTns. LTR retroTns are similar to retroviruses and contain several hundred nucleotide terminal repeats at both the 5' and 3' ends (Figure 2.1A). While some *Drosophila* LTR retroTns have *gag* and *env* genes homologous to retroviruses, others have more divergent ORFs that function in retroTn mobility (Figure 2.1). Non-LTR retroTns lack these terminal repeats and sequences homologous to the *env* gene (Figure 2.2A), but have conserved RTs (Figure 2.2). Both LTR and non-LTR retroTns often have an internal promoter located in the 5'

untranslated region (UTR) and a 3′ UTR containing a polyadenylation signal (Gogvadze and Buzdin 2009 and this work). The initial transposition step for all retroTns is RNA polymerase II (RNAPII)-dependent transcription of the entire element followed by translation of each independent ORF in different reading frames from this single, polygenic messenger RNA (mRNA).

Eukaryotic cells have evolved several noncoding RNA-mediated mechanisms to control further genomic spread of retroTns. In humans, mobility of the LINE-1 (L1) retroTn is regulated by both canonical RNA interference (RNAi) (Yang and Kazazian 2006) and endogenous small interfering (esi)RNA-mediated chromatin modifications (Chen *et al.* 2012). Similarly, in *Drosophila*, two distinct RNAi-like processes for silencing Tns have been elucidated. In the germline, the Piwi-interacting RNA (piRNA) pathway generates small RNAs that suppress Tns by inducing heterochromatin formation (Vagin *et al.* 2006; Aravin *et al.* 2007; Brennecke *et al.* 2007; Sentmanat and Elgin 2012; Le Thomas *et al.* 2013). In somatic cells, esiRNAs silence retroTns via a Dicer 2 (Dcr-2)/Argonaute 2 (Ago2)-dependent mechanism (Chung *et al.* 2008; Czech *et al.* 2008; Ghildiyal *et al.* 2008; Saito and Siomi 2010; Xie *et al.* 2013). Global analysis of small RNA libraries generated from embryo-derived *Drosophila* somatic cells (S2) (Schneider 1972) showed that 86% of esiRNAs mapped to Tns; esiRNAs mapping to LTR retroTns were highly enriched (Ghildiyal *et al.* 2008). Dcr-2 is required for generation of esiRNAs (Czech *et al.* 2008; Okamura *et al.* 2008a,b) and retroTn expression increases following RNAi depletion of Dcr-2 (Ghildiyal *et al.* 2008; Marques *et al.* 2010).

The production of esiRNAs by Dcr-2 requires a double-stranded RNA (dsRNA)

precursor (Tomari *et al.* 2007; Ghildiyal *et al.* 2008; Marques *et al.* 2010). While dsRNAs generated by hybridization of natural antisense transcripts and their sense counterparts are substrates for Dcr-2 in *Drosophila* (Czech *et al.* 2008; Okamura *et al.* 2008a), retroTn dsRNA precursors have not been systematically investigated. As *Drosophila* does not encode an RNA-dependent RNA polymerase to generate a complementary strand, the origin of the antisense (AS) transcript necessary to form the dsRNA retroTn precursor is unknown. Here, we provide evidence that both non-LTR and LTR retroTns produce sense (S) and AS transcripts from intraelement transcription start sites (tss) with canonical *Drosophila* promoters. We then use a novel polyA+/− fractionation followed by strand-specific RT-qPCR technique to show that most S and AS retroTn transcripts are not enriched for polyadenylation. Finally, increases in AS retroTn transcript levels in Dmel-2 cells RNAi depleted of Dcr-2 indicate that AS and S transcripts are substrates for Dcr-2.

RESULTS

To investigate Tn AS transcription, we performed strand-specific high throughput sequencing (HTS) of rRNA-depleted RNA (>200 nt) from control Dmel-2 cells (LacZ). These libraries were prepared in triplicate and sequenced on an Illumina HiSeq, resulting in an average read depth of 101.5× and >98% of reads mapping to the *Drosophila* genome (Table 2.1). Surprisingly, 41.9% of the reads mapped nonuniquely (Table 2.1), indicating that a large percentage of transcripts are derived from non-rRNA repetitive sequences.

To compare abundance of S and AS Tn transcripts, we visualized nonunique RNA-seq reads using the UCSC genome browser. Because Tns are highly conserved and

multicopy, RNA-seq reads corresponding to S and AS Tn transcripts map to more than one genomic location. Therefore, the normalized reads per million value identified for each Tn generally represents total cellular S or AS transcription of all copies of that element. Only Tns having more than three full-length annotated elements in the *Drosophila* genome were investigated.

Table 2.1: HTS Mapping Statistics

Sample	total # reads	% mapping	read depth	% unique	*p*-value	% non-unique	*p*-value
RNA-seq							
LacZ1	30020327	98.5	99.7	57.0	6.949E-05	41.500	9.582E-05
LacZ2	35561677	98.4	118.1	56.3		42.100	
LacZ3	26072570	98.4	86.6	56.4		42.000	
Dcr2-1	22668363	98.2	75.3	51.8		46.500	
Dcr2-2	23826043	98.0	79.2	51.3		46.700	
Dcr2-3	24236601	98.0	80.5	51.1		46.900	
smRNA-seq							
LacZ1	341316	99.1		64.5	1.209E-03	34.600	1.544E-03
LacZ2	288286	99.1		63.4		35.700	
LacZ3	285543	99.2		66.1		33.000	
Dcr2-1	207361	99.7		77.6		22.100	
Dcr2-2	272897	99.7		76.5		23.1	
Dcr2-3	196054	99.7		76.7		23.0	

Table 2.1. The total number of reads, percent mapping, read depth, percent unique and percent non-unique are shown for each hight-throughput sequencing sample. Read depth = total # reads*100/30.1 Mb. A Students T-test was performed to deterimin if differences observed between % unique or % non-unique for Dcr2 and LacZ samples were statistically significant. The p-values for these tests are indicated.

LTR retroTns generate the majority of AS Tn transcripts

This analysis revealed S and AS nonunique reads for the majority of Tns examined (Table 2.2), while little to no AS transcription of non-Tn genes is evident (Table 2.3) These observations are consistent with previous data that *Drosophila* does not exhibit AS transcription upstream of mRNA genes (Lapidot and Pilpel 2006; Nechaev *et al.* 2010; Core *et al.* 2012). Tn nonunique RPM show that the most abundant and active *Drosophila* Tn class, LTR retroTns, is highly expressed in the S and AS direction (Kaminker *et al.* 2002; Kofler *et al.* 2015) (Table 2.2). Non-LTR retroTn and TIR DNA Tns are generally transcribed at lower levels (Table 2.2). These data are consistent with previous analyses showing that LTR retroTn-derived esiRNAs are more prevalent in S2 cells than esiRNAs originating from non-LTR retroTn or TIR DNA Tns (Ghildiyal *et al.* 2008).

While AS transcription is observed for all but the "I" family of Tns, the ratio of S to AS nonunique reads differs dramatically when S and/or AS RPM are >100. Non-LTR retroTns in the Jockey family and LTR retroTns in the Gypsy family have generally low S/AS ratios (2.88 and 3.34, respectively) while LTR retroTns in the Pao and copia families and TIR DNA Tns in the Pogo family have much higher S/AS ratios (6.81, 132.22, and 28.04, respectively) (Table 2.2). LTR retroTns generating the most esiRNAs in S2 cells all belong to the Gypsy family of retroTns (Chung *et al.* 2008; Czech *et al.* 2008; Kawamura *et al.* 2008). Because Gypsy retroTns Dm297, blood, and mdg1 generate abundant esiRNAs and produce ample AS transcripts (Table 2.4) these retroTns were chosen for further investigation. Additionally, only Tns in the Jockey family of non-LTR retroTns generate both S and AS transcripts (Table 2.1) and esiRNAs

(Kawamura *et al.* 2008). Jockey family members jockey and juan produce the highest levels of AS RNAs (Table 2.5). Therefore, to explore the importance of LTR sequences in retroTn AS transcription, juan and jockey were chosen for further study. TIR DNA Tns were not further investigated.

Table 2.2 *Drosophila* Transposons Sorted by Class.

Class	Family	No. Tns	%S	%AS	Avg S RPM	Avg AS RPM	S/AS
Non-LTR	Jockey	5	80.0	80.0	302.0	104.8	2.88
	I	1	0.0	0.0	0.0	0.0	—
	R1	2	0.0	50.0	0.0	32.0	—
LTR	Gypsy	18	66.7	72.2	871.3	261.1	3.34
	Pao	3	100.0	100.0	1046.7	153.7	6.81
	Copia	1	100.0	100.0	64258.0	486.0	132.22
TIR	ProtoP	1	100.0	100.0	80.0	31.0	2.58
	Tc1	5	0.0	40.0	0.0	24.5	—
	Transib	1	100.0	100.0	48.0	21.0	2.29
	Pogo	1	100.0	100.0	785.0	28.0	28.0

Table 2.2. Transposons sorted by class and family were analyzed. No. Tns, the number of individual Tns within a family having three or more full-length elements; %S or %AS, the percentage of Tns included in column 3 with S or AS nonuniquely mapping RNA-seq reads; AVG S or AVG AS, the average normalized nonunique S or AS read count (RPM) for each Tn family; and S/AS, the ratio of S RPM to AS RPM.

Table 2.3 Highly Transcribed Genes Show Little AS Transcription

Gene	S RPM	AS RPM
GAPDH1	573	3
GAPDH2	615	1
Groucho	48	3
Armadillo	372	0
Pumillio	208	0
Succinate Dehydrogenase A (SdhA)	108	0
RpL 32	3850	11
RpL 0	2329	0
Stem Loop Binding Protein (SLBP)	69	2
Cleavage and Polyadenylation Specificy Protein (CPSF) 100	35	3

Table 2.3. Sense and Antisense RPM for highly expressed genes.

Table 2.4 List of LTR Retrotransposons

LTR

retroTn	Family	Size (bp)	Genomic location	S	AS
17.6{}790	Gypsy	7494	2R:5614174-5621667	+++(1120)	+++(469)
17.6{}804	Gypsy	7494	2R:6835588-6843081	+++	+++
17.6{}1287	Gypsy	7475	3R:6629950-6637424	+++	+++
297{}832	Gypsy	6992	2R:10972539-10979530	+++++(3848)	++++(571)
297{}388	Gypsy	6997	2L:16153791-16160787	+++++	++++
297{}407	Gypsy	6978	2L:19147425-19154402	+++++	++++
3S18{}853	Pao	6130	2R:14463358-14469487	++++(2171)	+(95)
3S18{}4	Pao	6127	X:322507-328633	++++	+
3S18{}35	Pao	6127	X:3309106-3315232	++++	+
412{}880	Gypsy	7567	2R:19801877-19809443	+(54)	+(40)
412{}881	Gypsy	7521	2R:20034646-20042166	+	+
412{}882	Gypsy	7428	2R:20064814-20072241	+	+
blood{}852	Gypsy	7413	2R:14375381-14382793	++++(1531)	+++++(694)
blood{}856	Gypsy	7412	2R:15603415-15610826	+++	+++++
blood{}280	Gypsy	7443	2L:347941-355383	+++	+++++
Burdock{}770	Gypsy	6412	2R:3703232-3709643	+(188)	++(152)
Burdock{}783	Gypsy	6413	2R:5038862-5045274	+	++
Burdock{}514	Gypsy	6413	2L:21691882-21698294	+	++
copia{}631	Copia	5145	2R:1105258-1110402	>+++++(64258)	+++(486)
copia{}837	Copia	5151	2R:12427198-12432348	>+++++	+++
copia{}840	Copia	5146	2R:13124069-13129214	>+++++	+++
diver{}782	Pao	6133	2R:4665472-4671604	++(785)	++(163)
diver{}839	Pao	6132	2R:13063915-13070046	++	++
diver{}873	Pao	6112	2R:18467972-18474083	++	++
HMS-Beagle{}318	Gypsy	7062	2L:6991487-6998548		
HMS-Beagle{}333	Gypsy	7062	2L:9973781-9980842		
HMS-Beagle{}333	Gypsy	7072	2L:12558375-12565446		
Invader2{}633	Gypsy	5075	2R:1115288-1120362		
Invader2{}563	Gypsy	5045	2L:22329289-22334333		
Invader2{}1169	Gypsy	5056	3L:23264861-23269916		
invader3{}695	Gypsy	5474	2R:1509626-1515099		
invader3{}240	Gypsy	5382	X:21818924-21824305		
invader3{}751	Gypsy	5477	2R:2358324-2363800		
mdg1{}831	Gypsy	7367	2R:10903017-10910383	+++(1192)	+++++(788)
mdg1{}859	Gypsy	7355	2R:15802405-15809759	+++	+++++
mdg1{}885	Gypsy	7451	2R:20615462-20622912	+++	+++++
mdg3{}119	Gypsy	5520	X:13357733-13363252	+++(1153)	++(140)
mdg3{}144	Gypsy	5520	X:16386734-16392253	+++	++
mdg3{}291	Gypsy	5520	2L:1801273-1806792	+++	++
opus{}760	Gypsy	7525	2R:2839724-2847248	+(106)	++(124)
opus{}821	Gypsy	7602	2R:9615692-9623293	+	++
opus{}127	Gypsy	7604	X:14445732-14453335	+	++
Quasimodo{}352	Gypsy	7387	2L:12781858-12789244		+(56)
Quasimodo{}360	Gypsy	7379	2L:13449517-13456895		+
Quasimodo{}1186	Gypsy	7355	3R:84350-91704		+
roo{}796	Pao	9109	2R:6064440-6073548	+(184)	++(203)
roo{}806	Pao	9116	2R:6897375-6906490	+	++
roo{}828	Pao	9094	2R:10354854-10363947	+	++
rover{}1212	Gypsy	7320	3R:713256-720575		
rover{}1275	Gypsy	7412	3R:4732111-4739522		
rover{}133	Gypsy	7470	X:14928180-14935649		
springer{}300	Gypsy	7510	2L:3251549-3259058		
springer{}1464	Gypsy	7543	3R:26900006-26907548		
springer{}59	Gypsy	7510	X:4990361-4997870		
Stalker{}174	Gypsy	7256	X:19691436-19698691	+(66)	+(35)
Stalker{}1277	Gypsy	7230	3R:5130306-5137535	+	+
Stalker{}1427	Gypsy	7255	3R:22384105-22391359	+	+
Stalker2{}1505	Gypsy	8119	4:340670-348788	++(414)	+(80)
Stalker2{}22	Gypsy	7883	X:1848974-1856856	++	+
Stalker2{}1042	Gypsy	7895	3L:18703914-18711808	++	+
tirant{}797	Gypsy	8527	2R:6473118-6481644	++(602)	++(203)
tirant{}833	Gypsy	8425	2R:11006877-11015301	++	++
tirant{}834	Gypsy	8527	2R:11228251-11236777	++	++
Transpac{}32	Gypsy	5249	X:2969722-2974970	+(182)	+(42)
Transpac{}1439	Gypsy	5248	3R:23688731-23693978	+	+
Transpac{}362	Gypsy	5249	2L:13522313-13527561	+	+

Table 2.4. Family, size, genomic location, S and AS Transcription levels are shown for each individual LTR retrotransposon. The '+' represents relative transcription among Tns and the number shown within the () is the normalized nonunique read count (RPM) corresponding to that individual Tns.

Table 2.5 List of Non-LTR Retrotransposons and TIR Transposons

non-LTR

RetroTn	Family	Size (bp)	Genomic location	S (RPM)	AS (RPM)
BS{}707	Jockey	5128	2R:1984289-1989416		
BS{}1260	Jockey	5142	3R:3869566-3874707		
BS{}1292	Jockey	5122	3R:7666843-7671964		
Doc{}772	Jockey	4719	2R:3756749-3761467	+(203)	+(33)
Doc{}819	Jockey	4726	2R:9024402-9029127	+	+
Doc{}827	Jockey	4721	2R:10281792-10286512	+	+
F{}731	Jockey	4699	2R:2199985-2204683	+(78)	+(37)
F{}755	Jockey	4707	2R:2382090-2386796	+	+
F{}763	Jockey	4710	2R:3084132-3088841	+	+
I{}769	I	5133	2R:3495977-3501109		
I{}129	I	5371	X:14530558-14535928		
I{}18	I	1727	X:1461750-1463476		
jockey{}817	Jockey	5010	2R:8707006-8712015	+(265)	++(108)
jockey{}838	Jockey	4959	2R:13034810-13039768	+	++
jockey{}277	Jockey	5006	2L:47514-52519	+	++
Juan{}768	Jockey	4236	2R:3322453-3326688	++(662)	+++++(241)
Juan{}138	Jockey	4226	X:15307142-15311367	++	+++++
Juan{}1190	Jockey	4232	3R:234195-238426	++	+++++
Rt1a{}1276	R1	5177	3R:5104562-5109738		
Rt1a{}905	R1	5193	3L:1424822-1430014		
Rt1a{}1390	R1	5175	3R:15570024-15575198		
Rt1b{}334	R1	5171	2L:10138214-10143384		+(32)
Rt1b{}1218	R1	5027	3R:1164177-1169203		+
Rt1b{}1288	R1	5170	3R:6783807-6788976		+

TIR

Tn	Family	Size (bp)	Genomic location	S (RPM)	AS (RPM)
1360{}1226	protop	1107	3R:1648157-1649263	+(80)	+(31)
1360{}136	protop	1107	X:15167794-15168900	+	+
1360{}1498	protop	1084	4:315271-316354	+	+
Bari1{}1534	Tc1	1728	4:860624-862351		
Bari1{}282	Tc1	1728	2L:770516-772243		
Bari1{}1409	Tc1	1728	3R:19384173-19385900		
HB{}276	Tc1	1573	X:22243764-22245336		
HB{}512	Tc1	1636	2L:21673477-21675112		
HB{}761	Tc1	1633	2R:2876633-2878265		
hopper{}82	Transib	1433	X:7739814-7741246	+(48)	+(21)
hopper{}105	Transib	1435	X:11162651-11164085	+	+
hopper{}1432	Transib	1432	X:21841727-21843158	+	+
mariner2{}1130	Tc1	1110	3L:23131582-23132691		
mariner2{}767	Tc1	983	2R:3281653-3282635		
mariner2{}522	Tc1	879	2L:22004569-22005447		
pogo{}297	Pogo	2122	2L:2933354-2935475	++(785)	+(28)
pogo{}400	Pogo	2134	2L:17912012-17914134	++	+
pogo{}1294	Pogo	2122	3R:7848660-7850781	++	+
S{}173	Tc1	1731	X:19604159-19605889		+(23)
S{}758	Tc1	1735	2R:2551739-2553473		+
S{}1207	Tc1	1704	3R:508954-510657		+
FB{}1449	Tc1	1310	3R:24905161-24906470		+(26)
FB{}2296	Tc1	1592	2L:22250953-22252544		+
FB{}5386	Tc1	1325	4:601549-602873		+

Table 2.5. Family, size, genomic location, S and AS Transcription levels are shown for each individual Tns. The '+' represents relative transcription among Tns and the number shown within the () is the normalized nonunique read count (RPM) corresponding to that individual Tns.

LTR retroTn AS transcription initiates from within or near LTRs

Bedgraphs of nonunique, strand-specific RNA-seq reads mapping to Dm297 (Figure 2.1B), blood (Figure 2.1C), and mdg1 (Figure 2.1D, right half of mdg1{}1720) representative full-length elements are shown. These data indicate that both S and AS transcripts are distributed across the elements including the three ORFs

(Figure 2.1, B–D). Dm297, blood, and mdg1 AS reads tend to be concentrated in the LTRs, while S RPM are higher in the ORFs (Figure 2.1, B–D). In all three cases, total S transcript levels were higher than AS (Figure 2.1, B–D, red numbers). Some sequences were removed by splicing (data not shown).

To identify S and potential AS transcription start sites (tss), we remapped publicly available short-capped RNA high-throughput sequencing datasets (Nechaev *et al.* 2010; Henriques *et al.* 2013) and filtered to isolate only nonunique reads. Potential S and AS tss were observed for all three LTR retroTns (Figure 2.1, B–D, blue). RPM for tss on the S strand were higher than RPM for AS tss for Dm297 and blood (Figure 2.1, B and C, blue numbers). These data correlate with S and AS transcription levels for each retroTn. For Dm297 and blood, S and AS transcription could begin near the 3' end of the LTR (Figure 2.1, B and C, top, blue) and transcription initiating from this location could result in S and AS RNAs spanning the entire element (Figure 2.1, B and C, blue).

In contrast, no LTR AS tss were observed for full-length mdg1 elements. An AS tss is visible (Figure 2.1D, right, bottom, blue); however, transcription from this tss would only produce AS RNAs of the 5' LTR. One mdg1 element, mdg1{}1720, was identified that could produce the observed AS transcripts. Mdg1{}1720 consists of two tandemly repeated mdg1 elements with an inverted and centralized LTR in the RT ORF of the first

mdg1 retroTn (Figure 2.1D). Transcription initiating from the non-LTR tss in the downstream mdg1 element could result in AS transcription of the first mdg1 retroTn. While only nonuniquely mapping AS reads are shown in Figure 2.1D, AS reads corresponding to unique sequences in the upstream mdg1 repeat are observed (data not shown), indicating that this specific retroTn, mdg1{}1720, is transcribed.

We next performed strand-specific RT-qPCR (Purcell *et al.* 2006; Vashist *et al.* 2012) to confirm S and AS transcription of Dm297, blood, and mdg1. Each potential transcript was reverse transcribed with a strand-specific, gene-specific primer having a unique nucleic acid tag. The unique tag provided a primer binding site for the downstream qPCR reaction to ensure detection of only the transcript of interest. A list of the primer sequences used can be found in Table 4.1 in the Materials and Methods section. Random priming was evaluated in the absence of an RT primer and no target transcripts were detected (data not shown). For Dm297, blood, and mdg1, we detected both S and AS transcription using several primer sets spanning the coding sequence. We calculated the difference between S and AS Cts (ΔCt(S − AS)) for Dm297, blood, and mdg1 (Table 2.7). AS transcripts were less abundant in all cases and the differences between S and AS transcription correlated with RNA-seq data (Figure 2.1, B–D).

Lastly, we performed Northern blot analysis to detect S and AS transcripts. Sequences of Northern blot probes can be found in Table 4.2 in the Materials and Methods section**.** RNA from Dmel-2 cells was transferred to a membrane and probed with radioactively labeled Dm297, blood, and mdg1 complementary S and AS probes. S (Figure 2.1E) and AS (Figure 2.1F) transcripts were observed for Dm297, blood, and

mdg1. S Dm297 (6995 nt) appears slightly smaller than its predicted size while S blood (7410 nt) and mdg1 (7480 nt) are slightly larger (Figure 2.1E). No other prominent bands were detected. The Dm297 S transcript is most abundant while mdg1 and blood S transcripts are much less prevalent. Each lane contained equal amounts of 28S rRNA (Figure 2.1E, bottom). These data correlate with LTR retroTn S transcript levels observed in the RNA-seq data (Figure 2.1, B–D). We observe multiple Dm297 and blood AS transcripts resulting in a smear between 8 and 10 kb, while a single AS mdg1 transcript is present at ∼8 kb (Figure 2.1F). An ∼2 kb mdg1 AS transcript is also visible (Figure 2.1F). The sizes of these RNAs support AS transcription of full-length Dm297 and blood LTR retroTns from the bioinformatically identified tss (Figure 2.1, B–D). These data also indicate that the mdg1 AS tss (Figure 2.1D, first and third blue peaks) is functional, producing the predicted ∼8 kb and ∼2 kb RNAs, while the inverted LTR S tss is not active in this context. Multiple Dm297 AS transcripts may result from inefficient RNAPII termination. A lighter exposure (Figure 2.1F, bottom "L") of the AS transcripts reveals that the ∼8 kb AS mdg1 transcript is most abundant, while Dm297 and blood AS RNAs are less prevalent, mirroring the RNA-seq data (Figure 2.1, B–D). S2 culture cells have amplified Tn content (Potter *et al.* 1979; Junakovic *et al.* 1988; Wen *et al.* 2014) and RNA-seq reads cannot be mapped to the S2-specific mdg1 copies as the S2 genome has not been sequenced. Regardless, the data presented here support AS transcription of mdg1{}1720 from non-LTR tss.

To gain a more complete view of AS retroTn transcription, we considered the genomic context of each annotated full-length Dm297, blood, and mdg1 element (Table

2.6) Individual retroTns were found both intergenically and within introns. Intronic Dm297 and mdg1 retroTns were more often AS to their host coding gene, while blood elements were in the S orientation. These data indicate that transcription of intronic LTR retroTns, together with RNAs produced from intraelement tss, could contribute to AS transcript abundance.

We also investigated S and AS transcription of individual Dm297, blood, and mdg1 elements (Table 2.6). If unique intraelement RNA-seq reads and RNA-seq reads corresponding to the retroTn-intergenic/intronic sequence junction could be identified, we concluded that the individual Dm297, blood, or mdg1 element was transcribed. S transcription was confirmed for 9/18 (50%) of Dm297, 2/15 (13%) of mdg1, and 9/22 (41%) of blood retroTns (Table 2.6). AS transcription was verified for 4/18 (22%) of Dm297, 1/15 (7%) of mdg1, and 3/22 (14%) of blood elements (Table 2.6). Of all mdg1 retroTns, only AS transcription of mdg1{}1720 could be confirmed using this analysis. These numbers of transcribed individual elements are probably an underestimate of the total as not all elements have mutations allowing observation of unique internal reads. Also, RPM for unique reads were low, reflecting less RNA from one individual element as compared to nonunique RPM corresponding to total transcription of all individuals of a retroTn type. Collectively, the RNA-seq analyses, strand-specific RT-qPCR, and Northern blots indicate that individual LTR retroTns in Dmel-2 cells undergo convergent S and AS transcription and that transcription can initiate within or near Dm297, blood, and mdg1 LTRs.

Fig. 2.1 Bedgraphs and Northern Blot Analysis of LTR RetroTns.

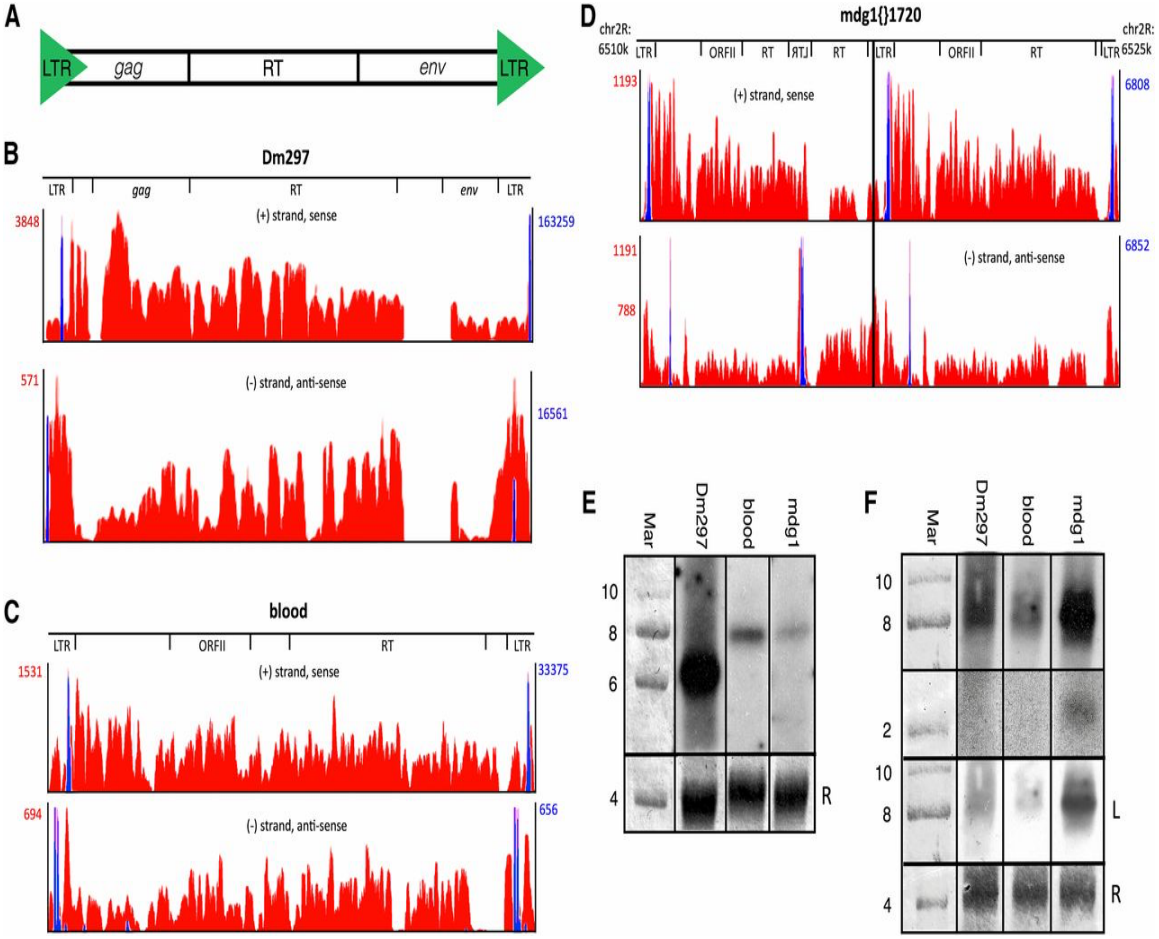


Fig. 2.1 LTR retroTns Dm297, blood, and mdg1{}1720 produce AS transcripts from intraelement tss in or near the LTRs. (A) Schematic of *Drosophila* LTR retroTns. (B–D) Bedgraphs representing S (top) and AS (bottom) nonunique RNA-seq reads mapping to each LTR retroTn are shown in red. Peak reads per million (RPM) are listed to the left (red numbers). For mdg1, two AS RPM values are listed; the top is the RPM for mdg1{}1720 and the bottom is the RPM for only the downstream canonical mdg1 element (right of the black line). Only the chromosome location of mdg1{}1720 is shown as Dm297 and blood bedgraphs are representative examples. Relative locations of specific ORFs are shown above the bedgraphs. Nonunique small-capped RNA-seq reads representing tss are overlaid in blue and RPM values are listed to the right (blue numbers). (E) Representative Northern blots of S LTR retroTn transcripts. The probe used for each blot is indicated above. The first lane is methylene blue-stained RNA marker; the sizes of bands are shown to the left of the blots. Methylene blue-stained 28S rRNA is used as a loading control (bottom) and is marked with an "R." (F) Representative Northern blots of AS LTR retroTn transcripts. The top two panels are from the same longer exposure film while the third panel ("L") is a lighter exposure. Other details are as in E.

Table 2.6 Individual LTR and Non-LTR Transposons.

LTR Transposons

retroTn	Length (bp)	Genomic location	UI-S	UI-AS	S-S	S-AS	intergenic	intron	S/AS to Tn?
297{}832	6992	2R:10972539-10979530	x		x		x		
297{}388	6997	2L:16153791-16160787	x	x	x	x	x		
297{}407	6978	2L:19147425-19154402	x	x	x			brat	S
297{}1440	6998	3R:23701477-23708474	x		x		x		
297{}48	6978	X:4051980-4058957	x		x	x		Fas2	AS
297{}92	6995	X:9692611-9699605			x			CG32698	AS
297{}98	6996	X:10419159-10426154			x			spri	AS
197{}109	7013	X:11527505-11534517		x				ptp10D	AS
297{}327	6997	2L:8594520-8601516	x					Sema-1a	AS
297{}338	6999	2L:10876598-10883596		x	x	x	x		
297{}346	6995	2L:12067385-12074379	x	x			x		
297{}376	6995	2L:15586465-15593459	x		x		x		
297{}766	6996	2R:3149747-3156742	x		x	x	x		
297{}897	6996	3L:402740-409735	x	x	x	x	x		
297{}950	6995	3L:7786867-7793861			x	x		CG32369	S
297{}1107	6917	3L:22983017-22989933		x	x	x	x		
297{}323	6917	2L:7977135-7984051	x		x			snoo	AS
297{}1286	6917	3R:6167388-6174304	x			x	x		
mdg1{}831	7367	2R:10903017-10910383	x		x			hbs	AS
mdg1{}859	7355	2R:15802405-15809759	x	x			x		
mdg1{}885	7451	2R:20615462-20622912						CG33988	AS
mdg1{}299	7372	2L:3202794-3210165	x				x		
mdg1{}305	7384	2L:4601759-4609142	x			x	x		
mdg1{}1280	7365	3R:5663885-5671249	x	x	x			Teh1	AS
mdg1{}1403	7335	3R:17586899-17594233			x			CG42335	AS
mdg1{}1442	7373	3R:23999447-24006819				x		CG34354	S
mdg1{}900	7369	3L:925231-932599	x					Glut1	S
mdg1{}1047	7390	3L:19110198-19117587	x			x	x		
mdg1{}1678	7403	2R:6546757-6554159	x				x		
mdg1{}29	7353	X:2738988-2746340	x	x			x		
mdg1{}914	7369	3L:2966778-2974144	x				x		
mdg1{}1610	7273	3L:17724006-17731278		x				C cn	AS
mdg1{}1720	18802	2R:6509904-6528705	x	x		x		CG11883	S
blood{}852	7413	2R:14375381-14382793	x		x			CG30116	AS
blood{}856	7412	2R:15603415-15610826	x			x	x		
blood{}280	7443	2L:347941-355383	x		x	x	x		
blood{}285	7409	2L:1220184-1227592	x	x	x	x		CG42329	S
blood{}289	7408	2L:1679032-1686439		x	x			chinmo	S
blood{}335	7413	2L:10156377-10163789		x			x		
blood{}344	7411	2L:11713562-11720972			x	x	x		
blood{}356	7410	2L:12933905-12941314				x	x		
blood{}375	7412	2L:15446106-15453517	x		x		x		
blood{}468	7411	2L:20303216-20310626	x				x		
blood{}472	7411	2L:20576274-20583684	x		x		x		
blood{}488	7415	2L:21347605-21355019	x					Tsp39D	S
blood{}1369	7417	3R:11444508-11451924	x		x		x		
blood{}1376	7409	3R:13401108-13408516	x		x		x		
blood{}1389	7411	3R:15531296-15538706			x		x		
blood{}1462	7412	3R:26505296-26512707	x	x		x	x		
blood{}1470	7413	3R:27823769-27831181	x		x	x	x		
blood{}218	7415	X:21407620-21415034	x	x			x		
blood{}409	7417	2L:19347541-19354957	x		x			dnt	AS
blood{}959	7410	3L:8491452-8498861					x		
blood{}1092	7413	3L:22548925-22556337	x	x	x		x		
blood{}1094	7395	3L:22610345-22617739	x	x	x	x	x		

Non-LTR Transposons

retroTn	Length (bp)	Genomic location	UI-S	UI-AS	S-S	S-AS	intergenic	intron	S/AS to Tn?
juan{}768	4236	2R:3322453-3326688			x		x		
juan{}138	4226	X:15307142-15311367	x	x	x			CG18210	AS
juan{}1190	4232	3R:234195-238426	x	x	x	x		CG32944	AS
juan{}257	4235	X:21953676-21957910	x	x			x		
juan{}83	4235	X:7925528-7929762				x	x		
juan{}266	4230	X:22099389-22103618			x	x	x		
juan{}2251	4236	2R:2298595-2302830				x	x		
jockey{}765	5014	2R:3123380-3128393	x	x			x		
jockey{}807	4985	2R:6907459-6912443	x	x				luna	S
jockey{}817	5010	2R:8707006-8712015		x			x		
jockey{}838	4959	2R:13034810-13039768	x	x			x		
jockey{}277	5006	2L:47514-52519			x			CG31973	AS
jockey{}307	5002	2L:4918233-4923234	x		x			hoe1	AS
jockey{}477	5017	2L:20891618-20896634			x			CG9339	AS
jockey{}1238	5017	3R:2335803-2340819	x				x		
jockey{}1447	5015	3R:24618648-24623662					x		
jockey{}51	5018	X:4625918-4630935	x	x			x		
jockey{}261	5023	X:21966895-21971917	x	x	x		x		
jockey{}973	5011	3L:9569071-9574081						CG32048	AS
jockey{}1086	5087	3L:21742484-21747570						Syn1	S
Jockey{}1630	14127	2R:14245969-14260095					x		

Table 2.6 Size and genomic location of all full-length juan and jockey elements are shown. The presence of unique internal S or AS reads (UI-S or UI-AS) or, S or AS reads corresponding to retroTn-external sequence junctions produced by splicing (S-S or S-AS) are shown with an 'x.' Whether the retroTn is intergenic or intragenic is shown. If the retroTn is contained within an intron, the gene name is shown. The last column describes whether an intronic retroTn is S or AS to the S strand of the host gene. 'S' indicates that S retroTn RNA is also the S mRNA while 'AS' indicates that the AS retroTn transcript is also the S mRNA.

Table 2.7 Strand Specific RT qPCR of RetroTns

LTR:

Tn-amplicon	Ave. ΔCt(S-AS)	SD
Dm297-RT	-5.23	0.18
Dm297-env	-4.38	0.12
blood-ORFII	-4.31	0.10
blood-RT	-0.86	0.08
mdg1-ORFII	-3.02	0.13
mdg1-RT	-2.00	0.07

Non-LTR:

Tn-amplicon	Ave. ΔCt(S-AS)	SD
juan-ORF1	-2.44	0.12
juan-RT	-2.01	0.22
jockey-gag	-0.46	0.47
jockey-RT	1.86	0.23

Table 2.7 Strand specific pPCR with primers to multiple retroTn ORFs (first column) was performed in triplicate. The average difference between S and AS Ct values and standard deviation (SD) were calculated.

Non-LTR retroTns juan and jockey produce AS transcripts

To investigate the role of LTRs in AS transcription, we examined non-LTR retroTns juan and jockey. Similar to LTR elements, strand-specific nonunique RNA-seq reads map the entire length of jockey and juan (Figure 2.2, B and C, red); however, non-LTR S and AS transcripts are less abundant than corresponding LTR retroTn RNAs. Additionally, jockey is the only retroTn investigated for which the AS transcript is more highly expressed than the S transcript.

A juan S tss is observed at the 5' end of the retroTn (Figure 2B, top, blue) and initiation of transcription from this location could result in a single complete element S transcript. Several AS tss were also observed; however, these tss cannot be responsible for reads mapping to the 3' half of juan (Figure 2.2B, bottom, blue). The source of these AS RNA-seq reads is unclear, although unique AS reads were identified for both intergenic and intragenic juan elements (Table 2.6), indicating AS transcription of individual elements (see previous section). Two S tss are observed for jockey. Transcription beginning at these tss could produce a S transcript the entire length of the element (Figure 2.2C, top, blue). For jockey, a potential AS tss is observed at the 5' end, but the number of reads mapping to this tss do not correlate with higher AS RPM (Figure 2.2C, bottom, blue). Collectively, these data indicate that non-LTR retroTns are transcribed in both S and AS directions, albeit at lower levels than LTR retroTns.

To verify S and AS transcription, we performed strand-specific qPCR of non-LTR retroTns juan and jockey as described in the previous section. S and AS transcription were detected for juan and jockey and the $\Delta Ct(S - AS)$ mirrored the ratio of S to AS RPM

of each ORF investigated (Table 2.6). Additionally, Northern blot analysis revealed juan S and AS, and jockey AS transcription (Figure 2.2, D and E); however, the S jockey transcript was not visible presumably because of its low abundance. Previously, S jockey transcripts initiating from an internal promoter were identified in *Drosophila* cell culture (Mizrokhi *et al.* 1988). A probe to the S juan transcript (4236 nt) reveals one band ~5 kb, while AS jockey (5020 nt) and AS juan probes show smears indicating multiple transcripts (Figure 2.2, D and E). Juan AS RNAs range from ~2 kb to ~4 kb (Figure 2.2E). If juan AS transcription initiates from bioinformatically identified tss (Figure 2.2B) and RNAPII termination is inefficient, multiple ~2 to ~4 kb AS transcripts could be produced. Jockey AS RNAs range from ~7 kb to greater than 10 kb (Figure 2.2D). We hypothesize that smaller transcripts in this range could be produced from the bioinformatically identified tss (Figure 2.2C). Additionally, one jockey retroTn (jockey{}1630), has a LTR retroTn, roo (9092 nt), inserted in the *gag* ORF making this element 14,127 nt. Transcripts originating from an observed tss (data not shown) at the 5' end of jockey{}1630 could result in jockey AS RNAs greater than 10 kb. These data suggest that non-LTR retroTns juan and jockey are transcribed in both the S and AS directions from intraelement tss.

Figure 2.2 Bedgraphs and Northern Blot Analysis of Individual Non-LTR RetroTns.

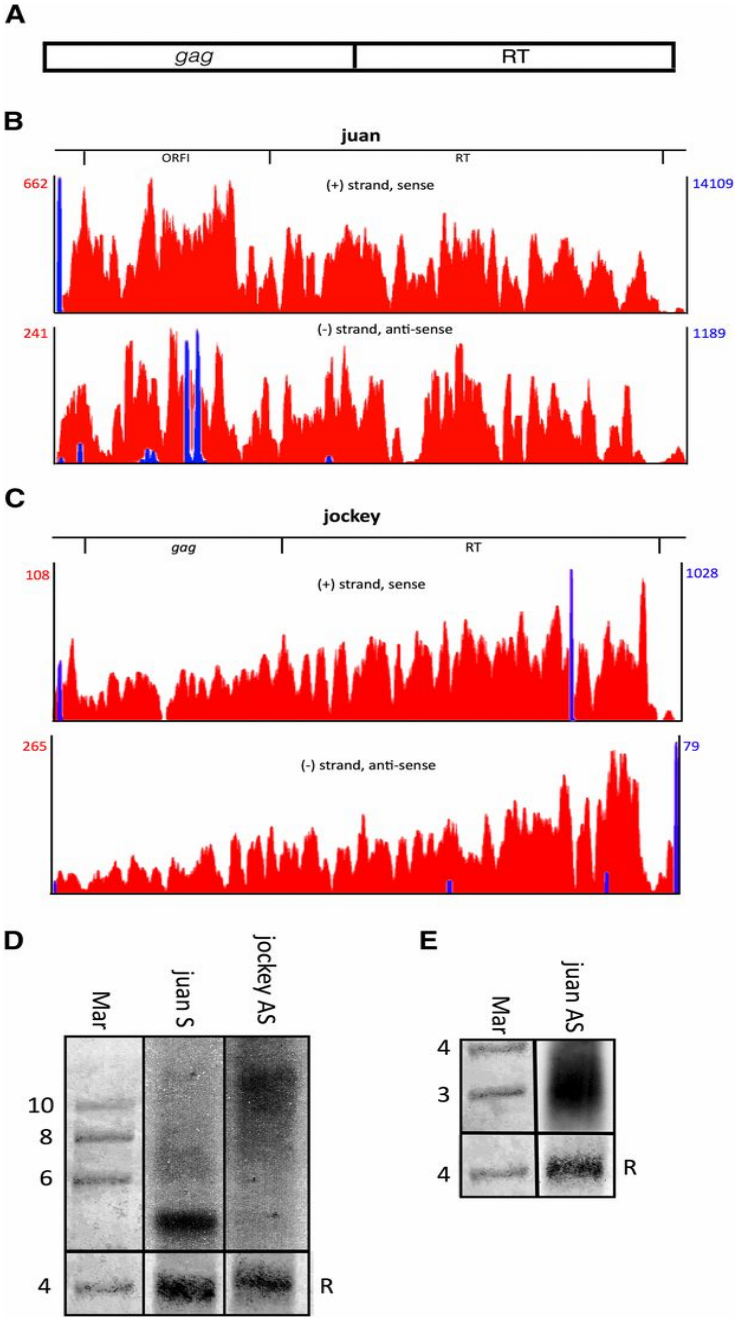


Fig. 2.2 Non-LTR retroTns juan and jockey produce AS transcripts from intraelement tss. (A) Schematic of non-LTR retroTns in *Drosophila*. (B and C) Bedgraphs representing S (top) and AS (bottom) nonunique RNA-seq reads mapping to each non-LTR retroTn are shown in red. Other details are as described in Figure 2.1. (D and E) Representative Northern blots of juan S and jockey AS (D), and juan AS (E) transcripts are shown. Details are as in Figure 2.1E.

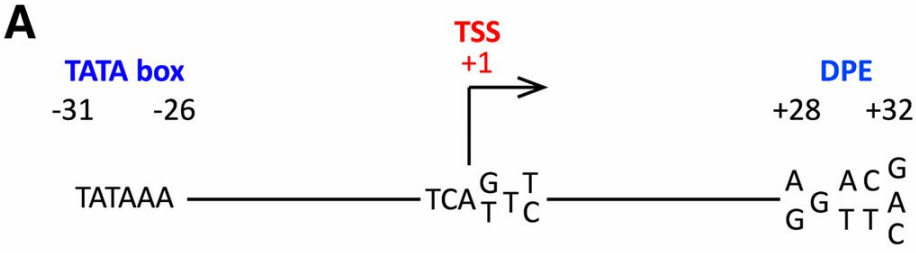
S and AS tss have canonical *Drosophila* promoter elements

We next wanted to determine if the observed LTR and non-LTR retroTn tss were flanked by traditional *Drosophila* promoter elements. *Drosophila* transcription initiates at T-C-A+1-G/T-T-T/C (where A+1 is the tss) within a promoter composed of a TATA box (−31 to −26) and/or a downstream promoter element (DPE) located between +28 to +32 (Butler and Kadonaga 2002) (Figure 2.3A). The TATA box or

DPE occur in core promoters 29% and 26% of the time, respectively, while 14% contain both a TATA box and a DPE (Butler and Kadonaga 2002). Further evaluation of Dm297 revealed near-canonical tss and DPEs in appropriate locations for both S and AS promoters (+28 and +29, respectively, Figure 2.3B). Blood S and AS transcripts initiate from tss having two noncanonical bases but also have canonical DPEs +28 from S and AS tss (Figure 2.3B). The mdg1 LTR has a canonical tss with a perfect DPE +28 downstream, while the AS tss not located in the LTR has two nonideal bases and an inappropriately spaced DPE (Figure 2.3B). These data support previous characterization of the mdg1 S promoter (Arkhipova and Ilyin 1991). The non-LTR retroTn juan has a near canonical S tss and a canonical AS tss. Both tss have canonical DPEs +28 bases downstream of the tss (Figure 2.3B). These data support bonafide S tss for all three LTR-retroTns and non-LTR retroTn juan. Finally, promoter analysis revealed no canonical initiation site or DPE for either S or AS jockey tss. We hypothesize that this promoter is unique compared to more canonical core *Drosophila* promoters.

Figure 2.3 *Drosophila* Promoter Element Analysis

A



B

Tn	TSS	TSS 5'-3'	S DPE	AS DPE	S DPE Pos.	AS DPE Pos.
Dm297	S-TSS	T**T**AGTC	AGTCG	-	28	-
Dm297	AS-TSS	**G**CAGTC	-	GGATG	-	29
blood	S-TSS	T**T**AG**A**C	AGA**A**C	-	28	-
blood	AS-TSS	T**G**A**C**TT	-	GGT**C**A	-	28
mdg1	S/AS-TSS	TCAGTT	AGATC	-	28	-
mdg1	AS-TSS_2	TCA**CA**T		GGTCA		38
juan	S-TSS	**A**CAGTC	GGATG	-	28	-
juan	AS-TSS	TCAGTT		AGTTA	-	28
jockey	S-TSS		-	-	-	-
jockey	AS-TSS		-	-	-	-

Fig 2.3 S and AS tss have canonical *Drosophila* RNAPII promoter elements. (A) A schematic representing canonical *Drosophila* promoter elements is shown. (B) HTS analysis at nucleotide resolution of LTR and non-LTR retroTns tss is depicted. The tss, tss sequence, S DPE, AS DPE, and position of each DPE are shown for each retroTn. Bold nucleotides represent divergence from canonical nucleotide/s shown in A.

LTR and non-LTR retroTn AS transcripts lack strong polyadenylation

Our data suggest that S and AS LTR and non-LTR retroTns are convergently transcribed from canonical *Drosophila* promoters. As these RNAs are likely RNAPII transcripts (Gogvadze and Buzdin 2009), we wanted to determine their polyadenylation status. S retroTn transcripts have canonical polyadenylation signals (Gogvadze and Buzdin 2009; this work) and polyadenylation of these RNAs has previously been reported (Gogvadze and Buzdin 2009). We first fractionated total RNA using an oligo d(T) column and then performed strand-specific qPCR to the RT ORF of each retroTn on total RNA, polyA+ RNA and polyA− RNA. We used the amount of transcript in total RNA to normalize polyA+ and polyA− fractions by subtracting polyA+ and polyA− Cq values from those of total RNA (ΔCq = total-polyA +/−) (Figure 2.4A). We then determined a fold enrichment of polyadenylation by calculating the difference between these ΔCq values (ΔΔCq = [(total-polyA +)-(total-polyA −)]) (Figure 2.4B). 18s rRNA (polyA−) and Actin (polyA+) were used as controls. Total RNA was efficiently separated into polyA+ and polyA− fractions as 18s S transcripts were >70-fold less in the polyA+ fraction than in total RNA (Figure 2.4A, 18s +) and actin S transcripts were ∼6-fold increased in the polyA+ fraction as compared to total RNA (Figure 2.4A, actin +). 18s S RNAs were more than 100-fold depleted in polyA+ transcripts, while Actin S RNAs were approximately 10-fold enriched in polyA+ transcripts (Figure 2.4B), indicating that our assay to assess polyadenylation was working properly.

S Dm297 and mdg1 transcripts were ∼10- and ∼5-fold more, respectively, in the polyA+ fraction than in total RNA and ∼2-fold less in the polyA− fraction than in total

RNA (Figure 2.4A, Dm297 and mdg1, blue). These S transcripts are enriched for polyadenylation at least as much (mdg1) if not more (Dm297) than the polyadenylated Actin control (Figure 2.4B). Dm297 and mdg1 AS transcripts, and blood, juan, and jockey S and AS transcripts were between ∼1.5- and ∼3-fold more in both polyA+ and polyA− fractions than in total RNA, indicating a mixture of both polyA+ and polyA− transcripts (Figure 2.4A). ΔΔCq calculations suggest that blood, juan, and jockey S transcripts are enriched for polyadenylation although much less than Dm297 and mdg1 (Figure 2.4B). None of the AS transcripts are highly enriched with polyA+ transcripts. Blood, mdg1, and juan are slightly enriched in polyA− RNAs (Figure 2.4B). These data suggest that while all retroTn S transcripts have polyadenylation signals, only Dm297 and mdg1 S transcripts are polyadenylated. Bioinformatic assessment did not reveal strong polyadenylation sites for any of the AS transcripts examined (data not shown). Collectively, these data support a hypothesis that retroTn AS transcripts are not strongly polyadenylated.

Figure 2.4 Polyadenylation Status of LTR and Non-LTR Transposons

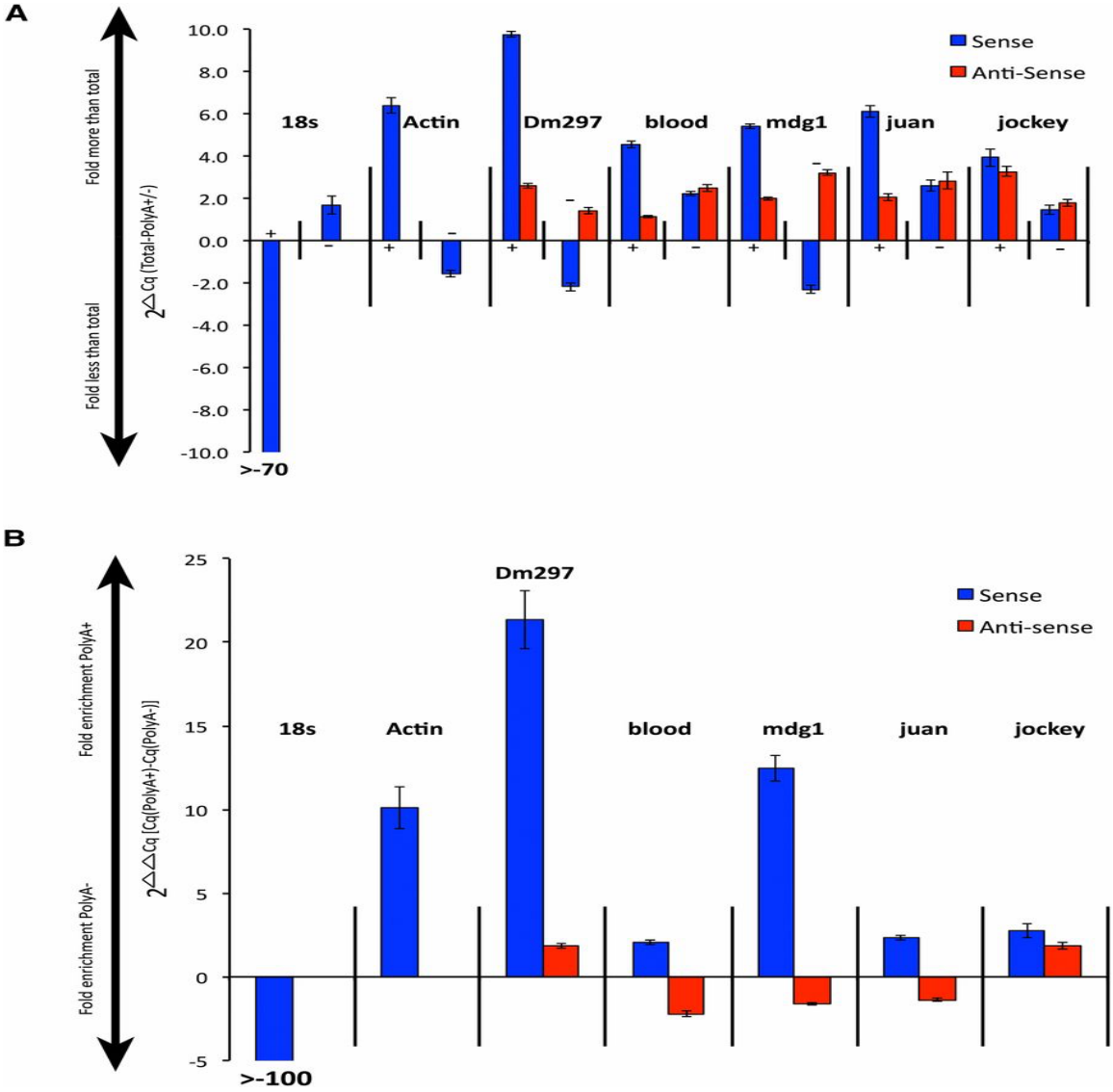


Fig. 2.4 LTR and non-LTR retroTn AS transcripts lack strong polyadenylation. (A) Graph of S and AS transcript fold differences in polyA+ or polyA− fractions compared to total RNA. $2^{\Delta Cq}$ is the *y*-axis and represents (total RNA – polyA+ or polyA−). PolyA+ or polyA− fractions are indicated along the *x*-axis as + or – signs. S transcripts are blue bars and AS transcripts are red bars. The name of each retroTn or control is listed above the appropriate group. Error bars represent standard deviation of strand-specific qPCR technical triplicates. (B) Graph of direct comparison of polyA+ to polyA− levels for each retroTn S/AS transcript pair. Fold enrichment values for polyA+ or polyA− fractions are shown along the *y*-axis ($2^{\Delta\Delta Cq}$) where the ΔΔCq equals [(total-polyA +)-(total-polyA −)]). S and AS transcripts are shown along the *x*-axis; S bars are blue and AS bars are red. Error bars represent standard deviation of strand-specific qPCR technical triplicates.

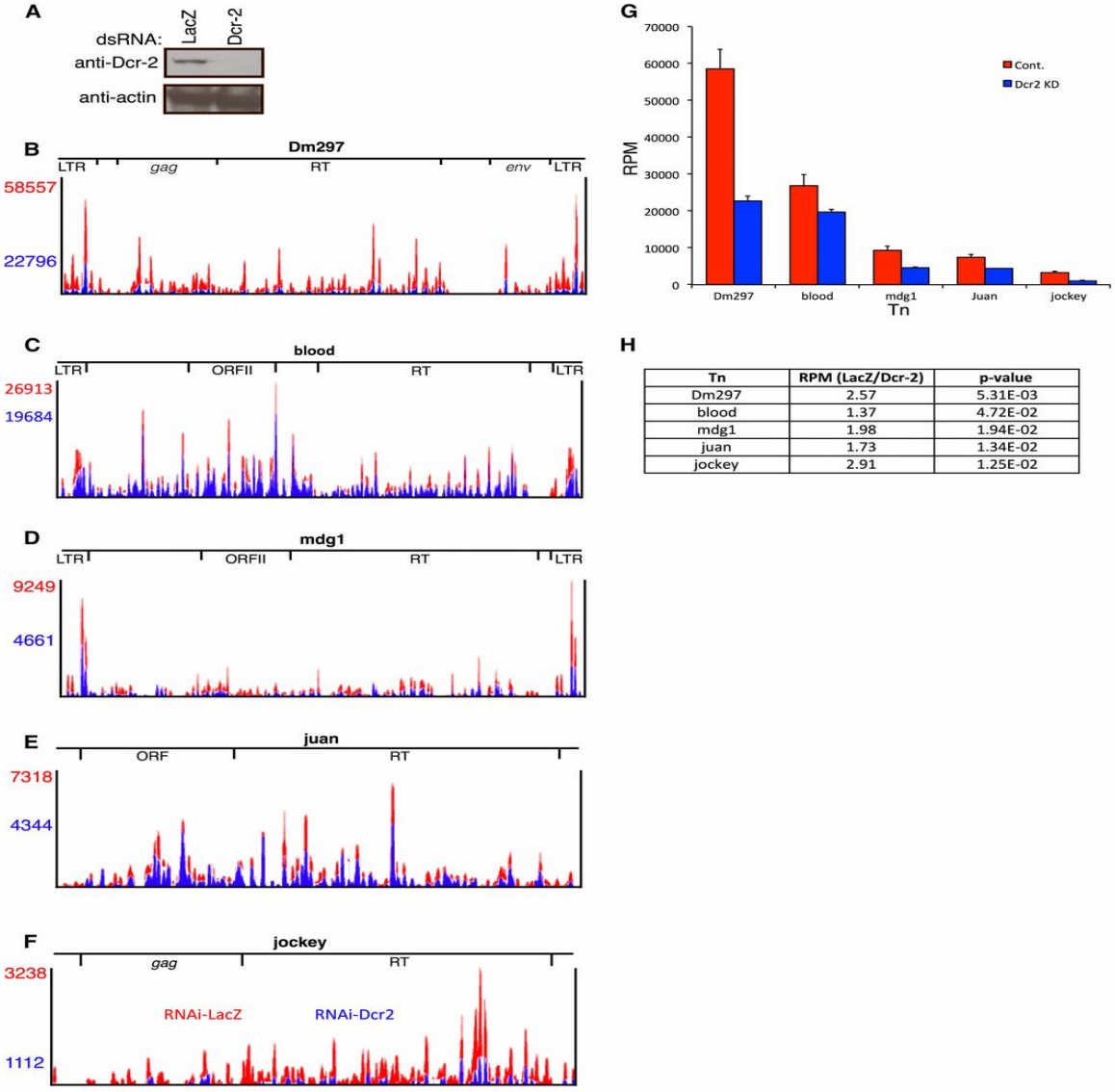
Dcr-2 depletion decreases retroTn-derived esiRNA levels

Previous studies show that esiRNAs, many of which map to retroTns, are cleaved from long dsRNA precursors by Dcr-2 (Tomari *et al.* 2007; Ghildiyal *et al.* 2008; Chung *et al.* 2008; Kawamura *et al.* 2008; Siomi *et al.* 2008). Knockdown of Dcr-2 results in decreased esiRNA levels and a corresponding increase in precursor RNAs (Chung *et al.* 2008; Ghildiyal *et al.* 2008). Small RNA (<200 nt) high-throughput sequencing (HTS) libraries were constructed in triplicate from Dcr-2-depleted and control (LacZ) cells (Figure 2.5A). Greater than 99% of reads from all six libraries mapped to the *Drosophila* genome (Table 2.1). Dcr-2 knockdown resulted in a statistically significant ($P = 0.00154$) decrease in nonunique reads (22.7%) compared to the LacZ control (34.4%) (Table 2.1), indicating that Dcr-2 is required for global production of nonuniquely mapping esiRNAs.

Nonunique siRNA-seq reads map across Dm297, blood, mdg1, juan, and jockey for both the LacZ control and the Dcr-2-depleted samples (Figure 2.5, B–F) and esiRNA patterns are generally similar for both the control and the Dcr-2 knockdown. RPM vary considerably among the five retroTns with the most esiRNAs mapping to Dm297 and blood, and fewer mapping to mdg1, juan, and jockey (Figure 2.5, B–F, red numbers). Both Dm297 and mdg1 have a higher concentration of reads mapping to LTRs, as previously described (Chung *et al.* 2008; Ghildiyal *et al.* 2008) (Figure 2.5, B and D). RPM for esiRNAs mapping to all five retro Tns were decreased in Dcr-2-depleted Dmel-2 cells (Figure 2.5, B–F). Average RPM calculated from triplicate sequencing experiments for each retroTn in both control and Dcr-2-depleted samples (Figure 2.5G) were used to determine the ratio of esiRNAs in the LacZ control as compared to the Dcr-2 knockdown (Figure 2.5H). Dcr-2 depletion led to statistically significant reduction of the number of

esiRNAs mapping to Dm297 (2.6-fold), jockey (2.9-fold), mdg1 (2.0-fold), juan (1.7-fold), and blood (1.4-fold) (Figure 2.5H). These data indicate that depletion of Dcr-2 causes a decrease in retroTn-derived esiRNA levels without changing the specific esiRNAs produced. These data strengthen the previous hypothesis that Dcr-2 produces retroTn-derived esiRNAs in Dmel-2 cells.

Figure 2.5 Effects of Dcr-2 Depletion on RetroTns-Derived EsiRNAS.



H

Tn	RPM (LacZ/Dcr-2)	p-value
Dm297	2.57	5.31E-03
blood	1.37	4.72E-02
mdg1	1.98	1.94E-02
juan	1.73	1.34E-02
jockey	2.91	1.25E-02

Fig 2.5 Dcr-2 depletion decreases retroTn-derived esiRNA levels. (A) Representative Western blot of Dcr-2 depletion. The antibody used is indicated to the left of blots and dsRNA for RNAi is labeled above blots. (B–F) Bedgraphs of esiRNAs mapping to retroTns in control and Dcr-2-depleted Dmel-2 cells. The control is red and the Dcr-2 knockdown is in blue. RPM values are listed to the left of bedgraphs and are color coordinated. Relative locations of specific ORFs are shown above the bedgraphs. (G) Graphs of esiRNA levels in control and Dcr-2-depleted Dmel-2 cells (control, red; Dcr-2 knockdown, blue). RPM values are on the *y*-axis and retroTns are indicated along the *x*-axis. Error bars represent standard deviation of technical triplicates. (H) A table reporting the ratio of esiRNA levels (RPM) between the control and Dcr-2 knockdown for each retroTn (middle column) is shown. RetroTn is indicated in the left column and *P*-value (unpaired *t*-test) is indicated in the right column.

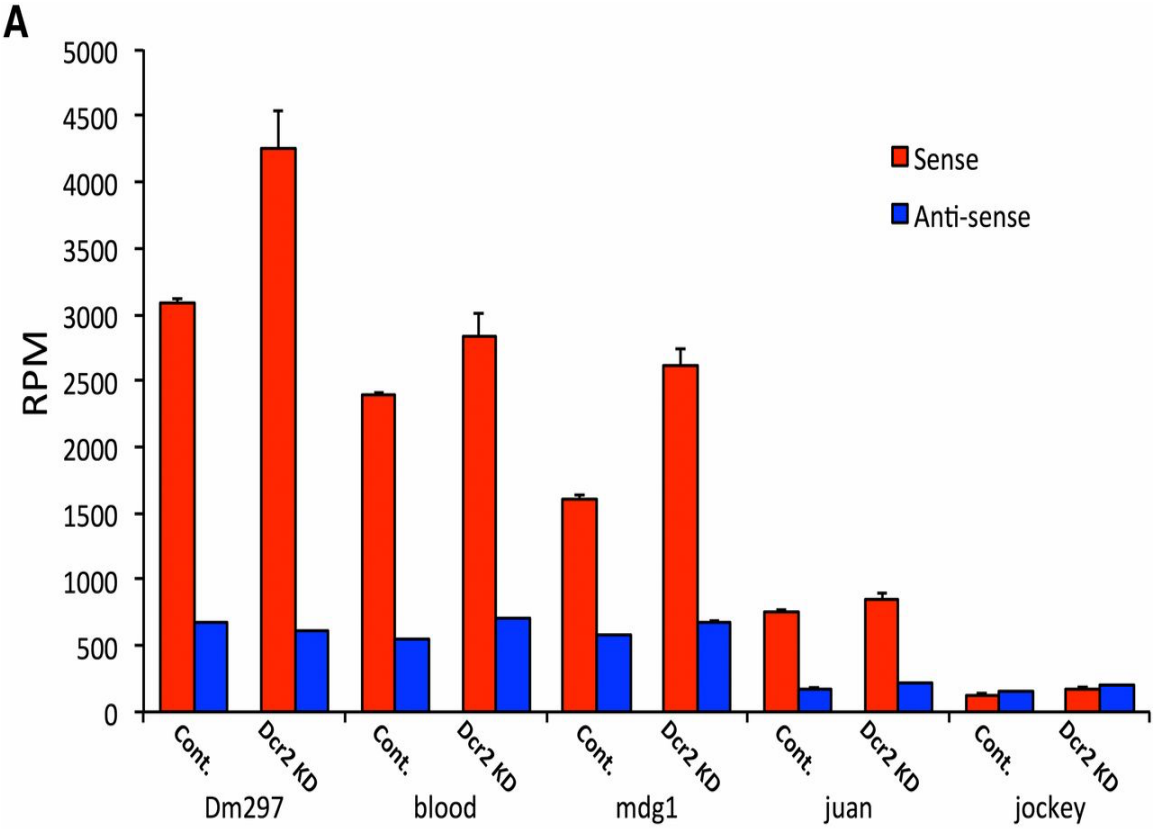
Sense and antisense retroTn transcript levels increase with Dcr-2 knockdown

Previous RT-qPCR studies suggest that some retroTn transcript levels increase after knockdown of Dcr-2 in S2 cells (Chung *et al.* 2008; Ghildiyal *et al.* 2008). To determine S and AS retroTn RNA levels globally, we performed strand-specific RNA-seq on Dcr-2-depleted, large RNA (>200 nt) resulting in an average read depth of 78.3× and ≥98% of reads mapping to the *Drosophila* genome (Table S1). Dcr-2 depletion resulted in a lower read depth as compared to the control (Table 2.1). The percentage of nonuniquely mapping reads was significantly increased in the Dcr-2 knockdown (46.7%) compared to the LacZ knockdown (41.9%) ($P = 9.6 \times 10{-}5$, Table 2.1)

We compared Dcr-2 knockdown and LacZ control RPM for Dm297, blood, mdg1, juan, and jockey (Figure 2.6A). Generally, S retroTn transcript levels were increased in the Dcr-2-depleted samples, as previously reported (Figure 2.6, A, red, and B) (Chung *et al.* 2008; Ghildiyal *et al.* 2008). We observed a similar trend for AS retroTn transcripts except for Dm297, which showed a slight reduction in AS transcript levels following Dcr-2 knockdown (Figure 2.6, A, blue, and B). These results suggest that Dcr-2 generates esiRNAs from dsRNA precursors consisting of S and AS retroTn transcripts

Figure 2.6 Effect of Dcr-2 Depletion on RetroTns Transcript Levels

A



B

Tn	Sense RPM Dcr2/RPM LacZ	Sense p-value	Anti-sense RPM Dcr2/RPM LacZ	Anti-sense p-value
Dm297	1.38	1.76E-02	0.92	3.02E-04
blood	1.18	4.29E-02	1.29	2.39E-04
mdg1	1.63	3.34E-03	1.18	7.11E-04
Juan	1.11	7.04E-02	1.23	2.16E-01
jockey	1.36	8.99E-04	1.35	1.52E-04

Figure 2.6 (A) A graph of retroTn transcript levels upon RNAi depletion of Dcr-2 is shown (sense, red; antisense, blue). RPM values are shown on the *y*-axis and retroTn (control *vs.* Dcr-2 knockdown) is indicated on the *x*-axis. Error bars represent standard deviation of technical triplicates. (B) A table reporting the ratio of retroTn S and AS transcript levels (RPM) between the control and Dcr-2 knockdown for each retroTn is shown. RetroTn is indicated in the first column. *P*-values (unpaired *t*-test) are reported for these comparisons.

DISCUSSION

Understanding the mechanisms that balance retroTn amplification and repression in eukaryotes is critical, as misregulation can lead to detrimental genomic damage. Many retroTns are active in *Drosophila* (Kofler *et al.* 2015), providing a unique opportunity to understand molecular mechanisms of retroTn repression. In *Drosophila* somatic cells, silencing of retroTns requires a dsRNA precursor that is processed into esiRNAs by Dcr-2 (Tomari and Zamore 2005; Ghildiyal *et al.* 2008; Marques *et al.* 2010). To better understand the origin of this retroTn-derived dsRNA precursor, we performed RNA-seq, strand-specific qPCR, and Northern blotting of control Dmel-2 cells. Most Tns produce S and AS transcripts, although S and AS expression are highest for LTR retroTns (Table 2.1). Bioinformatic analysis of representative LTR retroTns Dm297 and blood, a specific mdg1 element (mdg1{}1720), and representative non-LTR retroTns juan and jockey showed S and AS transcripts originating from intraelement transcription start sites for all elements investigated (Figure 2.1 and Figure 2.2). These initiation sites are generally canonical RNAPII transcription start sites with conserved DPEs (Figure 2.3). Collectively, these data suggest that AS retroTn RNAs are convergently transcribed from these start sites. Interestingly, we also observed that AS transcripts derived from retroTns are not strongly polyadenylated (Figure 2.4). By sequencing small RNAs, we determined that esiRNAs are globally derived from locations of retroTn S/AS convergent transcription and that Dcr-2 knockdown decreases esiRNA levels (Figure 2.5). Consistently, we showed that both S and AS retroTn transcript levels increase when Dcr-2 is knocked down (Figure 2.6). Taken together, these data support a model in which AS retroTn transcripts hybridize to their S counterparts forming dsRNAs that are substrates

for esiRNAs production by Dcr-2 (Figure 2.7).

Drosophila retroTns are convergently transcribed from independent, canonical S and AS tss

Unlike in mammals, *Drosophila* RNAPII transcription does not initiate bidirectionally from promoters to generate AS transcripts (Lapidot and Pilpel 2006; Nechaev *et al.* 2010; Core *et al.* 2012). Also, the >100 predicted overlapping *cis*-natural pairs in *Drosophila* are most often complementary ORF 3' UTRs (Okamura *et al.* 2008a), not S transcript-noncoding AS RNA pairs as in other organisms (Pelechano and Steinmetz 2013). Because protein coding genes do not produce AS transcripts in *Drosophila*, mechanisms of AS transcription and downstream regulatory functions have not been fully elucidated. AS retroTn transcription of *Drosophila* telomere LTR retroTns has been observed previously (Danilevskaya *et al.* 1999). Herein, we provide the first evidence of global AS transcription of LTR and non-LTR retroTns, and TIR DNA Tn families in *Drosophila* (Table 2.2, Table 2.4, Table 2.5). Additionally, we identify and quantitate AS transcription of individual Tns and specific elements (Figure 2.1, Figure 2.2, and Table 2.6).

Bioinformatically identified AS transcription start sites and promoter analysis provide the first clues about how retroTn AS transcripts are produced (Figure 2.3). Interestingly, Dm297, blood, and mdg1 AS transcription start sites and promoter elements are located within the retroTn and AS transcripts initiating from these locations could explain all observed AS RNA-seq reads, suggesting that external sequences are not required for LTR retroTn AS transcription. Evidence to support this hypothesis comes from identifying several individual, intergenic LTR

retroTns that are transcribed in the AS direction (Table 2.6). It seems unlikely that multiple intergenic LTR retroTns simultaneously evolved external promoters in different genomic locations and is more plausible that the observed internal Dm297, blood, and mdg1 AS transcription start sites are functional. AS transcription start sites are observed for non-LTR retroTn juan, but these cannot be responsible for transcription of the 5' end of the element (Figure 2.2). The AS transcription start site identified for jockey does not have adequate normalized read counts to account for the amount of AS transcription (Figure 2.2). We hypothesize that the additional AS jockey and 5' juan transcripts originate from intragenic elements oppositely oriented to mRNA S transcripts (Table 2.6). Collectively, these data suggest that LTR retroTns are transcribed from intraelement transcription start sites, while some non-LTR retroTns RNAs are generated indirectly by transcription of protein coding genes.

AS transcripts can arise from bidirectional transcription at RNAPII promoters (Core *et al.* 2008; Guil and Esteller 2012) or convergent transcription from strand-specific promoters (Gullerova and Proudfoot 2012). Bidirectional transcription initiates in both directions from one promoter, while convergent transcription requires independent transcription start sites. Bidirectional transcription of a retroTn from a single promoter would not result in a full-length dsRNA (Figure 2.7). Therefore, our results suggest that double-stranded retroTn RNAs are derived from convergent transcription of S and AS RNAs from independent transcription start sites (Figure 2.7). Transcriptional gene silencing mediated by convergent transcription is highly efficient in both fission yeast and mammalian cells (Gullerova and Proudfoot 2012). Our data

support a model in which formation of dsRNAs by convergent transcription of retroTns is the first step in *Drosophila* somatic cell retroTn silencing.

Production of dsRNAs by convergent transcription is a novel retroTn regulatory mechanism

A well-studied mammalian non-LTR retroTn, L1, initiates AS transcription in the S RNA 5' UTR in humans (Speek 2001; Nigumann *et al.* 2002) and the S RNA ORF1 in mice (Li *et al.* 2014). Most full-length intragenic L1 elements are oriented AS to protein coding genes (Szak *et al.* 2002). Therefore, AS transcription from the identified transcription start sites proceeds into neighboring mRNAs forming fusion transcripts that regulate expression of numerous genes (Speek 2001; Nigumann *et al.* 2002; Mätlik *et al.* 2006; Cruickshanks and Tufarelli 2009) and affect mobility of L1 elements (Li *et al.* 2014). The L1 retroTn is closely related to *Drosophila* non-LTR retroTns jockey and juan (Mizrokhi *et al.* 1988; Speek 2001). No AS transcription start sites were observed in jockey analogous to L1 AS transcription start sites (Figure 2.2C). Juan AS transcription start sites were located in the S RNA ORFI (Figure 2.2C), but only two full-length juan elements (of seven total) are intragenic (Table 2.6) limiting the impact of a potential L1-like AS fusion transcript regulatory mechanism. Additionally, juan AS transcripts were identified upstream of the observed transcription start sites. Together, these data indicate that the mechanism of *Drosophila* non-LTR retroTn AS transcript initiation and the functions of these AS RNAs may differ from their mammalian L1 counterparts.

In fission yeast and the *Drosophila* germline, movement of repetitive sequences and retroTns, respectively, are repressed by a transcriptional gene silencing mechanism wherein siRNAs induce heterochromatin formation (Huisinga and Elgin 2009; Wang and

Elgin 2011; Sienski *et al.* 2012; Huang *et al.* 2013; Le Thomas *et al.* 2013; Rozhkov *et al.* 2013). In *Schizosaccharomyces pombe*, siRNAs are produced from RNA-dependent RNA polymerase generated dsRNA precursors by Dicer 1 (Volpe *et al.* 2002; Yu *et al.* 2014; Holoch and Moazed 2015). In *Drosophila*, retroTns are silenced in the germline by piRNAs cleaved from single-stranded substrates and amplified via a mechanism that does not include long dsRNAs (Saito *et al.* 2006; Aravin *et al.* 2007; Gunawardane *et al.* 2007). Therefore, our proposed model that esiRNAs are generated from hybridized convergently transcribed S and AS retroTn transcripts, is novel as other mechanisms do not require dsRNA substrates, utilize an RNA-dependent RNA polymerase to produce dsRNA substrates, or use AS transcripts to regulate gene expression in a way that does not require siRNAs.

Lack of AS retroTn polyadenylation may lead to nuclear retention of dsRNAs

Efficient polyadenylation of transcripts can promote export to the cytoplasm and removal of polyA signals may cause nuclear retention of RNAs (Dower *et al.* 2004). We propose that convergently transcribed S and AS retroTn transcripts hybridize in the nucleus forming dsRNAs. Because only Dm297 and mdg1 S transcripts are polyadenylated (Figure 2.4), all double-stranded retroTn RNAs investigated would contain at least one polyA− component, encouraging nuclear retention of these dsRNAs. As the number of retroTn AS transcripts is often significantly less than the number of S transcripts (Figure 2.1 and Figure 2.2), unhybridized S RNAs are exported to the cytoplasm for translation, leading to a balance of repression and expansion of retroTns. A nuclear pool of Dcr-2 (Cernilogar *et al.* 2011, and data not shown) may use nuclear-

retained retroTn dsRNA as substrates for esiRNAs biogenesis.

Dcr-2 generates esiRNAs from dsRNAs derived from convergent S and AS transcription of retroTns

LTR and non-LTR retroTns produce esiRNAs from dsRNA precursors through Dcr-2-dependent mechanisms in *Drosophila* somatic cells (Ghildiyal *et al.* 2008; Kawamura *et al.* 2008; Siomi *et al.* 2008). Here, we show that expression of both S and AS retroTn transcripts is regulated by Dcr-2. Specifically, depletion of Dcr-2 leads to reduction in esiRNA levels (Figure 2.5) and a corresponding increase in both S and AS retroTn transcript levels (Figure 2.6). The mechanism of Dcr-2 mediated AS silencing is likely similar to S silencing as S and AS esiRNAs are often equally abundant (Ghildiyal *et al.* 2008).

Others have hypothesized that esiRNAs are processed from double-stranded LTR hairpins because of higher concentrations of small RNA-seq reads from LTRs (Chung *et al.* 2008; Ghildiyal *et al.* 2008). Generally, our data do not support this model. EsiRNA reads from retroTns span the entire element, suggesting that LTR hairpins cannot be the only dsRNA substrates (Figure 2.5). Additionally, S and AS RNA-seq reads map to all regions of retroTns, indicating that the entire element has the potential to form a dsRNA precursor. Also, we observe convergent transcription of non-LTR retroTns and Dcr-2-regulated esiRNAs mapping to these retroTns (Figure 2.5). Thus, an LTR is not required for dsRNA formation and subsequent siRNA biogenesis.

Mechanisms of AS transcription and esiRNA biogenesis are conserved in tissue culture and *Drosophila*

The data presented here were collected in Dmel-2 cells, a derivative of Schneider

2 (S2) cells, a somatic cell line derived from *Drosophila* embryos (Schneider 1972). Previous parallel investigation of esiRNA biogenesis in S2 cells and *Drosophila* heads indicated no mechanistic differences between these two tissues (Ghildiyal *et al.* 2008). Most importantly, esiRNAs were equally derived from S and AS retroTn strands and mapped evenly across retroTn precursors (Ghildiyal *et al.* 2008), indicating that a full-element dsRNA precursor is required to generate the observed esiRNAs in both fly tissues and cell lines. These data are consistent with our results that retroTns in S2 cells produce AS transcripts (Figure 2.1 and Figure 2.2).

Previous studies show that *Drosophila* tissue culture cells have amplified Tn content (Potter *et al.* 1979; Tchurikov *et al.* 1981; Maisonhaute *et al.* 2007; Wen *et al.* 2014) and hypothesize that this amplification is necessary for creating immortal cell lines (Junakovic *et al.* 1988). Once established, Tn location and number appear stable in *Drosophila* Kc and S2 cell lines (Junakovic *et al.* 1988). This amplification is reflected as a greater portion of retroTn-derived esiRNAs mapping to Tns in S2 cells than in *Drosophila* heads (Ghildiyal *et al.* 2008). While having more Tn copies in tissue culture potentially increases the absolute levels of S and AS retroTn transcripts (and esiRNAs generated from dsRNA precursors) the molecular mechanisms required to produce AS transcripts and generate esiRNAs appear conserved between flies and culture cells (Ghildiyal *et al.* 2008). Additionally, a higher concentration of esiRNAs and dsRNA precursors is a tremendous advantage of the S2 cell system.

In conclusion we show, for the first time, that *Drosophila* retroTns are transcribed in the AS direction from intraelement transcription start sites. We observed

convergent transcription of S and AS transcripts in all retroTns investigated, suggesting that this is a global dsRNA formation mechanism in *Drosophila*. The experiments described here will provide the basis for future mechanistic studies of retroTn AS transcription and allow determination of the role of convergent transcription in retroTn gene silencing.

Figure 2.7 Proposed Model of Sense and Anti-Sense Transcription

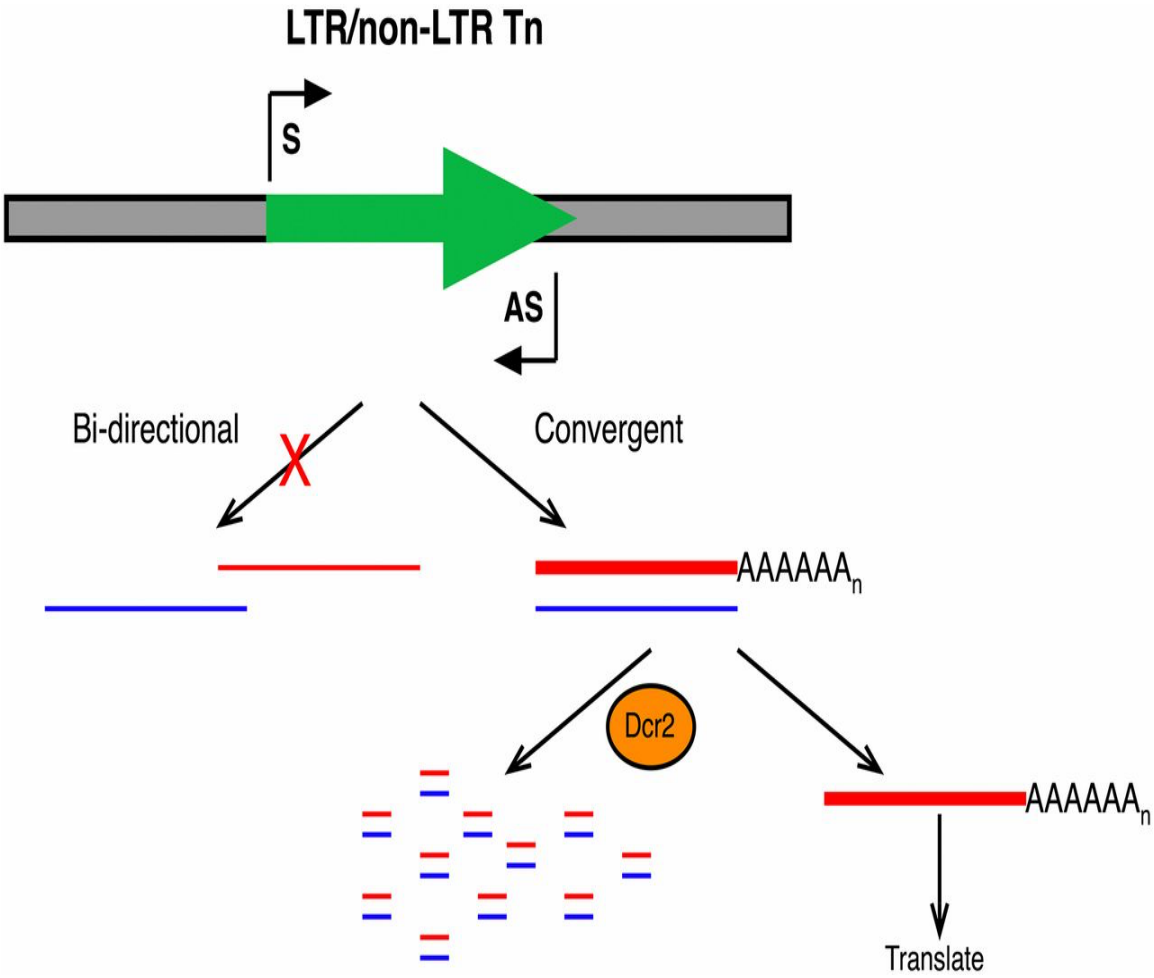


Fig 2.7 Dcr-2 generates esiRNAs from dsRNAs derived from convergent S and AS transcription of retroTns. Shown is a model depicting convergent S and AS transcription (arrows, black, "S" and "AS," respectively) of retroTns (arrow, green) in *Drosophila*. S transcripts (red) are polyadenylated and more abundant (thick line) compared to AS transcripts (blue, thin line). AS transcripts act as a molecular sponge isolating a portion of S transcripts resulting in the formation of dsRNA Dcr-2 substrates. Some S transcripts are translated promoting mobility of retroTns.

REFERENCES

Aravin A. A., Hannon G. J., Brennecke J., 2007 The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science **318**: 761–764.

Arkhipova I. R., Ilyin Y. V., 1991 Properties of promoter regions of mdg1 Drosophila retrotransposon indicate that it belongs to a specific class of promoters. EMBO J **10**: 1169–1177.

Bingham P. M., Kidwell M. G., Rubin G. M., 1982 The molecular basis of P-M hybrid dysgenesis: The role of the P element, a P-strain-specific transposon family. Cell **29**: 995–1004.

Brennecke J., Aravin A. A., Stark A., Dus M., Kellis M., Sachidanandam R., Hannon G. J., 2007 Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell **128**: 1089–1103.

Brouha B., Schustak J., Badge R. M., Lutz-Prigge S., Farley A. H., Moran J. V., Kazazian H. H., 2003 Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci USA **100**: 5280–5285.

Butler J. E. F., Kadonaga J. T., 2002 The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev **16**: 2583–2592.

Cernilogar F. M., Onorati M. C., Kothe G. O., Burroughs A. M., Parsi K. M., Breiling A., Sardo Lo F., Saxena A., Miyoshi K., Siomi H., Siomi M. C., Carninci P., Gilmour D. S., Corona D. F. V., Orlando V., 2011 Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature **480**: 391–395.

Chen L., Dahlstrom J. E., Lee S.-H., Rangasamy D., 2012 Naturally occurring endo- siRNA silences LINE-1 retrotransposons in human cells through DNA methylation. Epigenetics **7**: 758–771.

Chung W.-J., Okamura K., Martin R., Lai E. C., 2008 Endogenous RNA interference provides a somatic defense against Drosophila transposons. Curr. Biol. **18**: 795– 802.

Cordaux R., Batzer M. A., 2009 The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. **10**: 691–703.

Core L. J., Waterfall J. J., Lis J. T., 2008 Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science **322**: 1845–1848.

Core L. J., Waterfall J. J., Gilchrist D. A., Fargo D. C., Kwak H., Adelman K., Lis J. T., 2012 Defining the status of RNA polymerase at promoters. Cell Rep **2**: 1025–1035.

Cruickshanks H. A., Tufarelli C., 2009 Isolation of cancer-specific chimeric transcripts

induced by hypomethylation of the LINE-1 antisense promoter. Genomics **94**: 397–406.

Czech B., Malone C. D., Zhou R., Stark A., Schlingeheyde C., Dus M., Perrimon N., Kellis M., Wohlschlegel J. A., Sachidanandam R., Hannon G. J., Brennecke J., 2008 An endogenous small interfering RNA pathway in Drosophila. Nature **453**: 798–802.

Danilevskaya O. N., Traverse K. L., Hogan N. C., DeBaryshe P. G., Pardue M. L., 1999 The two Drosophila telomeric transposable elements have very different patterns of transcription. Molecular and Cellular Biology **19**: 873–881.

Deininger P., 2011 Alu elements: know the SINEs. Genome Biol. **12**: 236–247.

Dower K., Kuperwasser N., Merrikh H., Rosbash M., 2004 A synthetic A tail rescues yeast nuclear accumulation of a ribozyme-terminated transcript. RNA **10**: 1888– 1899.

Ghildiyal M., Seitz H., Horwich M. D., Li C., Du T., Lee S., Xu J., Kittler E. L. W., Zapp M. L., Weng Z., Zamore P. D., 2008 Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science **320**: 1077–1081.

Gogvadze E., Buzdin A., 2009 Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. **66**: 3727–3742.

Grant G. R., Farkas M. H., Pizarro A. D., Lahens N. F., Schug J., Brunk B. P., Stoeckert C. J., Hogenesch J. B., Pierce E. A., 2011 Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics **27**: 2518–2528.

Guil S., Esteller M., 2012 Cis-acting noncoding RNAs: friends and foes. Nat Struct Mol Biol **19**: 1068–1075.

Gullerova M., Proudfoot N. J., 2012 Convergent transcription induces transcriptional gene silencing in fission yeast and mammalian cells. Nat Struct Mol Biol **19**: 1193–1201.

Gunawardane L. S., Saito K., Nishida K. M., Miyoshi K., Kawamura Y., Nagami T., Siomi H., Siomi M. C., 2007 A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science **315**: 1587–1590.

Henriques T., Gilchrist D. A., Nechaev S., Bern M., Muse G. W., Burkholder A., Fargo D. C., Adelman K., 2013 Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Molecular Cell **52**: 517–528.

Holoch D., Moazed D., 2015 RNA-mediated epigenetic regulation of gene expression. Nat. Rev. Genet. **16**: 71–84.

Huang X. A., Yin H., Sweeney S., Raha D., Snyder M., Lin H., 2013 A major epigenetic

programming mechanism guided by piRNAs. Dev. Cell **24**: 502–516.

Huisinga K. L., Elgin S. C. R., 2009 Biochimica et Biophysica Acta. BBA - Gene Regulatory Mechanisms **1789**: 3–16.

Junakovic N., Di Franco C., Best-Belpomme M., Echalier G., 1988 On the transposition of copia-like nomadic elements in cultured Drosophila cells. Chromosoma **97**: 212– 218.

Kaminker J. S., Bergman C. M., Kronmiller B., Carlson J., Svirskas R., Patel S., Frise E., Wheeler D. A., Lewis S. E., Rubin G. M., Ashburner M., Celniker S. E., 2002 The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. **3**: 1–20.

Kawamura Y., Saito K., Kin T., Ono Y., Asai K., Sunohara T., Okada T. N., Siomi M. C., Siomi H., 2008 Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature **453**: 793–797.

Kent W. J., Sugnet C. W., Furey T. S., Roskin K. M., Pringle T. H., Zahler A. M., Haussler D., 2002 The human genome browser at UCSC. Genome Res. **12**: 996– 1006.

Kidwell M. G., Kidwell J. F., Sved J. A., 1977 Hybrid dysgenesis in Drosophila melanogaster: A syndrome of aberrant traits including mutation, sterility and male recombination. Genetics **86**: 813–833.

Kofler R., Nolte V., Schlötterer C., 2015 Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet **11**: e1005406.

Lapidot M., Pilpel Y., 2006 Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms. EMBO Rep. **7**: 1216-1222.

Le Thomas A., Rogers A. K., Webster A., Marinov G. K., Liao S. E., Perkins E. M., Hur J. K., Aravin A. A., Tóth K. F., 2013 Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev **27**: 390– 399.

Lee M.-C., Marx C. J., 2013 Synchronous waves of failed soft sweeps in the laboratory: remarkably rampant clonal interference of alleles at a single locus. Genetics **193**: 943–952.

Lerat E., Rizzon C., Biémont C., 2003 Sequence divergence within transposable element families in the Drosophila melanogaster genome. Genome Res. **13**: 1889– 1896.

Li J., Kannan M., Trivett A. L., Liao H., Wu X., Akagi K., Symer D. E., 2014 An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. Nucleic Acids Res **42**: 4546–4562.

Maisonhaute C., Ogereau D., Hua-Van A., Capy P., 2007 Amplification of the 1731 LTR

retrotransposon in Drosophila melanogaster cultured cells: origin of neocopies and impact on the genome. Gene **393**: 116–126.

Marques J. T., Kim K., Wu P.-H., Alleyne T. M., Jafari N., Carthew R. W., 2010 Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. Nat Struct Mol Biol **17**: 24–30.

Martin M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet j. **17**: 10.

Mätlik K., Redik K., Speek M., 2006 L1 antisense promoter drives tissue-specific transcription of human genes. J. Biomed. Biotechnol. **2006**: 71753.

McClintock B., 1950 The origin and behavior of mutable loci in maize. Proc Natl Acad Sci USA **36**: 344–355.

Mills R. E., Bennett E. A., Iskow R. C., Devine S. E., 2007 Which transposable elements are active in the human genome? Trends Genet. **23**: 183–191.

Mizrokhi L. J., Georgieva S. G., Ilyin Y. V., 1988 jockey, a mobile Drosophila element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. Cell **54**: 685–691.

Nechaev S., Fargo D. C., Santos dos G., Liu L., Gao Y., Adelman K., 2010 Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science **327**: 335–338.

Nigumann P., Redik K., Mätlik K., Speek M., 2002 Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics **79**: 628–634.

Okamura K., Balla S., Martin R., Liu N., Lai E. C., 2008a Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nat Struct Mol Biol **15**: 581–590.

Okamura K., Chung W.-J., Ruby J. G., Guo H., Bartel D. P., Lai E. C., 2008b The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature **453**: 803–806.

Pelechano V., Steinmetz L. M., 2013 Gene regulation by antisense transcription. Nat. Rev. Genet. **14**: 880–893.

Picard G., Bregliano J. C., Bucheton A., Lavige J. M., Pélisson A., Kidwell M. G., 1978 Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster. Genet. Res. **32**: 275–287.

Potter S. S., Brorein W. J., Dunsmuir P., Rubin G. M., 1979 Transposition of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. Cell **17**: 415–427.

Purcell M. K., Hart S. A., Kurath G., Winton J. R., 2006 Strand-specific, real-time RT- PCR assays for quantification of genomic and positive-sense RNAs of the fish rhabdovirus, Infectious hematopoietic necrosis virus. J. Virol. Methods **132**: 18–24.

Rozhkov N. V., Hammell M., Hannon G. J., 2013 Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev **27**: 400–412.

Rubin G. M., Kidwell M. G., Bingham P. M., 1982 The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. Cell **29**: 987–994.

Saito K., Siomi M. C., 2010 Small RNA-mediated quiescence of transposable elements in animals. Dev. Cell **19**: 687–697.

Saito K., Nishida K. M., Mori T., Kawamura Y., Miyoshi K., Nagami T., Siomi H., Siomi M. C., 2006 Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. Genes Dev **20**: 2214–2222.

Santos dos G., Schroeder A. J., Goodman J. L., Strelets V. B., Crosby M. A., Thurmond J., Emmert D. B., Gelbart W. M., FlyBase Consortium, 2015 FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large- scale migration of genome annotations. Nucleic Acids Res **43**: D690–7.

Schneider I., 1972 Cell lines derived from late embryonic stages of Drosophila melanogaster. J Embryol Exp Morphol **27**: 353–365.

Seitz H., Ghildiyal M., Zamore P. D., 2008 Argonaute loading improves the 5' precision of both MicroRNAs and their miRNA* strands in flies. Curr. Biol. **18**: 147–151.

Sentmanat M. F., Elgin S. C. R., 2012 Ectopic assembly of heterochromatin in Drosophila melanogaster triggered by transposable elements. Proceedings of the National Academy of Sciences **109**: 14104–14109.

Sienski G., Dönertas D., Brennecke J., 2012 Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell **151**: 964–980.

Siomi M. C., Saito K., Siomi H., 2008 How selfish retrotransposons are silenced in Drosophila germline and somatic cells. FEBS Lett. **582**: 2473–2478.

Speek M., 2001 Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Molecular and Cellular Biology **21**: 1973–1985.

Sullivan K. D., Steiniger M., Marzluff W. F., 2009 A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. Molecular Cell **34**: 322–332.

Szak S. T., Pickeral O. K., Makalowski W., Boguski M. S., Landsman D., Boeke J. D., 2002 Molecular archeology of L1 insertions in the human genome. Genome Biol. **3**: 52.

Tchurikov N. A., Ilyin Y. V., Skryabin K. G., Ananiev E. V., Bayev A. A., Krayev A. S., Zelentsova E. S., Kulguskin V. V., Lyubomirskaya N. V., Georgiev G. P., 1981 General properties of mobile dispersed genetic elements in Drosophila melanogaster. Cold Spring Harb. Symp. Quant. Biol. **45 Pt 2**: 655–665.

Tomari Y., Zamore P. D., 2005 Perspective: machines for RNAi. Genes Dev **19**: 517– 529.

Tomari Y., Du T., Zamore P. D., 2007 Sorting of Drosophila small silencing RNAs. Cell **130**: 299–308.

Vagin V. V., Sigova A., Li C., Seitz H., Gvozdev V., Zamore P. D., 2006 A distinct small RNA pathway silences selfish genetic elements in the germline. Science **313**: 320– 324.

Vashist S., Urena L., Goodfellow I., 2012 Development of a strand specific real-time RT-qPCR assay for the detection and quantitation of murine norovirus RNA. J. Virol. Methods **184**: 69–76.

Volpe T. A., Kidner C., Hall I. M., Teng G., Grewal S. I. S., Martienssen R. A., 2002 Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. Science **297**: 1833–1837.

Wang S. H., Elgin S. C. R., 2011 Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proceedings of the National Academy of Sciences **108**: 21164–21169.

Wen J., Mohammed J., Bortolamiol-Becet D., Tsai H., Robine N., Westholm J. O., Ladewig E., Dai Q., Okamura K., Flynt A. S., Zhang D., Andrews J., Cherbas L., Kaufman T. C., Cherbas P., Siepel A., Lai E. C., 2014 Diversity of miRNAs, siRNAs, and piRNAs across 25 Drosophila cell lines. Genome research **24**: 1236-1250.

Xie W., Donohue R. C., Birchler J. A., 2013 Quantitatively increased somatic transposition of transposable elements in Drosophila strains compromised for RNAi. PLoS ONE **8**: e72163.

Yang N., Kazazian H. H., 2006 L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol **13**: 763–771.

Yu R., Jih G., Iglesias N., Moazed D., 2014 Determinants of heterochromatic siRNA

biogenesis and function. Molecular Cell **53**: 262–276.

BIOINFORMATIC ANALYSIS OF SENSE AND ANTISENSE EXPRESSION FROM TERMINAL INVERTED REPEAT TRANSPOSONS IN *DROSOPHILA* SOMATIC CELLS

CONTRIBUTION

I am First Author on the work described below, which has been published in the Journal *Fly*, March, 2016. All experiments detailed were performed by myself. Analysis of sequence data and preparation of the manuscript was performed by Dr. Steiniger and myself.

SUMMARY

Understanding regulation of transposon movement in somatic cells is important as mobile elements can cause detrimental genomic rearrangements. Generally, transposons move via one of two mechanisms; retrotransposons utilize an RNA intermediate, therefore copying themselves and amplifying throughout the genome, while terminal inverted repeat transposons (TIR Tns) excise DNA sequences from the genome and integrate into a new location. Our recently published work indicates that retrotransposons in *Drosophila* tissue culture cells are actively transcribed in the antisense direction. Our data support a model in which convergent transcription of retrotransposons from intra element transcription start sites results in complementary RNAs that hybridize to form substrates for Dicer-2, the endogenous small interfering (esi)RNA generating enzyme. Here, we extend our previous analysis to TIR Tns. In contrast to retrotransposons, our data show that antisense TIR Tn RNAs result from transcription of intronic TIR Tns oriented antisense to their host genes. Also, disproportionately less esiRNAs are generated from TIR transcripts than from

retrotransposons and transcription of very few individual TIR Tns could be confirmed. Collectively, these data support a model in which TIR Tns are regulated at the level of Transposase production while retrotransposons are regulated with esiRNA post-transcriptional mechanisms in *Drosophila* somatic cells.

INTRODUCTION

Active transposons (Tns) and transposon derived sequences comprise approximately 22% of the *Drosophila melanogaster* genome.[1-3] Movement of these Tns plays an important role in evolution, but also causes genomic instability;[4] therefore, regulation of Tn expansion is important to maintain an appropriate balance. In *Drosophila*, mutations linked to P element insertions cause hybrid dysgenesis syndrome.[5-8] Intensive study of P elements has contributed to a molecular understanding of class II terminal inverted repeat (TIR) transposition mechanisms[9,10] and how Tn movement is regulated *in vivo*.[11-14]

Transposons are classified based on the identity of their nucleic acid intermediates. Transposons having an RNA intermediate and encoding a reverse transcriptase are retrotransposons (retroTn). Tns utilizing a cut-and-paste mechanism with a DNA intermediate and having inverted repeat end sequences are called terminal inverted repeat (TIR) Tns. *Drosophila melanogaster* has several classes of both retroTns and TIR Tns,[1] although retroTns appear to be more active.[3,15]

Movement of both retroTns and TIR Tns must be regulated to ensure genomic stability. In *Drosophila*, two non-coding RNA mediated post-transcriptional silencing mechanisms have been elucidated. The piwi-interacting RNA (piRNA) pathway generates

small RNAs that suppress Tn mobility by inducing heterochromatin formation in the germline.[16-22] These siRNAs are produced from a single stranded RNA precursor. In somatic cells, endogenous small interfering (esi)RNAs silence retroTns using an Argonaute 2 (Ago2)-dependent mechanism.[23-27] Most of these esiRNAs are generated from double stranded Tn derived RNA precursors by Dicer-2 (Dcr2).[23,28,29] Increased movement of retroTns is observed in somatic tissues from *Drosophila* mutants lacking RNAi components.[27] How these silencing mechanisms function in the fly to suppress Tn movement are poorly understood. One transcriptional regulatory mechanism has been identified for P element transposition. Alternative splicing prevents expression of the P element Transposase (Tnp) in somatic cells.[12] Therefore, P elements are only mobile in the *Drosophila* germline as functional P element Tnp is only produced in this tissue.[11,12]

To understand the origins of Dcr2 substrates in *Drosophila* tissue culture cells,[30] we performed small RNA-seq and RNA-seq on wild type and Dcr2 depleted samples. Our analyses of these data revealed that many individual retroTns are transcribed in both the sense (S) and antisense (AS) direction from intra element transcription start sites with canonical *Drosophila* RNA polymerase II promoters. These S and AS RNAs are substrates for Dcr2 as their levels are increased in the Dcr2 depleted sample. Correspondingly, the number esiRNAs generated from these substrates decreases when Dcr2 is knocked down. This work was recently published in *Genetics*.[15]

Here we extend this in-depth analysis to include TIR Tns *pogo* and 1360 (*hoppel* or *ProtoP*). The *pogo* TIR Tn is a member of the Tc1/*mariner* superfamily of Tns with 21 base pair terminal inverted repeats.[31,32] 1360 is believed to be derived from an ancient P

element-like Tn.[33] These elements were chosen for further investigation as S and AS transcription of these Tns was observed previously.[15] We conclude that while a few AS RNA-seq reads are observed for these TIR Tns, very few esiRNAs are generated from these transcripts indicating that TIR Tn movement is minimally regulated by esiRNAs in *Drosophila melanogaster* somatic cells. Additionally, we discovered that unlike retroTns, few individual TIR Tns are transcribed in either the S or AS direction. Finally, analyses of 1360 and *pogo* transcripts allow insight into the transposition mechanisms of these TIR Tns.

RESULTS

Ratios of full-length to truncated Tns differ for TIR and retroTns

1360 or *Hoppel* is the most abundant TIR Tn in the *Drosophila* genome with 304 annotated copies,[34,35] while 48 *pogo* TIR Tns have been documented.[35] A full-length 1360 element is predicted to be 1107 base pairs (bp) while a full-length *pogo* element is 2122 bp. As a first step towards insight into the molecular details of TIR Tn mobility in a genomic context, variation in sizes among the annotated 1360 and *pogo* Tns were first examined. Approximately one-third of 1360 Tns are greater than 1 kb (Table 2.8), although only three are the predicted 1107 bp (data not show). Generally, the 1360 elements vary tremendously in size; often differing by only 1 bp in the 1360 Tns less than 40 bp and by only 10s of bp for the 1360 Tns greater than 40 bp (data not shown). Rarely are two 1360 elements the same size. In contrast, *pogo* elements are restricted to four sizes: 2120-2123 bp, 1067-1491 bp, 704 bp or 186-187 bp. Examination of length distributions for previously investigated non-LTR retrotransposons (retroTns) Juan and

Jockey and LTR retroTns blood, mdg1 and 297 revealed a much higher percentage of full-length elements than observed for *pogo* and 1360 (data not shown).[15] The variability of length distributions may support differences in mechanisms that control retroTn and TIR Tn movement.

AS TIR Tn transcripts are not produced from intraelement tss

To investigate S and AS *pogo* and 1360 Tn transcription and potential esiRNA biogenesis, small RNA-seq and RNA-seq data sets from control *Drosophila* tissue culture (Dmel-2) cells[15] were mapped to the *Drosophila* genome followed by visualization of non-unique and unique reads using the UCSC genome browser (http://genome.ucsc.edu, Dm6 assembly, August 2014).[36,37] Examples of representative full-length intergenic *pogo* and 1360 elements show RNA-seq reads mapping to both S and AS strands, although the number of normalized AS reads (reads per million (RPM)) is low (Figures 2.8A-B, red).

We also re-mapped publically available short-capped RNA sequencing data to identify potential S and AS *pogo* and 1360 transcription start sites (tss).[38,39] No intra element tss accounting for S or AS transcripts could be identified for canonical *pogo* elements (Figure 2.8A, blue) indicating that the non-unique reads mapping to this intergenic element originated from a different pogo TIR Tn. While intra element tss are present in canonical 1360 elements, these tss initiate transcription into flanking sequences rather than producing S and AS 1360 TIR Tn RNAs (Figure 2.8B, blue). Like *pogo*, these data show that the observed non-unique reads mapping to this representative intergenic 1360 TIR Tn were produced by a different individual Tn.

If the S and AS TIR Tn RNA-seq reads are not produced from intra element tss, what could be the source of these transcripts? To further investigate the origins of low level AS TIR Tn transcription, we first examined Tn flanking sequences for all elements. No AS tss were observed for *pogo* and 1360 TIR Tns in regions surrounding the Tns (data not shown). Next we investigated the genomic locations of *pogo* and 1360 TIR Tns larger than 1 kb. About half of the one hundred and twelve 1360 elements larger than 1 kb are between genes (intergenic) and half are within introns of protein coding genes (intragenic) (Table 2.8). ~70% of *pogo* TIR Tns greater than 1 kb are intergenic while the other ~30% are intragenic (Table 2.8). Further examination of intragenic TIR Tns greater than 1 kb revealed that the orientation of the Tn to the mRNA was AS ~60% of the time for both *pogo* and 1360 meaning that transcription of the protein coding gene would produce AS TIR Tn RNAs for 60% of the TIR Tns (Table 2.8) greater than 1 kb. As the AS RPMs for both *pogo* and 1360 are low (Figures 2.8A-B), we hypothesize that these RNAs are generated indirectly from transcription of protein coding genes with AS oriented intragenic TIR Tns.

Previous experiments show that endogenous small interfering (esi)RNAs are produced from hybridized (double stranded (ds)) retroTn S and AS transcripts by Dicer-2 (Dcr2).[15] To investigate the potential for S and AS 1360 and *pogo* transcripts to generate esiRNAs, we visualized smRNA-seq reads from control Dmel-2 cells corresponding to representative 1360 and *pogo* TIR Tns (Figures 2.8A-B, bottom). Very few esiRNAs were observed for either *pogo* or 1360.

Table 2.8. Analysis of size classes of 1360 and *pogo* TIR Tns

Element	Size	% total	% >1kb inter	% >1kb intra	% >1kb intra (S)	% >1kb intra (AS)	%trans. int. tss
1360	>1kb	36.8	47.4	52.6	42.5	57.5	1.3
	1kb-40bp	52.0					
	<40bp	11.2					
pogo	>2kb	10.4					
	1-1.5kb	23.0	68.7	31.3	40.0	60.0	0.0
	704bp	4.2					
	187-186bp	58.3					
	186-40bp	4.2					

The percent (**% total**) of 1360 and *pogo* TIR Tns in each **size** class is shown in the left three columns. The remainder of the analysis was only performed on Tns greater than 1 kb. The percentages of intergenic and intragenic 1360 and *pogo* Tns (**%>1kb inter** and **%>1kb intra**) are shown in columns three and four. The percent of intragenic 1360 and *pogo* Tns having mapped sense (S) and antisense (AS) RNA-seq reads (**%>1kb intra (S)** and **%>1kb intra (AS)**) are shown in columns five and six. Finally, the percent of 1360 and *pogo* TIR Tns greater than 1 kb for which S or AS transcription from an internal transcription start site (tss) could be confirmed (**%trans. int. tss**) is reported in the last column.

Fig 2.8 Sense and Antisense Bedgraphs of 1360 and *pogo* TIR Tns

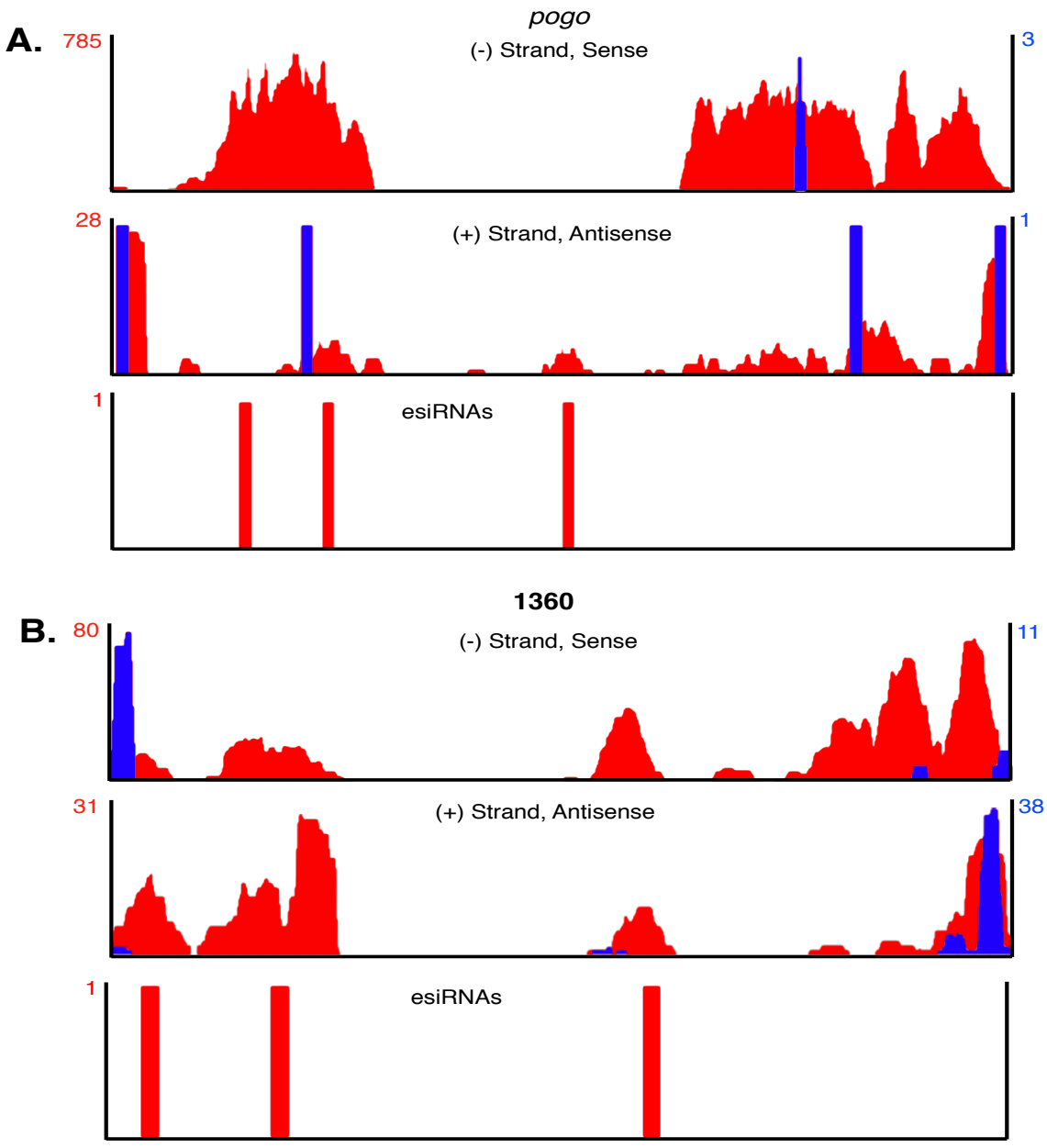


Fig. 2.8 Sense and antisense 1360 and pogo TIR Tn transcripts are not produced by intra element transcription start sites. (A-B) Bedgraphs representing sense (top) and antisense (middle) non-unique RNA-seq reads mapping to a representative full-length pogo TIR Tn (A) or 1360 TIR Tn (B) are shown in red. Peak reads per million (RPM) are listed to the left (red numbers). Non-unique small-capped RNA-seq reads representing transcription start sites are overlaid in blue and RPM values are listed to the right (blue numbers). Non- unique endogenous small-RNA seq reads mapping to pogo and 1360 TIR Tns are shown below the RNA-seq reads (A-B).

Pogo{}4759 is the only actively transcribed pogo Tn in the Drosophila genome

Drosophila pogo elements fall into four size classes: 2.1 kb, 1.1-1.4 kb, 704 bp and 186-187 bp. Previous analyses indicate that single internal deletions are responsible for the 2.1 kb to 1.1-1.4 kb size reduction resulting in Tns with similar ends, but differing internal structure.[31] This observation is confirmed by aligning all annotated *Drosophila pogo* elements (data not shown). Comparing *pogo* TIR Tns representative of the 2.1 kb (pogo{}297), 1.1-1.4 kb (pogo{}4759), 704 bp (pogo{}1454), and 186-187 bp (pogo{}718) size classes reveals deletion of pogo{}297 nucleotides (nts) 788 to 1440 in pogo{}4759, deletion of an additional ~300 nts flanking the pogo{}4759 internal deletion in pogo{}1454, and loss of all but 93 nts at each end of the Tn in pogo{}718 (Figure 2.9A, left).

Visualization of RNA-seq reads corresponding to pogo{}297, pogo{}4759, and pogo{}1454 show no non-unique or unique reads mapping to pogo{}297 nts 788 to 1440 (Figure 2.9A, right), nor is there evidence that these sequences have been removed by splicing (data not shown). In contrast, non-unique RNA-seq reads map the entire length of pogo{}4759 and pogo{}1454 (Figure 2.9A, right). Additionally, unique reads corresponding to a splice junction are clearly evident for pogo{}4759 and a strong tss (13,114 RPM) is present just upstream of the intergenic pogo{}4759 element (Figure 2.9B). These data, together with a lack of observed intra element S *pogo* tss (Figure 2.8A), support a model in which none of the five 2.1 kb *Drosophila* pogo elements are transcribed, but that active transcription of pogo{}4759 accounts for all non-unique reads mapping to *pogo* transposons.

Fig 2.9 Bedgraphs Representing the *pogo* TIR Tns

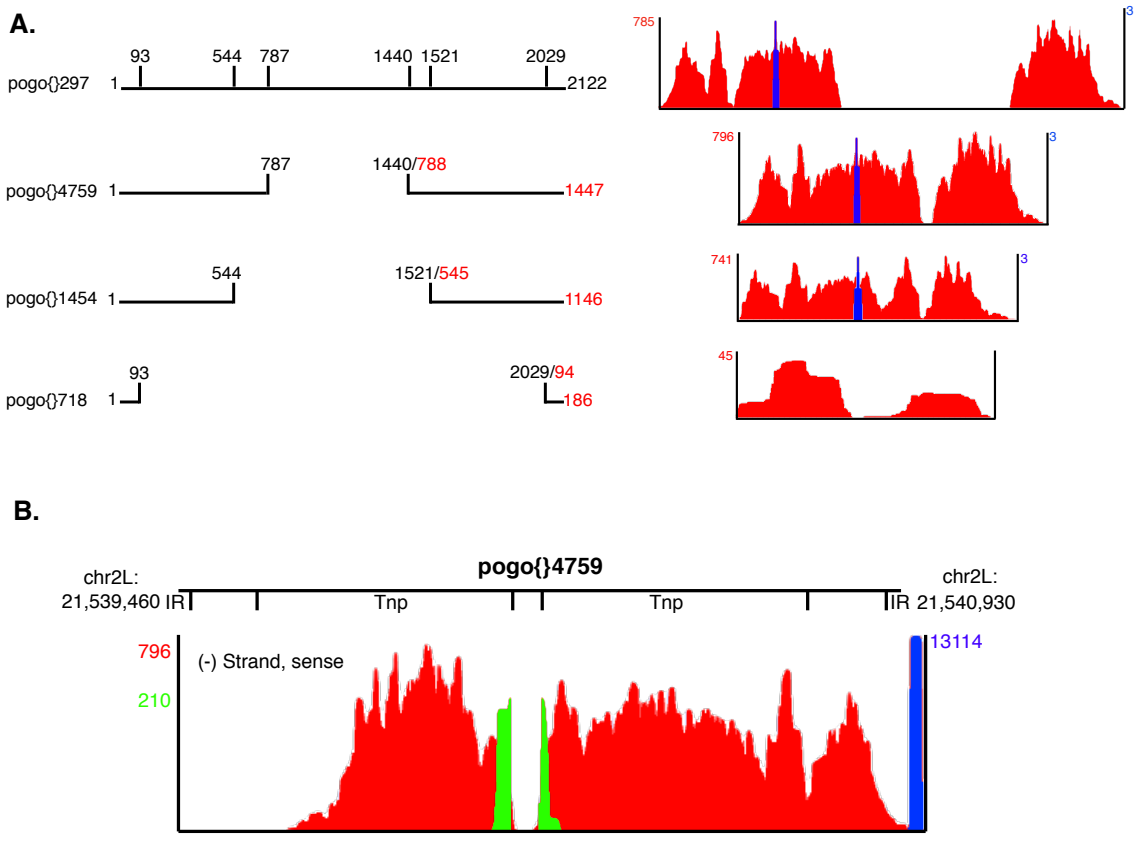


Fig 2.9 Pogo{}4759 is the only transcribed pogo element in the Drosophila genome. (A) Pogo TIR Tns representing the 4 different size classes of pogo elements are shown. Nucleotide deletion positions are labeled above each schematic. To the right of each Tn, non- unique RNA-seq (red) and small-capped RNA-seq (blue) reads mapping to each pogo TIR Tn are displayed. (B) Non-unique RNA-seq reads (red), unique RNA-seq reads (green) and small-capped RNA-seq (blue) reads mapping to pogo{}4759 are shown with maximum normalized RPM displayed in corresponding colors. Relative locations of specific ORFs are shown above the bedgraphs with the chromosomal location of pogo{}4759.

EsiRNAs are generated from 1360 TIR Tns

Canonical 1360 TIR Tns produce very few esiRNAs although AS transcripts are evident that could potentially hybridize with S RNAs (Figure 2.8B). Upon further investigation we identified a few 1360 elements with low abundance intra element S and AS tss near the 3' end of the Tn; the presence of these tss correlates with increased transcription of this 1360 region. An example (1360{}1539) is shown in Figure 2.10A. Visualization of smRNA-seq reads corresponding to these sequences shows that esiRNAs are generated from these 1360 RNAs, albeit at low frequency (Figure 2.10A, bottom).

From these data, we conclude that the number of esiRNAs produced from TIR Tn dsRNA precursors is dramatically less than the expression level of S and AS TIR Tn transcripts. Normalized RNA-seq read counts for S and AS 1360{}1539 are ~170 while esiRNAs RPMs are 4 (Figure 2.10A). Because expression of RNAs from canonical 1360 elements is reduced compared to 1360{}1539 (Figure 2.8B), any esiRNAs produced from hybridized S and AS transcripts would be below the limit of detection of our assay.

Transcription from 1360 intra element tss creates fusion RNAs with neighboring sequences

As discussed previously, intra element tss were identified in canonical 1360 elements, but these tss initiate transcription into flanking sequences instead of towards the TIR Tn (Figure 2.8B). Further investigation revealed a few individual 1360 elements for which unique RNA-seq reads corresponding to sequences immediately surrounding the Tn could be identified. An example is shown in Figure 2.10B (1360{}1514). 1360{}1514 is an intergenic TIR Tn and therefore is not indirectly transcribed as a consequence of being in an intron. 1360{}1514 has two tss on the (-) strand that clearly

overlap non-unique RNA-seq reads at the 5' end of the element (Figure 2.10B). Further examination reveals unique RNA-seq reads (10 RPM) mapping to the 1360{}1514/flanking sequence indicating that transcription from the observed tss continues beyond the Tn into neighboring sequences (Figure 2.10B). To our knowledge, TIR Tn fusion transcripts have not been observed previously.

1360{}1533 may encode a P-element-like Transposase

Previous investigations defined an ancestral *Drosophila* Tn termed *ProtoP* from which 1360 elements derive.[33] The consensus sequence of this 4480 bp TIR Tn encodes an 864 amino acid Transposase (Tnp) with homology to the modern P-element Tnp.[33] Our examinations of the 304 annotated 1360 elements in the *Drosophila* genome yielded one element that might produce a functional Tnp. 1360{}1533, a 7854 bp 1360 element, is transcribed from the (-) strand and has multiple ORFs in all three reading frames (Figure 2.10C). The first reading frame encodes an ORF near the 3' end with homology to the P-element Tnp (582 amino acids) while the third reading frame encodes an ORF with an retroTn RNase H domain and separate homology to retroviral integrases (641 amino acids) (Figure 2.10C). Additionally, two intra element tss were identified in 1360{}1533 that would allow transcription of the proposed Tnp (Figure 2.10C). Unfortunately, all short-capped RNA-seq (defining tss) and RNA-seq reads mapping to 1360{}1533 were non-unique. Therefore, transcription of this specific 1360 element could not be confirmed bioinformatically.

Fig 2.10 Bedgraphs and TSS of 1360 TIR Tns

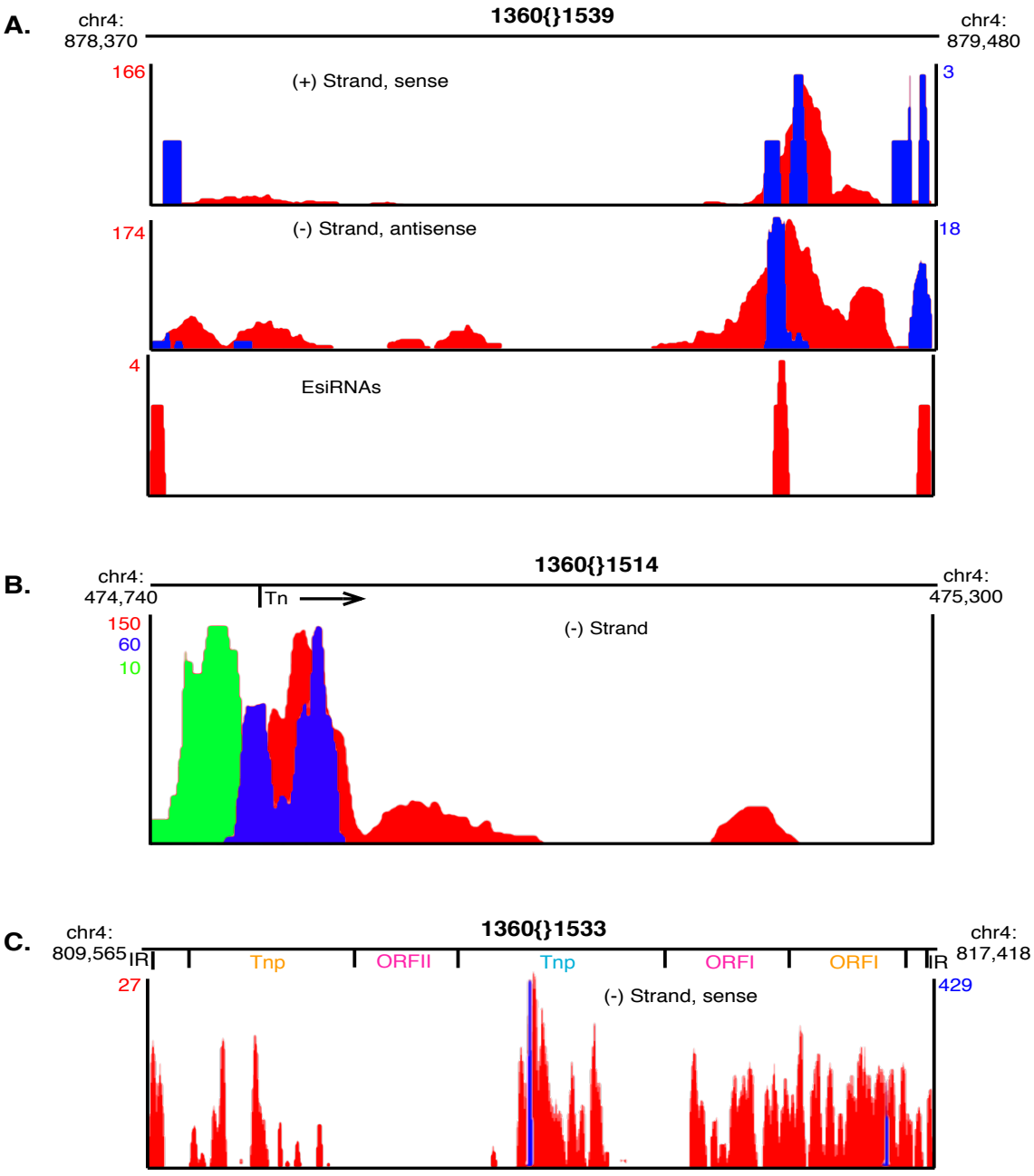


Fig 2.10 Diverse 1360 TIR Tns produce potential regulatory RNAs. (A-C) Non-unique RNA-seq reads (red), unique RNA-seq reads (green) and small-capped RNA-seq (blue) reads mapping to 1360{}1539 (A), 1360{}1514 (B) or 1360{}1533 (C) are shown with maximum normalized RPM displayed in corresponding colors. Chromosomal locations of each TIR Tn are shown above the bedgraphs. Relative locations of specific ORFs are shown for 1360{}1533.

DISCUSION

RetroTns and TIR Tns are differentially regulated

Recently, we published data supporting a model in which retroTns in *Drosophila* somatic cells are regulated by esiRNAs.[15] RetroTns are convergently transcribed in the sense and antisense direction primarily from intra element transcription start sites (Figure 2.11A).[15] Many full-length retroTns are present in the *Drosophila* genome and transcription of individual elements was confirmed for a large percentage of the elements investigated.[15] The sense and antisense transcripts produced from retroTns have the potential to hybridize, creating double stranded RNAs that are substrates for esiRNA biogenesis by Dcr2 (Figure 2.11A).[15] EsiRNAs restrict retroTn movement in *Drosophila* somatic cells by an unknown mechanism requiring RNAi factors.[27] The amount of Dcr2 precursor is determined by the expression level of the least transcribed retroTn strand. As the amount of antisense transcript is usually less, we proposed that the excess sense strand would be translated, providing proteins required for retroTn mobility (Figure 2.11A). Therefore, our model indicates that the potential for retroTn amplification is defined by the balance between inhibition by esiRNAs and translation of proteins required for retroTn mobility.

The analyses described here support a very different mechanism to limit TIR Tn movement in *Drosophila* somatic cells. Antisense TIR Tn RNAs are produced indirectly from intronic elements oriented antisense to sense mRNAs (Table 2.8, Figure 2.8). Because expression of these transcripts is considerably lower than for retroTns (Figures 2.8A-B), the potential for formation of double stranded RNA Dcr2 precursor is greatly reduced (Figure 2.11B). Additionally, the number of esiRNAs produced from potential

TIR Tn double stranded RNAs is proportionately less than was observed for retroTns (Figures 2.10A and 2.11B).[15] Therefore, many less total esiRNAs are generated from TIR Tns than from retroTns. The potential for production of protein(s) required for Tn movement is also dramatically different for TIR Tns. The number of full-length, actively transcribed TIR Tns in the *Drosophila* genome, is much lower than the number of full-length, actively transcribed retroTns (Table 2.8).[15] Potential Tnp ORFs could only be identified for one *pogo* TIR Tn (Figure 2.9B) and one 1360 element (Figure 2.10C). We hypothesize that lower functional TIR Tn Tnp copy number reduces active Tnp concentration.

Collectively, these data support post-transcriptional retroTn regulation, while TIR Tns are inhibited at the transcriptional level. In these models, the potential for retroTn movement is higher because retroTn sense transcript is more highly expressed. To balance this, large numbers of esiRNAs produced from hybridized sense and antisense retroTn RNAs inhibit retroTn mobility. In contrast, because less TIR Tn Tnp transcript is produced, no esiRNAs are required to inhibit TIR Tn movement. Future experiments will be required to test these hypotheses.

TIR Tn 1360 produces fusion transcripts

We observed sequencing reads mapping to junctions between 1360 TIR Tns and flanking sequences indicating that transcription initiates within the 5 prime end of the Tn and continues into neighboring sequences producing hybrid TIR Tn/flanking RNAs. To our knowledge, these fusion Tn/flanking sequence RNAs have not been reported previously in *Drosophila*. Interestingly, the mammalian LINE-1 retroTn produces a similar

fusion transcript by initiating AS transcription near the 5' end of the element.[40-42] LINE-1 elements are often intergenic and oriented AS to their host genes,[43] therefore, AS transcription results in LINE-1 RNA/mRNA fusion transcripts known to regulate expression of many genes.[40,41,44,45] Further investigation is required to determine if intragenic 1360 TIR Tns could regulate gene expression using a similar mechanism in *Drosophila*.

Fig 2.11 Models Depicting Tn Regulation in *Drosophila* S2 Cells

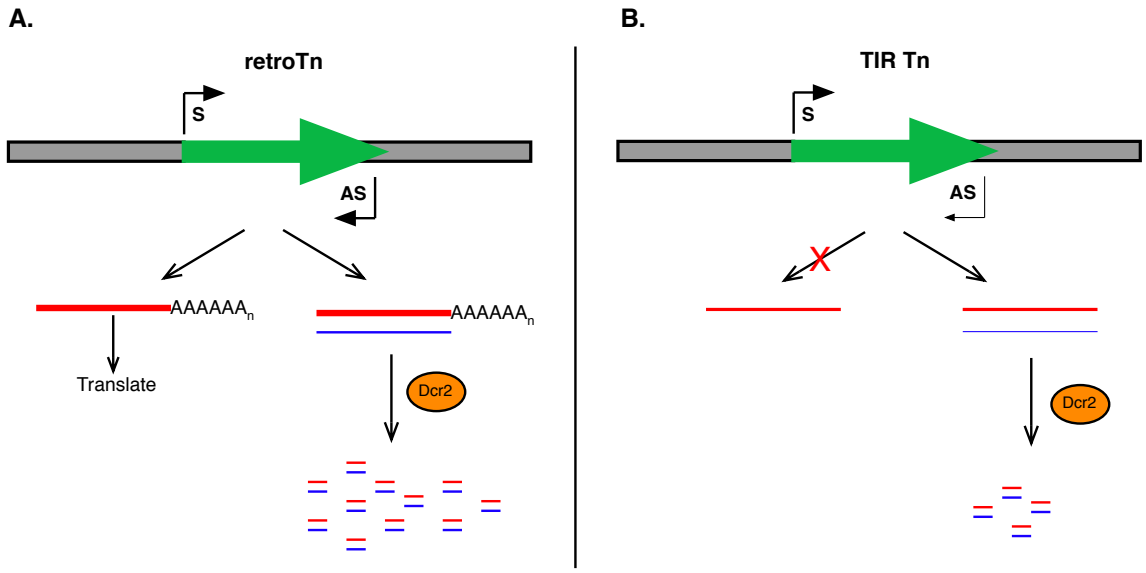


Fig 2.11. Models depicting Tn regulation in Drosophila somatic cells. (A) RetroTns (green arrow) produce both sense (S, red) and anti- sense (AS, blue) transcripts by convergent transcription. Hybridization of these RNAs creates a double stranded RNA substrate for bio- genesis of endogenous small interfering (esi)RNAs by Dcr2. These esiRNAs repress Tn movement via an unknown mechanism. The retroTn transcript is also translated providing proteins required for Tn movement and balancing Tn repression by esiRNAs. (B) TIR Tns also produce both S and AS transcripts, but the amount of AS transcript is~4-fold lower than the lowest expressed retroTn transcript investigated (thin blue line). Additionally, the number of esiRNAs produced from potential TIR Tn dsRNA substrates is dramatically less than for retroTns. While these mechanisms lead to limitations in repressing TIR Tn via the esiRNAs pathway, inhibition is less necessary as S transcription of TIR Tns Tnps is severely restricted.

References

1. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol 2002; 3:1–20.

2. Lerat E, Rizzon C, Biémont C. Sequence divergence within transposable element families in the Drosophila melanogaster genome. Genome Res 2003; 13:1889–96.

3. Kofler R, Nolte V, Schlötterer C. Tempo and Mode of Transposable Element Activity in Drosophila. PLoS Genet 2015; 11:e1005406.

4. Kazazian HH. Mobile elements: drivers of genome evolution. Science 2004; 303:1626–32.

5. Kidwell MG, Kidwell JF, Sved JA. Hybrid dysgenesis in Drosophila melanogaster: A syndrome of aberrant traits including mutation, sterility and male recombination. Genetics 1977; 86:813–833.

6. Picard G, Bregliano JC, Bucheton A, Lavige JM, Pélisson A, Kidwell MG. Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster. Genet Res 1978; 32:275–87.

7. Rubin GM, Kidwell MG, Bingham PM. The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. Cell 1982; 29:987–94.

8. Bingham PM, Kidwell MG, Rubin GM. The molecular basis of P-M hybrid dysgenesis: The role of the P element, a P-strain-specific transposon family. Cell 1982; 29:995–1004.

9. Kaufman PD, Rio DC. P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell 1992; 69:27–39.

10. Tang M, Cecconi C, Bustamante C, Rio DC. Analysis of P element transposase protein-DNA interactions during the early stages of transposition. J Biol Chem 2007; 282:29002–12.

11. Rio DC, Laski FA, Rubin GM. Identification and immunochemical analysis of biologically active Drosophila P element transposase. Cell 1986; 44:21–32.

12. Laski FA, Rio DC, Rubin GM. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. Cell 1986; 44:7–19.

13. Spradling AC, Bellen HJ, Hoskins RA. Drosophila P elements preferentially transpose to replication origins. Proceedings of the National Academy of Sciences 2011; 108:15948–53.

14. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 2008; 322:1387–92.

15. Russo J, Harrington AW, Steiniger M. Antisense Transcription of Retrotransposons in Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. Genetics 2016; 202:107–21.

16. Vagin VV, Sigova A, Li C, Seitz H, Gvozdev V, Zamore PD. A distinct small RNA pathway silences selfish genetic elements in the germline. Science 2006; 313:320–4.

17. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 2007; 128:1089–103.

18. Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science 2007; 318:761–4.

19. Sentmanat MF, Elgin SCR. Ectopic assembly of heterochromatin in Drosophila melanogaster triggered by transposable elements. Proceedings of the National Academy of Sciences 2012; 109:14104–9.

20. Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, Perkins EM, Hur JK, Aravin AA, Tóth KF. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 2013; 27:390–9.

21. Gu T, Elgin SCR. Maternal Depletion of Piwi, a Component of the RNAi System, Impacts Heterochromatin Formation in Drosophila. PLoS Genet 2013; 9:e1003780.

22. Haynes KA, Caudy AA, Collins L, Elgin SCR. Element 1360 and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. Curr Biol 2006; 16:2222–7.

23. Ghildiyal M, Seitz H, Horwich MD, Li C, Du T, Lee S, Xu J, Kittler ELW, Zapp ML, Weng Z, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 2008; 320:1077–81.

24. Chung W-J, Okamura K, Martin R, Lai EC. Endogenous RNA interference provides a somatic defense against Drosophila transposons. Curr Biol 2008; 18:795–802.

25. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R, et al. An endogenous small interfering RNA pathway in Drosophila. Nature 2008; 453:798–802.

26. Saito K, Siomi MC. Small RNA-mediated quiescence of transposable elements in

animals. Dev Cell 2010; 19:687–97.

27. Xie W, Donohue RC, Birchler JA. Quantitatively increased somatic transposition of transposable elements in Drosophila strains compromised for RNAi. PLoS ONE 2013; 8:e72163.

28. Tomari Y, Du T, Zamore PD. Sorting of Drosophila small silencing RNAs. Cell 2007; 130:299–308.

29. Marques JT, Kim K, Wu P-H, Alleyne TM, Jafari N, Carthew RW. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. Nat Struct Mol Biol 2010; 17:24–30.

30. Schneider I. Cell lines derived from late embryonic stages of Drosophila melanogaster. J Embryol Exp Morphol 1972; 27:353–65.

31. Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K. The pogo transposable element family of Drosophila melanogaster. Mol Gen Genet 1992; 232:126–34.

32. Plasterk RH, Izsvák Z, Ivics Z. Resident aliens: the Tc1/mariner superfamily of transposable elements. Trends Genet 1999; 15:326–32.

33. Kapitonov VV, Jurka J. Molecular paleontology of transposable elements in the Drosophila melanogaster genome. Proc Natl Acad Sci USA 2003; 100:6569–74.

34. Bartolomé C, Maside X, Charlesworth B. On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster. Mol Biol Evol 2002; 19:926–37.

35. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. Unlocking the secrets of the genome. Nature 2009; 459:927–30.

36. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res 2002; 12:996–1006.

37. Santos dos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM, FlyBase Consortium. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res 2015; 43:D690–7.

38. Nechaev S, Fargo DC, Santos dos G, Liu L, Gao Y, Adelman K. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science 2010; 327:335–8.

39. Henriques T, Gilchrist DA, Nechaev S, Bern M, Muse GW, Burkholder A, Fargo DC,

Adelman K. Stable pausing by RNA polymerase II provides an opportunity to target and integrate regulatory signals. Molecular Cell 2013; 52:517–28.

40. Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Molecular and Cellular Biology 2001; 21:1973–85.

41. Nigumann P, Redik K, Mätlik K, Speek M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics 2002; 79:628–34.

42. Li J, Kannan M, Trivett AL, Liao H, Wu X, Akagi K, Symer DE. An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. Nucleic Acids Res 2014; 42:4546–62.

43. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol 2002; 3:research0052.

44. Mätlik K, Redik K, Speek M. L1 antisense promoter drives tissue-specific transcription of human genes. J Biomed Biotechnol 2006; 2006:71753.

45. Cruickshanks HA, Tufarelli C. Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics 2009; 94:397–406.

CHAPTER 3: INSIGHTS INTO ESIRNA PRECURSORS

DROSOPHILA MELANOGASTER RETROTRANSPOSON AND INVERTED REPEAT-DERIVED ENDOGENOUS SIRNAS ARE DIFFERENTIALLY PROCESSED IN DISTINCT CELLULAR LOCATIONS

CONTRIBUTIONS

I am first author on the work described below**.** I performed all sequencing, nuclear fractionation, RT-qPCR, and endogenous immunoprecipitation experiments. With regard to IPs performed in stably expressing HA-tagged constructs, Dan Michalski performed the IPs and created the Symplekin, Cpsf73, and Cpsf100 stable cell lines while I created the stable Dcr-2 cell line. Dan Michalski also performed the S1 assay. Kaylyn Bauer performed the immunofluorescence microscopy. Dr. Michael McKain created the SMACR pipeline to analyze small RNA libraries and assisted Dr. Steiniger and myself with bioinformatic analysis. I would also like to thank Dr. Michael Hughes for advice regarding RNA-Seq library preparation, sequencing, and analysis. Dr. Steiniger performed the initial Symplekin immunoprecipitation followed by mass spectroscopy. Manuscript and figures were prepared by Dr. Steiniger and myself.

SUMMARY

Endogenous small interfering (esi)RNAs repress mRNA levels and retrotransposon (retroTn) mobility in *Drosophila* somatic cells. EsiRNAs are primarily generated from transposon and inverted repeat (hairpin) loci in *Drosophila* culture cells. After discovering a nucleus specific physical interaction between the essential esiRNA cleavage enzyme Dcr2 and Symplekin, a component of the core cleavage complex (CCC)

required for 3' end processing of mRNAs, we investigated cellular localization of esiRNA biogenesis and overlap between these pathways. We found that knockdown of CCC components perturbs esiRNA levels and that retroTn precursor transcripts are highly enriched in the nucleus while hairpin RNAs are predominantly cytoplasmic. Additionally, retroTn and hairpin-derived esiRNAs have distinct physical characteristics. Combined, these observations support a novel mechanism in which differences in localization of esiRNA precursors impacts esiRNA biogenesis; hairpin-derived esiRNAs are generated in the cytoplasm independent of Dcr2-Symplekin interactions, while retroTns are processed in the nucleus.

INTRODUCTION

In *Drosophila,* independent groups of small RNAs with overlapping function regulate gene expression using transcriptional and post-transcriptional mechanisms. PIWI-interacting RNAs (piRNAs) are found, most notably, in the germ line where they inhibit transposon (Tn) expression by inducing heterochromatin formation at complementary genomic Tn insertion sites (Brennecke et al., 2007; Fagegaltier et al., 2009; Gu and Elgin, 2013; Haynes et al., 2006; Savva et al., 2013; Sentmanat and Elgin, 2012; Xie et al., 2013). Micro RNAs (miRNAs) and endogenous small interfering RNAs (esiRNAs) are expressed ubiquitously; however miRNAs frequently inhibit translation of protein coding genes (Valencia-Sanchez et al., 2006), while esiRNAs are suggested to inhibit Tn mobility in *Drosophila* somatic cells (Fagegaltier et al., 2009; Savva et al., 2013; Xie et al., 2013) and potentially target mRNAs for degradation using a cytoplasmic RNAi mechanism (Czech et al., 2008; Marques et al., 2010). While PIWI mediated Tn

repression in germ cells and translational inhibition by miRNAs have been actively investigated, the molecular details of how esiRNAs regulate their targets have not been described.

21 nucleotide (nt) esiRNAs are generated from double stranded (ds) precursor RNAs by Dicer-2 (Dcr2) and function through association with Argonaute 2 (Ago2) in *Drosophila* somatic cells (Czech et al., 2008; Ghildiyal et al., 2008; Iwasaki et al., 2015; Kawamura et al., 2008; Okamura et al., 2008a; 2008b). esiRNAs produced in *Drosophila* tissues derive generally from *cis*-natural antisense transcripts (*cis*-NATs), inverted repeat containing single stranded RNAs (hairpins (hps)), and Tns (Czech et al., 2008; Ghildiyal et al., 2008; Okamura et al., 2008b; 2008a). In contrast, *Drosophila* culture cells generate esiRNAs predominantly from long terminal repeat (LTR) Tns and hps; few *cis*-NAT derived esiRNAs are observed in S2 cell derived datasets (Ghildiyal et al., 2008; Kawamura et al., 2008; Russo et al., 2016). Differences between Tn and hp-derived esiRNA biogenesis have not been previously investigated.

Drosophila LTR and non-LTR retroTns are transcribed in both the sense (S) and antisense (AS) directions from RNA polymerase II-like promoters (Russo et al., 2016); a subset of retroTn S and AS transcripts are polyadenylated (Russo et al., 2016). Additionally, hp substrates Esi1 (pseudogene CG18854) and Esi2 (CG44774) are polyadenylated (A. W. Harrington, data not shown). Therefore, the 3’ ends of potential esiRNA substrates are processed by the core cleavage complex (CCC) containing CPSF73, CPSF100 and Symplekin (Michalski and Steiniger, 2015; Ryan et al., 2004; Sullivan et al.,

2009). Potential connections between mRNA 3' end processing and esiRNA biogenesis are intriguing and have not been previously described.

esiRNAs regulate Tns and additional targets via multiple pathways: A canonical cytoplasmic post-transcriptional RNAi pathway in which esiRNAs hybridize to target mRNAs resulting in Ago2 cleavage, and/or transcriptional regulation by induction of heterochromatin in the nucleus. mRNA targets of hp derived esiRNAs have been identified (Czech et al., 2008) and transcript levels of these targets are elevated in *Dcr2* mutant flies (Marques et al., 2010) supporting the post-transcriptional model. Evidence is mounting that Tn derived esiRNAs also mediate heterochromatin formation in *Drosophila* nuclei (Fagegaltier et al., 2009; Haynes et al., 2006; Savva et al., 2013; Sentmanat and Elgin, 2012). *Dcr2* catalytic mutants regulate position effect variegation (Fagegaltier et al., 2009; Haynes et al., 2006), a measure of heterochromatin formation (Agranat et al., 2008; Sun et al., 2001). Additionally, Dcr2 promotes transcription of heat shock genes (Cernilogar et al., 2011) and has been observed in the nuclei of *Drosophila* larvae (Grimaud et al., 2006). These data are consistent with a nuclear pool of Dcr2 that could contribute to transcriptional regulation by induction of heterochromatin in addition to cytoplasmic Dcr2 acting in the RNAi pathway.

To define connections between differential Tn and hp-derived esiRNA processing and cellular location, and to investigate the potential link between mRNA 3' end cleavage and esiRNA biogenesis, interactions between CCC components and Dcr2 were characterized and esiRNAs in control and RNAi-depleted *Drosophila* tissue culture cells were analyzed. These experiments revealed that Dcr2 and the CCC interact, but only in

the nucleus, and that the CCC indirectly regulates esiRNA biogenesis by modulating dsRNA precursor levels. Additionally, Tn- and hp-derived esiRNAs are physically distinct and occupy different subcellular compartments. Tn-derived esiRNAs and their precursors are retained in the nucleus while hp-derived esiRNAs and their precursors are efficiently exported to the cytoplasm. Collectively, these data support a model in which esiRNAs regulate gene expression and Tn mobility via diverse compartmentalized mechanisms.

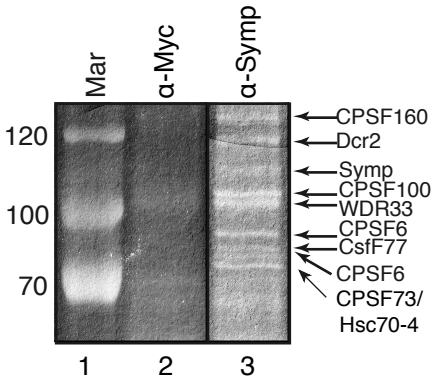
RESULTS

mRNA 3' end processing factor Symplekin interacts with Dcr2.

To identify potential novel CCC binding partners, we immunoprecipitated endogenous Symplekin from crude *Drosophila* culture cell nuclear extracts and identified co-immunoprecipitating proteins by mass spectrometry (Figure 3.1). The most abundant Symplekin interacting proteins in this assay were known CCC components CPSF73 and CPSF100 and additional mRNA 3' end processing proteins CPSF160, WDR33 (CG1109), (Chan et al., 2014; Schönemann et al., 2014) CPSF6 (CG7185), and CstF77 (Sabath et al., 2013). Surprisingly, Dcr2 and Hsc70, proteins known to act in siRNA biogenesis (Iwasaki et al., 2015; 2010) also interacted with Symplekin (Figure 3.1). To confirm this interaction, we performed the reverse immunoprecipitation. Dcr2 co-immunoprecipitated Symplekin and additional CCC factor components, CPSF73 and CPSF100, and R2D2, a known Dcr2 binding partner (Liu et al., 2003) (Figure 3.2A, lane 5). Additionally, Dcr2 co-immunoprecipitated with exogenously

Figure 3.1 Mass spectrometry (MS) identifies Symplekin binding partners

A.



B.

Sample	Name	FlybaseID	MW (Da)	# of Peptides	Function
a	CPSF160	FBgn0024698	164.7	33	3' end processing
b	Dicer-2	FBgn0034246	197.8	23	siRNA
c	Symp	FBgn0037371	132.1	47	3' end processing
d	CPSF100	FBgn0027873	85.4	31	3' end processing
e	WDR33	FBgn0046222	90.5	29	3' end processing
f	CPSF6	FBgn0035872	71.1	29	3' end processing/ poly(A) site selection
g	CstF77	FBgn0003559	84.5	32	3' end processing
h	CPSF6	FBgn0035872	71.1	22	3' end processing/ poly(A) site selection
i	CPSF73	FBgn0261065	76.8	20	3' end processing
	Hsc70-4	FBgn0266599	71.1	11	RISC loading

Figure 3.1 Mass spectrometry (MS) identifies Symplekin binding partners. (A) Endogenous Symplekin was immunoprecipitated from crude nuclear extracts and bound proteins were visualized on an SDS-PAGE gel stained with coomassie blue (lane 3). Markers (Mar, lane 1) are labeled in KDa (left). a-Myc (lane 2) is a non-specific antibody control. Individual bands were cut from the gel and proteins identified by MS. The primary protein in each band is labeled. (B) MS data for each gel slice (samples a-i) is represented with gene name, Flybase ID and known functions of each identified protein.

expressed, HA-tagged Symplekin, CPSF73 and CPSF100 (Figure 3.2B, Figure 3.3).

To determine which region of Symplekin interacts with Dcr2, we immunoprecipitated exogenously expressed HA-tagged Symplekin deletions from *Drosophila* culture cell lysates. The N-terminal region of Symplekin (amino acids 1-271) clearly interacts with endogenous Dcr2 while the C-terminal region (amino acids 272-1165) does not (Figure 3.2B, top, lanes 6 and 8, respectively). Reciprocal immunoprecipitation of endogenous Dcr2 reveals co-immunoprecipitation of HA-tagged Symp(1-271) (Figure 3.2B, bottom, lane 4). To investigate direct Dcr2-CCC interactions, we used a system in which Symplekin mutants are expressed in endogenous Symplekin RNAi-depleted cells. Unlike the interactions observed with full-length Symplekin, Dcr2 does not immunoprecipitate CPSF73 and CPSF100 when only the N-terminal region of Symplekin (HA-Symp(1-271)) is present (Figure 3.2C, top, lane 5). Additionally, very little CPSF73 and CPSF100 interact with Dcr2 in cells expressing HA-Symp (272-1165) (Figure 3.2C, bottom, lane 5). These data suggest that the Symplekin N-terminal region directly interacts with Dcr2 while CPSF73 and CPSF100 are present in this complex via interaction with the C-terminal region of Symplekin (Michalski and Steiniger, 2015).

The Dcr2-CCC complex is functionally distinct from the CCC.

CPSF73, CPSF100 and Symplekin tightly interact in the absence of RNA to form the CCC (Sullivan et al., 2009). When one member of the CCC is depleted, levels of the other factors are dramatically reduced (Sullivan et al., 2009). To determine if Dcr2 is a

bona fide CCC component, we investigated Symplekin and CPSF73 levels in a Dcr2 knockdown. When Dcr2 is depleted,

Symplekin and CPSF73 levels are unchanged (Figure 3.2D, lane 2). Dcr2 levels remain constant when CPSF73 or Symplekin are knocked down (Figure 3.2D, lanes 3 and 4).

Because CCC depletion causes 3' end misprocessing and Dcr2 interacts with the CCC, we determined the effects of Dcr2 depletion on mRNA 3' end processing. First, we mapped the 3' ends of endogenous H2A mRNAs in a Dcr2 depleted sample. No differences in mRNA 3' end processing were observed between Dcr2 knockdown and negative control samples (Figure 3.2E, left, compare lane 4 with lanes 2 and 3); Dcr2 depletion does not cause misprocessing of histone mRNA 3' ends as is observed when CCC components are knocked down (Figure 3.2E, left, compare lane 4 with lanes 5 and 6) (Michalski and Steiniger, 2015; Sullivan et al., 2009). Additionally, an RT-qPCR assay in which mRNA 3' end misprocessing is compared to total mRNA levels of a specific gene (Tatomer et al., 2014) was used to assay the effect of Dcr2 knockdown on mRNA 3' end misprocessing of polyadenylated genes. Very little misprocessing of a canonical polyadenylated mRNA (*sop*) was observed in a Dcr2 depleted sample as compared to the positive Symplekin control (Figure 3.2E, right). Collectively, these data support a model in which the Dcr2-CCC complex is functionally distinct from the CCC as mRNA 3' end processing is unaffected in the absence of Dcr2.

Figure 3.2 Dcr2 interacts with the N-terminal region of Symplekin; however, Dcr2 depletion does not affect mRNA 3' end processing.

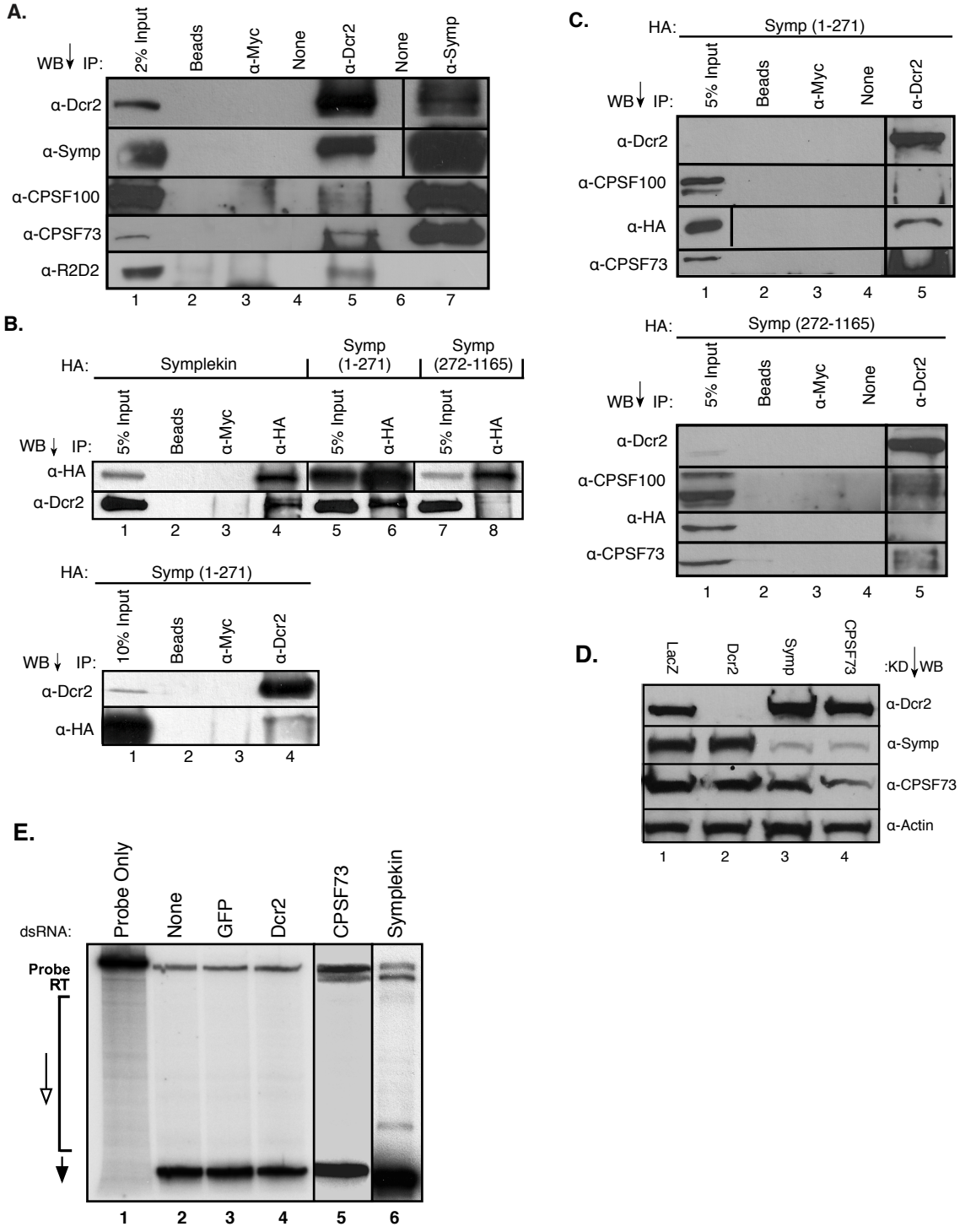
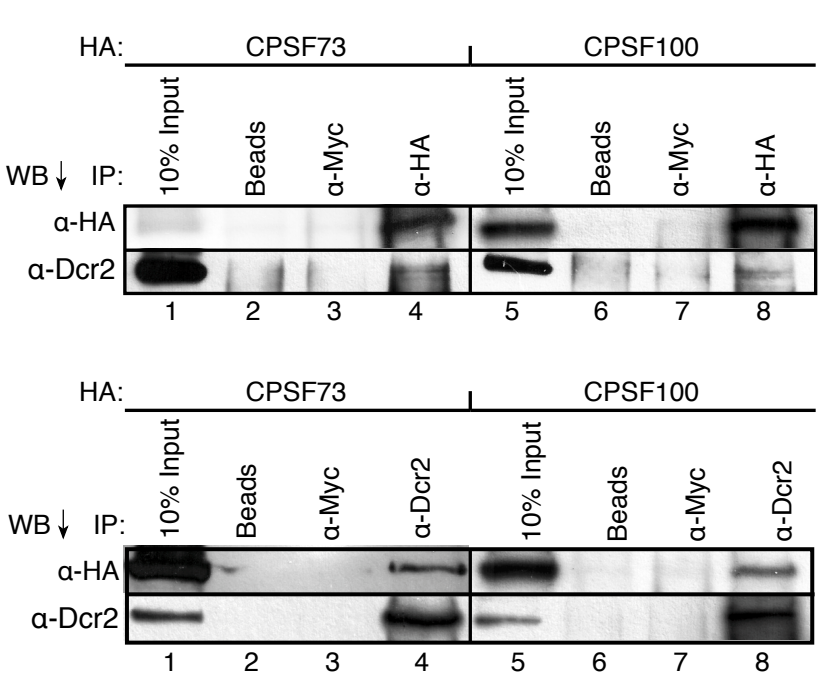


Figure 3.2 Dcr2 interacts with the N-terminal region of Symplekin; however, Dcr2 depletion does not affect mRNA 3' end processing. (A) Dcr2 co-immunoprecipitates (co-IP) with the core cleavage complex (CCC) and R2D2 in *Drosophila* culture cell crude nuclear extract. Antibodies used for immunoprecipitation (IP) are shown above. Antibodies used for western blot (WB) are listed to the left. 'Beads' and 'a-Myc' are negative controls. No sample was loaded in lanes labeled 'None.' (B) Dcr2 binds Symplekin amino acids 1-271 and not amino acids 272-1165. Exogenously expressed HA-tagged Symplekin deletions are defined above the blots. Other labels are as in (A). WB of full length Symplekin and Symplekin mutant IPs are the top figure while co-IP of Symp(1-271) with Dcr2 is shown in the bottom WB. (C) CCC components CPSF100 and CPSF73 do not interact with Dcr2 in the absence of full length Symplekin. WB of Symp(1-271) IP (top) and WB of Symp (272-1165) IP (bottom) from Symplekin RNAi-depleted samples are shown. WB are labeled as in (B). (D) Dcr2 depletion does not affect CCC component protein levels. RNAi-depleted proteins are listed above the blot. Antibodies used for WB are listed right. (E) Dcr2 RNAi-depletion does not cause mRNA 3' end misprocessing. An S1 nuclease assay was used to map histone (H)2A 3' ends (left). Knockdowns are shown at the top. Potential mRNA 3' end products are shown to the left: RT is the read-through misprocessed product, the open arrow marks the region of other misprocessed products, and the black arrow defines the properly processed product. RT-qPCR using primers that amplify misprocessed *sop* mRNAs (right) reveals very little misprocessed *sop* in Dcr2 knockdown samples. Knockdowns are shown on the x-axis. Degree of Misprocessing = 2^DDDCt(ORF-MP). The Symplekin KD degree of misprocessing is 86.2. Error bars represent one standard deviation.

Figure 3.3 Dcr2 binds exogenously expressed CCC components CPSF73 and CPSF100.



3.3 Dcr2 binds exogenously expressed CCC components CPSF73 and CPSF100. (A) HA-tagged, full-length CPSF73 and CPSF100 were immunoprecipitated from *Drosophila* culture cells stably expressing these proteins (top, lanes 4 and 8, respectively). Dcr2 co-immunoprecipitation with both CPSF73 and CPSF100 was identified by western blot (bottom, lanes 4 and 8, respectively). 'Beads' and a-Myc are negative controls. 10% input (lanes 1 and 5) was loaded for reference. (B) Dcr2 was immunoprecipitated from *Drosophila* culture cells expressing HA-tagged, full-length CPSF73 and CPSF100 (bottom, lanes 4 and 8 respectively). Co-immunoprecipitation of HA-CPSF73 and HA-CPSF100 was identified by western blot (top, lanes 4 and 8, respectively). Controls are as in (A).

Dcr2 interacts with the CCC in the nucleus.

To investigate subcellular localization of the Dcr2-CCC complex, Dmel-2 cells were first effectively separated into cytoplasmic and nuclear fractions using a novel, refined fractionation technique (See Materials and Methods, Chapter 5, Refined Nuclear and Cytoplasmic Fractionation). Western blots reveal pools of Dcr2 and Symplekin in both the nucleus and the cytoplasm (Figure 3.4A, lanes 1 and 3). While Dcr2 is primarily cytoplasmic and Symplekin is generally nuclear in accordance with their roles in RNAi and mRNA 3' end processing, respectively, an appreciable amount of each protein is found in the complementary subcellular compartment (Figure 3.4A, lanes 1 and 3). Additionally, immunofluorescence with antibodies to the endogenous proteins confirms the presence of both Symplekin and Dcr2 in the nucleus (Figure 3.4A). This assay also shows Symplekin and Dcr2 in the cytoplasm, consistent with their roles in cytoplasmic polyadenylation and RNAi, respectively (Barnard et al., 2004; Kim and Richter, 2006; Nishida et al., 2013).

Immunoprecipitations of endogenous Dcr2 from refined Dmel-2 nuclear and cytoplasmic fractions show that nuclear Dcr2 co-immunoprecipitates the CCC and R2D2 (Figure 3.4B, top, lane 5), while cytoplasmic Dcr2 only interacts with R2D2 (Figure 3.4B, bottom, lane 5); no interaction between cytoplasmic Dcr2 and the CCC is observed. Nuclear Symplekin co-immunoprecipitates Dcr2 and other CCC components, CPSF73 and CPSF100, but not R2D2 (Figure 3.4B, top, lane 7). Additionally, cytoplasmic Symplekin does not interact with Dcr2 (Figure 3.4B, bottom, lane 7). Together these data support a model in which Dcr2 forms distinct nuclear and cytoplasmic complexes.

Figure 3.4 Dcr2 only interacts with the CCC in the nucleus.

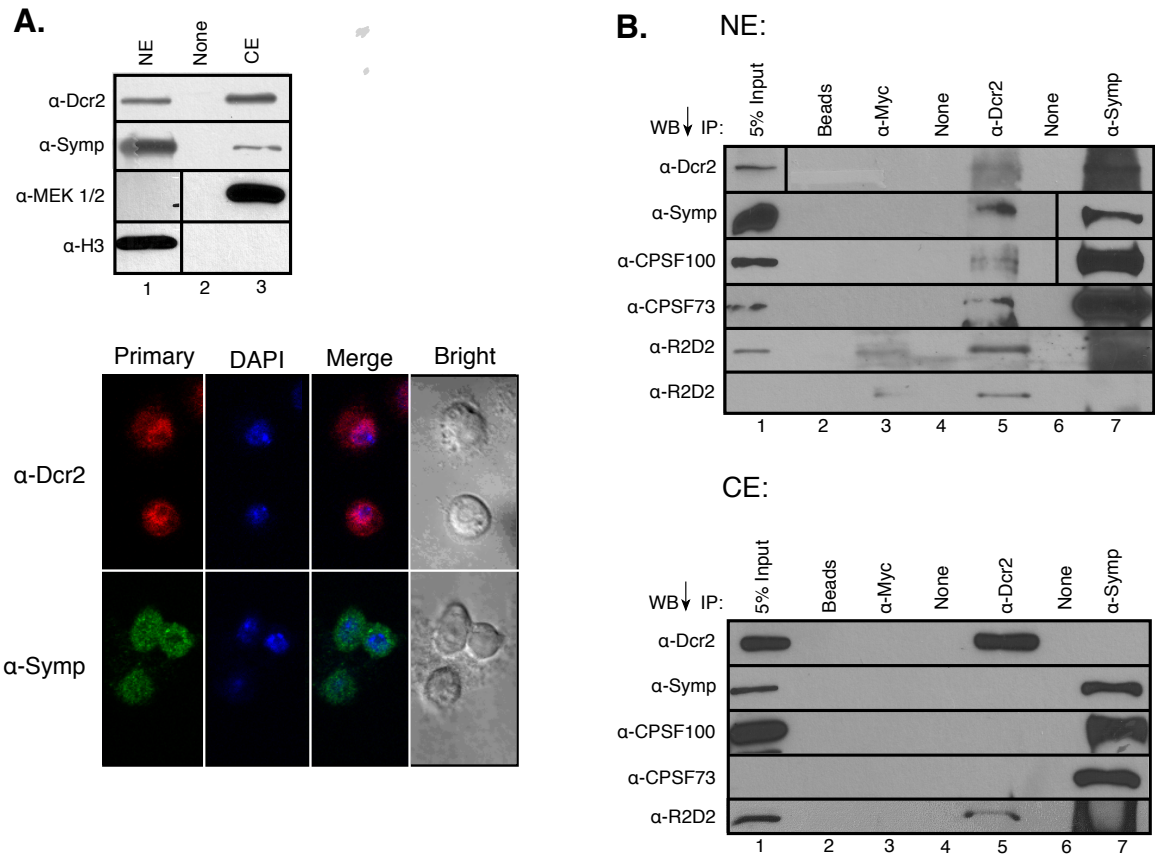


Figure 3.4 Dcr2 only interacts with the CCC in the nucleus. (A) Dcr2 is present in the nucleus. WB of refined nuclear (NE) and cytoplasmic extracts (CE) reveals a nuclear pool of Dcr2 (top). MEK 1/2 and H3 are cytoplasmic and nuclear controls, respectively. Immunofluorescence of *Drosophila* culture cells with anti-Dcr2 and anti-Symp antibodies shows both Dcr2 and Symplekin co-localizing with the DAPI stained nucleus (bottom). (B) Endogenous Dcr2 co-immunoprecipitates the CCC and R2D2 from refined NEs (top). No interaction between Dcr2 and the CCC is observed in CE (bottom). Antibodies used for IP are shown above. Antibodies used for WB are listed left. 'Beads' and 'a-Myc' are negative controls. A lighter exposure of the R2D2 WB is shown at the bottom of the top NE group.

The CCC indirectly regulates esiRNA abundance.

To investigate the role of the CCC in esiRNA biogenesis, CPSF73, Symplekin and Dcr2 were first independently RNAi-depleted from *Drosophila* culture cells. Then RNA was isolated and separated into large (>200 nts) and small (<200 nts) fractions, rRNA depleted and sequenced (Figure 3.5A). RNA-seq reads were mapped to the *Drosophila* genome and transcriptome using RUM (Grant et al., 2011) while smRNA-seq reads were mapped and analyzed using a novel pipeline termed **S**equence **M**apping, **A**nnotation, and **C**ounting for sm**R**NAs or SMACR (Figure 3.5B). Only siRNAs and miRNAs were further analyzed. Interestingly, ~40% of RNA-seq reads were non-unique although the samples were depleted of rRNAs (Figure 3.6). Also, the percentage of non-unique reads changes significantly with knockdown of Dcr2 and Symplekin (Figure 3.6). These data support previous claims that Dmel-2 culture cells have undergone Tn expansion (Maisonhaute et al., 2007; Potter et al., 1979; Tchurikov et al., 1981; Wen et al., 2014) and indicate that increased numbers of Tns may contribute to higher overall expression of repetitive sequences and abundance of Dcr2 dependent siRNAs. Tn expansion makes *Drosophila* culture cells an excellent system for studying esiRNAs biogenesis.

We first assessed how depletion of Dcr2 and CCC components CPSF73 and Symplekin affect siRNA dynamics in *Drosophila* culture cells. Normalized siRNA reads mapping to pre-miRNAs, non-coding (nc)RNAs, miRNAs, Tns, and two loci that produce RNAs capable of forming hairpin structures, Esi1/2 (hps), were added to give the total siRNA pool for each sample. The percentages of siRNAs mapping to miRNAs, Tns and hps were then calculated. Importantly, no statistically significant changes were observed

Figure 3.5 Workflow for high throughput sequencing and small RNA analysis

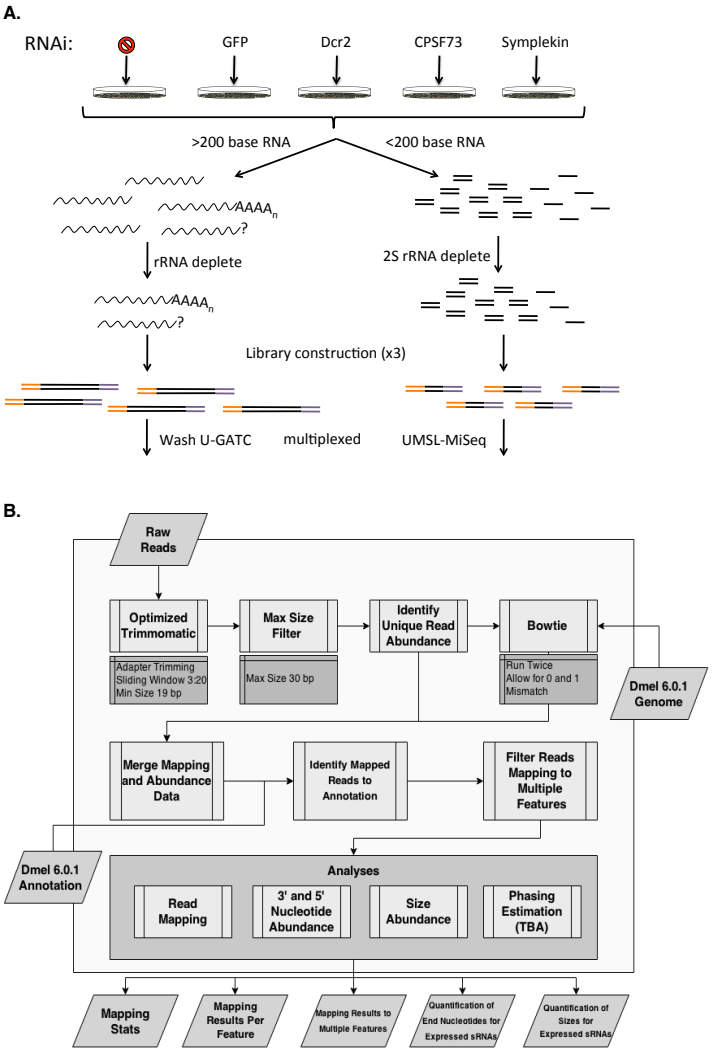


Figure 3.5 Work flow for high throughput sequencing and small RNA analysis (SMACR). (A) *Drosophila* cells were individually depleted of Dcr2, CPSF73, Symplekin, and GFP (or LacZ). An additional fifth sample was untreated. The untreated and GFP samples represent controls. RNA was isolated from each sample and fractionated into RNAs > than 200 nts and RNAs < 200 nts. Each sample was depleted of appropriate rRNAs followed by library construction in triplicate. RNA-seq was performed at Washington University while smRNA-seq was performed at University of Missouri-St. Louis. (B) Adapters were trimmed from the raw reads followed by filtering out all small RNAs larger than 30 nts. Small RNAs were mapped using Bowtie and were then sorted by feature: miRNA, transposon, hairpin, or non-coding RNA. The normalized read count of each unique small RNA mapping to each feature was calculated together with 3' and 5' and size abundance.

Figure 3.6 HTS statistics

A. RNA-seq

Sample	Total Reads	% Mapping	Read Depth	% Unique	% Non-Uniq	p-values
Blank1	24698945	98.8	82.1	58.3	41.7	0.112
Blank2	26712148	98.3	88.7	57.1	42.9	
Blank3	26318291	98.2	87.4	57.2	42.8	
LacZ1	30020327	98.5	99.7	57.0	43.0	
LacZ2	35561677	98.4	118.1	56.3	43.7	
LacZ3	26072570	98.4	86.6	56.4	43.6	
Dcr2-1	22668363	98.2	75.3	51.8	48.2	6.949E-05
Dcr2-2	23826043	98.0	79.2	51.3	48.7	
Dcr2-3	24236601	98.0	80.5	51.1	48.9	
CPSF73-1	27808742	98	92.4	58.2	41.8	0.521
CPSF73-2	22354433	98.1	74.3	56.6	43.4	
CPSF73-3	24986528	98.3	83.0	56.3	43.7	
Symp1	28837214	98.5	95.8	53.3	46.7	0.007
Symp2	25326064	98.2	84.1	53.2	46.8	
Symp3	31041518	97.2	103.1	51.8	48.2	

B. SiRNA-seq

Mismatch 0:

Sample	Total Reads	% Mapping
Blank1	296806	67.8
Blank2	257223	69.5
Blank3	144083	73.0
Blank4-BR	935588	84.8
CPSF73-1	328575	74.0
CPSF73-2	272747	78.2
CPSF73-3	292295	79.1
CPSF73-4-BR	857530	70.5
Dcr2-1	227654	71.8
Dcr2-2	312873	73.3
Dcr2-3	237995	71.7
Dcr2-4-BR	784740	81.7
LacZ1	390118	71.4
LacZ2	330015	73.1
LacZ3	329624	74.1
GFP-BR	928750	73.6
Sym1	468391	73.2
Sym2	332289	73.9
Sym3	389207	73.4
Sym4-BR	963513	79.0

Mismatch 1:

Sample	Total Reads	% Mapping
Blank1	296806	93.7
Blank2	257223	93.9
Blank3	144083	94.3
Blank4-BR	935588	96.9
CPSF73-1	328575	95.3
CPSF73-2	272747	95.3
CPSF73-3	292295	95.4
CPSF73-4-BR	857530	89.3
Dcr2-1	227654	95.5
Dcr2-2	312873	95.7
Dcr2-3	237995	95.6
Dcr2-4-BR	784740	95.1
LacZ1	390118	94.0
LacZ2	330015	94.0
LacZ3	329624	94.4
GFP-BR	928750	93.7
Sym1	468391	94.6
Sym2	332289	94.6
Sym3	389207	94.7
Sym4-BR	963513	94.3

Figure 3.6 HTS statistics. (A) Sample name, total number of reads, percent of reads mapping, read depth (# mapped reads/*Drosophila* transcriptome size (30.1 Mba)), percent unique and percent non-unique reads are shown for technical triplicates of each sample. A Student's T-test was used to determine if the observed differences in percentages of non-uniquely mapping reads between samples was statistically significant. Corresponding *p* values are shown in the last column. (B) Total number of reads and percent or reads mapping when zero mismatches are allowed (left) and one mismatch is allowed (right) for three technical triplicates and one biological replicate (BR) of each sample.

between untreated and LacZ dsRNA RNAi treated control samples (Figure 3.7A). When Dcr2 is depleted, the percentage of esiRNAs mapping to Tns and hps decreases significantly while the portion of miRNAs in the pool increases (Figure 3.7A). Biogenesis of hp esiRNAs is more dependent on Dcr2 than esiRNAs processed from Tn precursors as Dcr2 depletion reduces hp esiRNAs ~7.3 fold as compared to the control while Tn esiRNAs are only reduced ~1.3 fold (Figure 3.7A). Surprisingly, depletion of CCC components CPSF73 and Symplekin has differential effects on Tn and hp derived esiRNAs; the proportion of Tn derived esiRNAs increases slightly while the number of esiRNAs generated from hps decreases (Figure 3.7A). Knockdown of Symplekin and CPSF73 did not affect miRNA levels (Figure 3.7A). Together these data support a model in which esiRNAs are differentially processed from Tn and hp substrates in *Drosophila* culture cells.

To investigate potential explanations for the observed differences between Tn and hp derived esiRNA levels and differential effects of Dcr2 and CCC factor depletion on esiRNAs biogenesis, we first examined esiRNAs and precursor levels for individual Tns (LTR retroTn, mdg1, and non-LTR retroTn, jockey) and one hp locus (Esi2) in these samples as compared to a LacZ dsRNA RNAi control. The number of esiRNAs mapping to mdg1 and jockey increase in response to CCC depletion while esiRNAs generated from Esi2 decrease in these samples (Figure 3.7B). Dcr2 knockdown reduces esiRNAs mapping to mdg1, jockey and Esi2, although, Esi2 is most affected (Figure 3.7B). RetroTn esiRNA precursors consist of hybridized S and antisense AS retroTn transcripts (Russo et al., 2016). We previously reported that both S and AS mdg1 and jockey transcript levels are

elevated in Dcr2 depleted cells (Russo et al., 2016). Knockdown of CCC components Symplekin or CPSF73 results in little to no change in S mdg1 or jockey transcript abundance while the corresponding AS transcripts are significantly elevated (Figure 3.7B). esiRNA hp substrates can be formed solely by *cis* hybridization of inverted repeat sequences on the S strand. Examination of the Esi2 S transcript reveals that abundance is relatively unchanged when CCC factors are knocked down (Figure 3.7B). These data correlate with the whole genome analyses discussed above (Figure 3.7A).

We hypothesize that retroTn dsRNAs are generally blunt-ended with complementary S and AS strands while hp substrates having multiple inverted repeats are more complex and variable (Figure 3.7C). To determine if retroTn and hp esiRNA precursor structure could be altered by depletion of CCC factors, we investigated 3' end misprocessing of mdg1 and Esi2. Bedgraphs representing S and AS reads mapping to mdg1{}305 and surrounding sequences in CPSF73, Symplekin and LacZ depleted samples show few reads mapping beyond the 3' ends of S and AS transcripts (Figure 3.7D, top). In contrast, RNA-seq reads mapping downstream of Esi2 are readily detectable in Symplekin and CPSF73 knockdowns (Figure 3.7D, RNA-seq (-)) indicating read-through transcription of CG44774 (Sullivan et al., 2009). These read through sequences are AS to mRNAs transcribed from a gene immediately upstream of Esi2 (CG6903) (Figure 3.7D). Evidence that these sequences hybridize to form substrates for Dcr2 is apparent in esiRNAs generated from this region in Symplekin and CPSF73 knockdowns (Figure 3.7D, siRNA-seq). Additionally, the 3' ends of CG6903 mRNAs are also misprocessed in CCC depleted samples potentially providing sequences complementary to Esi2 (Figure 3.7D,

RNA-seq (+)). Together, these data indicate that CCC depletion does not alter retroTn Dcr2 substrate structure, while Esi2 structure could be altered by the presence of opposite strand sequences. Inefficient cleavage of these altered Esi2 structures could provide one explanation for the lowered Esi2 esiRNA levels observed in CPSF73 and Symplekin knockdowns (Figure 3.7B).

Figure 3.7 CCC depletion differentially affects esiRNA biogenesis from retroTns and inverted repeat loci

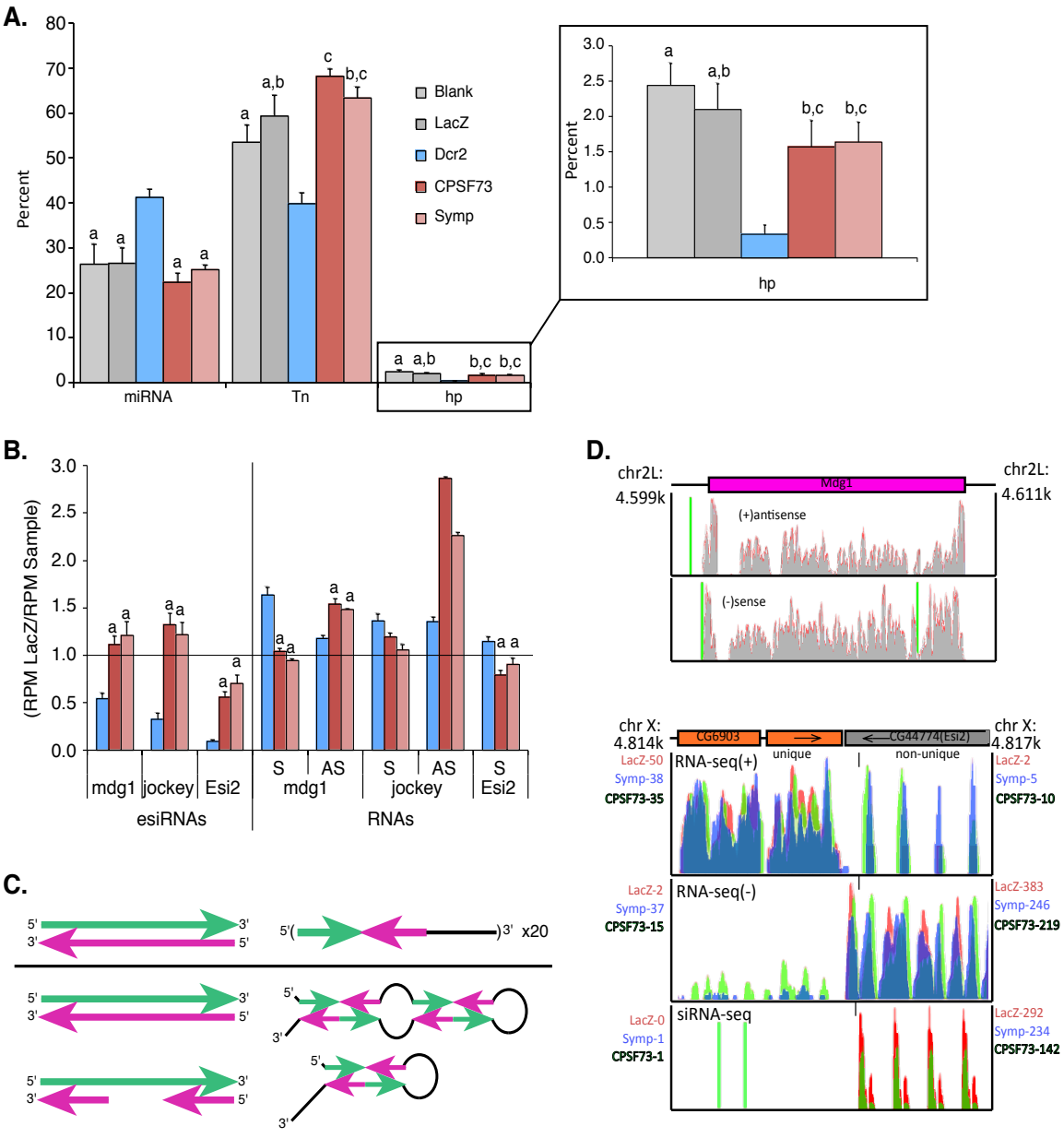


Figure 3.7 CCC depletion differentially affects esiRNA biogenesis from retroTns and inverted repeat loci. (A) Percentages of miRNAs, transpsoson (Tn)-derived and hairpin (hp)-derived siRNAs from Symplekin (pink), CPSF73 (red), Dcr2 (blue), and control (gray) samples are shown. Percentages are the total miRNA, Tn or hp normalized read count (RPM) divided by the total normalized read count (summed normalized miRNA, pre-miRNA, Tn, non-coding RNA, and hp RPMs). ANOVA was used to compare percentages for each sample within the miRNA, Tn, and hp groups. Letters represent statistically indistinguishable samples. Results of the hp analysis are magnified at right. Error bars represent one standard deviation. (B) RNAi-depleted samples are represented as in (A). RPMs of esiRNAs mapping to mdg1 and jockey retroTns and hp Esi2 in Dcr2, Symp and CPSF73 depleted samples were normalized to corresponding esiRNAs in LacZ samples (left). RPMs of RNA-seq reads mapping sense (S) and (AS) stands of mdg1 and jockey and the S Esi2 RNA in these samples were also normalized to corresponding esiRNAs in LacZ samples (right). Statistical analyses and error bars are as in (A). (C) Potential secondary structures for Tns (left) and hps (right). Complementary regions are shown in green and magenta. (D) Depletion of CCC components causes 3' end misprocessing of Esi2 substrates, but not S and AS mdg1 retroTn transcripts. RNA-/siRNA-seq reads from LacZ (control), Symplekin, and CPSF73 depleted samples were visualized using the UCSC genome browser. Schematics of the genomic region are shown above the bedgraphs. Non-unique S and AS mdg1 transcripts (top) from all three samples overlap (gray) with only a few unique reads (light green) flanking the retroTn. 3' end misprocessing is observed for both CG44774 (Esi2) and neighboring gene, CG6903 in Symplekin (RNA-seq, blue) and CPSF73 (RNA-seq, green) depleted samples, but not the LacZ control (red). esiRNAs (siRNA-seq) map to CG6903 in CCC knockdowns.

Hp and Tn-derived esiRNAs are differentially processed.

To investigate potential specific processing defects associated with Dcr2 or CCC knockdown, miRNAs and esiRNAs were filtered by length, 3' and 5' base. Generally, there were few differences among depleted samples as compared to controls (Figure 3.8) with the notable exception that Dcr2 depletion reduces the percentage of 21 nt esiRNAs, the most common length, (Figure 3.8A). This is not observed in non-Dcr2 dependent miRNA distributions in Dcr2 depleted samples (Figure 3.8). Negligible effects of CCC factor knockdown on esiRNAs size and end nucleotide preference indicate that CPSF73 and Symplekin do not directly affect Dcr2 catalytic activity.

Examination of length differences between esiRNAs and miRNAs in control samples reveals that miRNAs are almost equally distributed among 21, 22 and 23 nts, while 21 nt is the dominant length of Tn and hp-derived esiRNAs (Figure 3.9A). Unexpectedly, variations in length distributions were also observed between Tn and hp-derived esiRNAs. Approximately 75% of esiRNAs generated from Tns are 21 nt with 19, 20, 22 and 23-mers almost evenly comprising the remaining 25% (Figure 3.9A). In contrast, ~62% of hp-derived esiRNAs are 21 nt and ~23% are 22 nt; the proportion of 22-mers in the hp generated esiRNA pool is significantly greater than for Tn-derived esiRNAs (Figure 3.9A).

Dramatic differences between 3' and 5' base preference are also observed for miRNAs and esiRNAs. miRNAs and Tn generated esiRNAs predominantly end in an A, while the 3' base of hp derived esiRNAs is most often a G (Figure 3.9B). Although Tn generated esiRNAs and miRNAs most often end in A, the frequencies of other

Figure 3.8 Physical characteristics of miRNAs, Tn- and hp-derived esiRNAs in Symplekin, CPSF73, Dcr2 and control samples

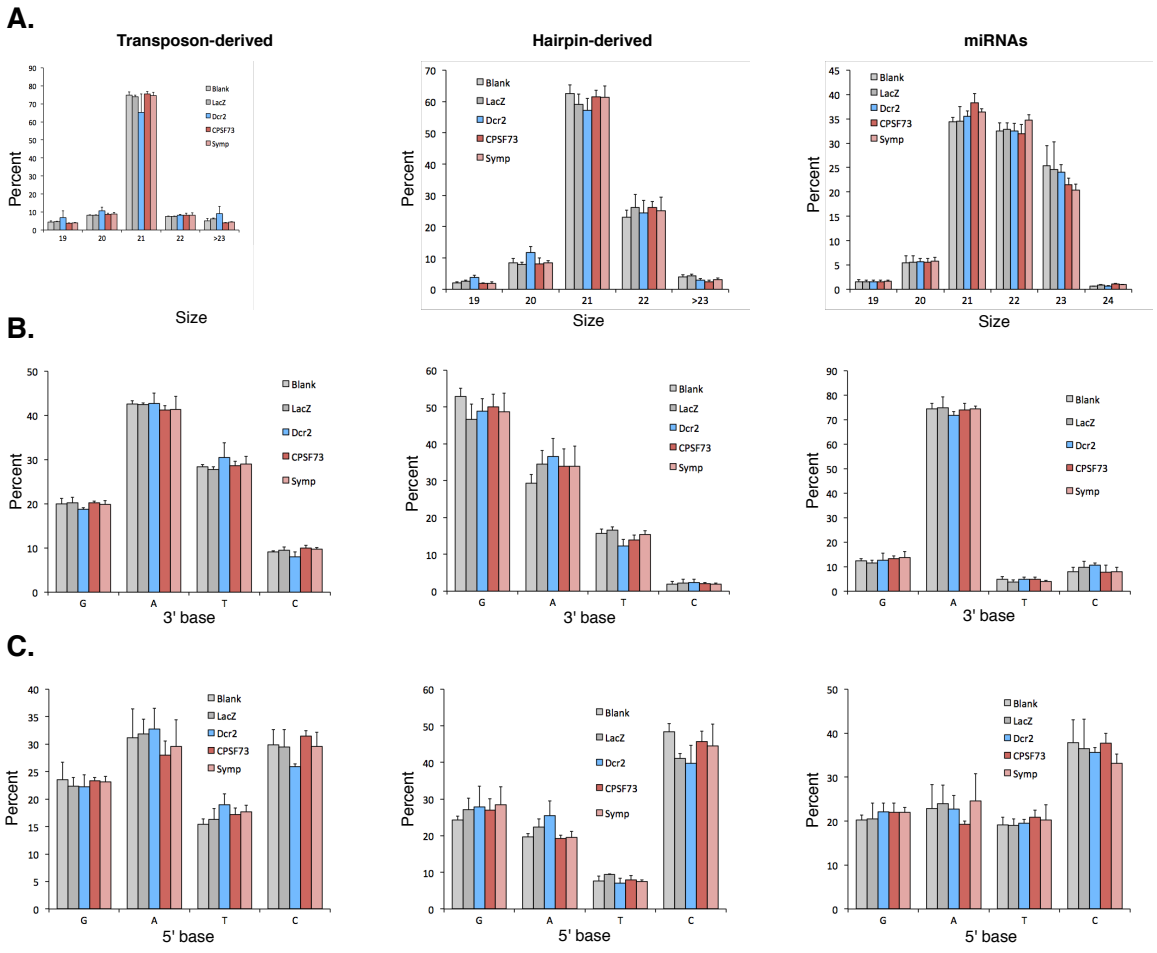


Figure 3.8 Physical characteristics of miRNAs, Tn- and hp-derived esiRNAs in Symplekin, CPSF73, Dcr2 and control samples. Mapped siRNAs were sorted by type and filtered by size (21-24 nts) (A), 3’ base (B), and 5’ base (C) for each sample. The abundance of normalized read counts in each category was then summed and the percentage of each individual category was calculated for all samples.

Percentages for each category were then plotted.nucleotides at the 3' end differ. Only ~40% of Tn generated esiRNAs have 3' As, while ~30% have Ts, ~20% have Gs and ~10% have Cs; ~75% of miRNAs have 3' As, while only ~5% have Ts, ~12% have Gs, and ~8% have Cs. Additionally, ~50% of hp generated esiRNAs have 3' Gs, ~30% have As, ~15% have Ts and ~5% have Cs (Figure 3.9B). Therefore, while the esiRNA 3' nt is generally more diverse than for miRNAs, significant differences between Tn and hp derived esiRNAs are also evident. Less difference between 5' nucleotide distributions of all three siRNA classes is observed. Approximately 25% of miRNAs and esiRNAs have a 5' G, while C is the most abundant siRNA 5' nucleotide (Figure 3.10A). Collectively, these data indicate that esiRNAs processed from Tn and hp substrates have diverse physical characteristics and support a model in which these two precursors are differentially processed in Dmel-2 cells.

RetroTn precursors and esiRNAs are retained in the nucleus.

To investigate potential differences in subcellular localization of retroTn and hp substrates, *Drosophila* culture cells were separated into refined nuclear and cytoplasmic fractions, total RNA was isolated and RT-qPCR was performed on Dm297 and mdg1 retroTn RNAs, and Esi1 and Esi2 containing substrate transcripts CG47744 and CG18854, respectively. A control canonical mRNA, GAPDH, was found to be slightly enriched in the nucleus (Figure 3.9C). Surprisingly, the retroTn transcripts were overwhelmingly enriched in the nucleus as compared to the GAPDH control; Dm297 and mdg1 were 254-229 fold and 370-221 fold nuclear, respectively, depending on which retroTn ORF was targeted by RT-qPCR (Figure 3.9C). Subsequent analyses of additional retroTns blood,

jockey and juan also show nuclear retention of these precursors (Figure 3.10B). CG47744 and CG18854 are not dramatically enriched in either the cytoplasm or nucleus, resembling the GAPDH control (Figure 3.9C).

To assess the cellular localization of retroTn, Esi1 and Esi2-derived esiRNAs, we measured levels of the most abundant esiRNAs from these precursors with custom Taqman assays. A cytoplasmic miRNA control (mir2A) is ~2.5 fold enriched in the cytoplasm (Figure 3.9D). Strikingly, both Dm297 and mdg1-derived esiRNAs are highly enriched in the nucleus while localization of Esi2.1 and Esi1.2 is slightly enhanced in the cytoplasm (Figure 3.9D). These data correlate with nuclear enrichment of retroTn precursors and mild cytoplasmic retention of hp substrates (Figure 3.9C). Together these data support a model where ds retroTn substrates are retained and processed in the nucleus while single stranded Esi1 and Esi2 precursors are exported to the cytoplasm for Dcr2-dependent generation of esiRNAs (Figure 3.9E).

Precursor levels increase in the nucleus after knock down of CCC components

Levels of both retrotransposon and hairpin precursors increase after knock down of CCC components. In both cases, this could be due to improper export from 3' end processing defects. In the case of the retrotransposon, the Sense strand is not only more abundant, but also more heavily polyadenylated. This makes the Antisense strand the limiting factor in dsRNA formation. Since members of the CCC have a lower effect on the AS strand, total dsRNA substrate would not likely increase while the detectable level of the Sense transcript would. With respect to hairpins precursor, decreased export also explains the increase in nuclear levels.

Figure 3.9 Transposon and hp Dcr2 substrates are differentially processed in different cellular compartments

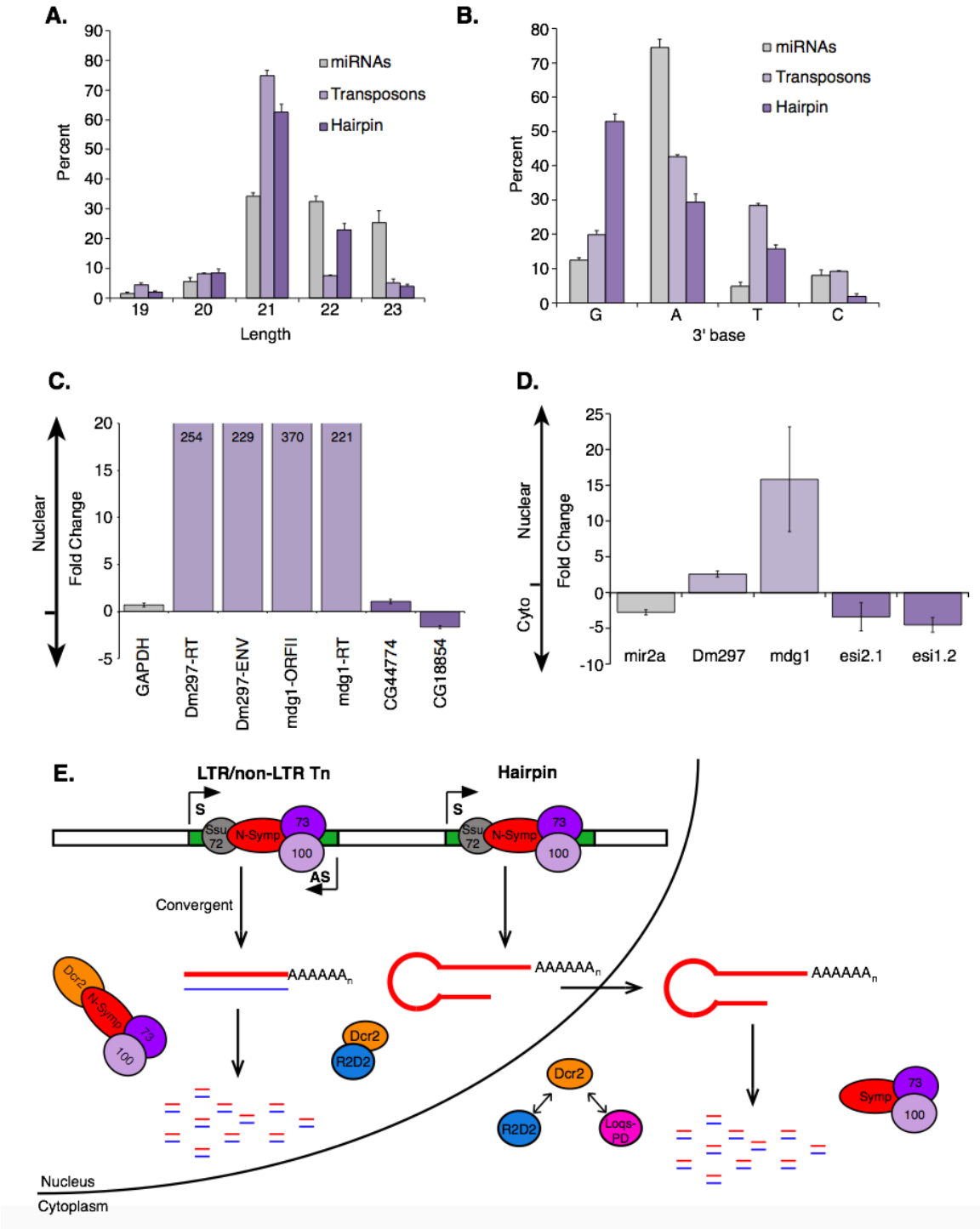


Figure 3.9 Transposon and hp Dcr2 substrates are differentially processed in different cellular compartments. (A) Percentage of miRNAs (gray), Transposons (light purple), and hairpin (purple) mapping esiRNAs in the LacZ control sample that are 19-23 nts. Error bars represent one standard deviation. (B) Percentage of miRNAs, Transposons, and hairpin mapping esiRNAs in the LacZ control sample that have a 3' G, A, T, or C. Colors and error bars are as in (A). (C) RT-qPCR of retroTn and Esi1/2 mRNAs isolated from refined nuclear and cytoplasmic fractions reveals nuclear retention of retroTn esiRNAs precursors. RT-qPCR targets are shown on the x-axis. CG44774 is the Esi1 precursor. CG18854 is the Esi2 precursor mRNA. Fold change is the average of three experiments and is calculated as $2^{(Ct(Nuclear)-Ct(Cytoplasm))}$. Error bars are as in (A). (D) Taqman qPCR of retroTn and Esi1/2 derived esiRNAs isolated from refined nuclear and cytoplasmic fractions shows nuclear retention of retroTn derived esiRNAs. Labels, calculations, and error bars are as in (C). (E) Data support a model in which double stranded retroTn transcripts are retained and processed to esiRNAs in the nucleus while RNAs containing inverted repeats are exported and processed in the cytoplasm. Dcr2 interacts with the N-terminal 271 amino acids of Symplekin in the nucleus, but not in the cytoplasm.

Figure 3.10 Distribution and location of miRNA, Tn- and hp-derived esiRNA 5' nucleotides and precursors.

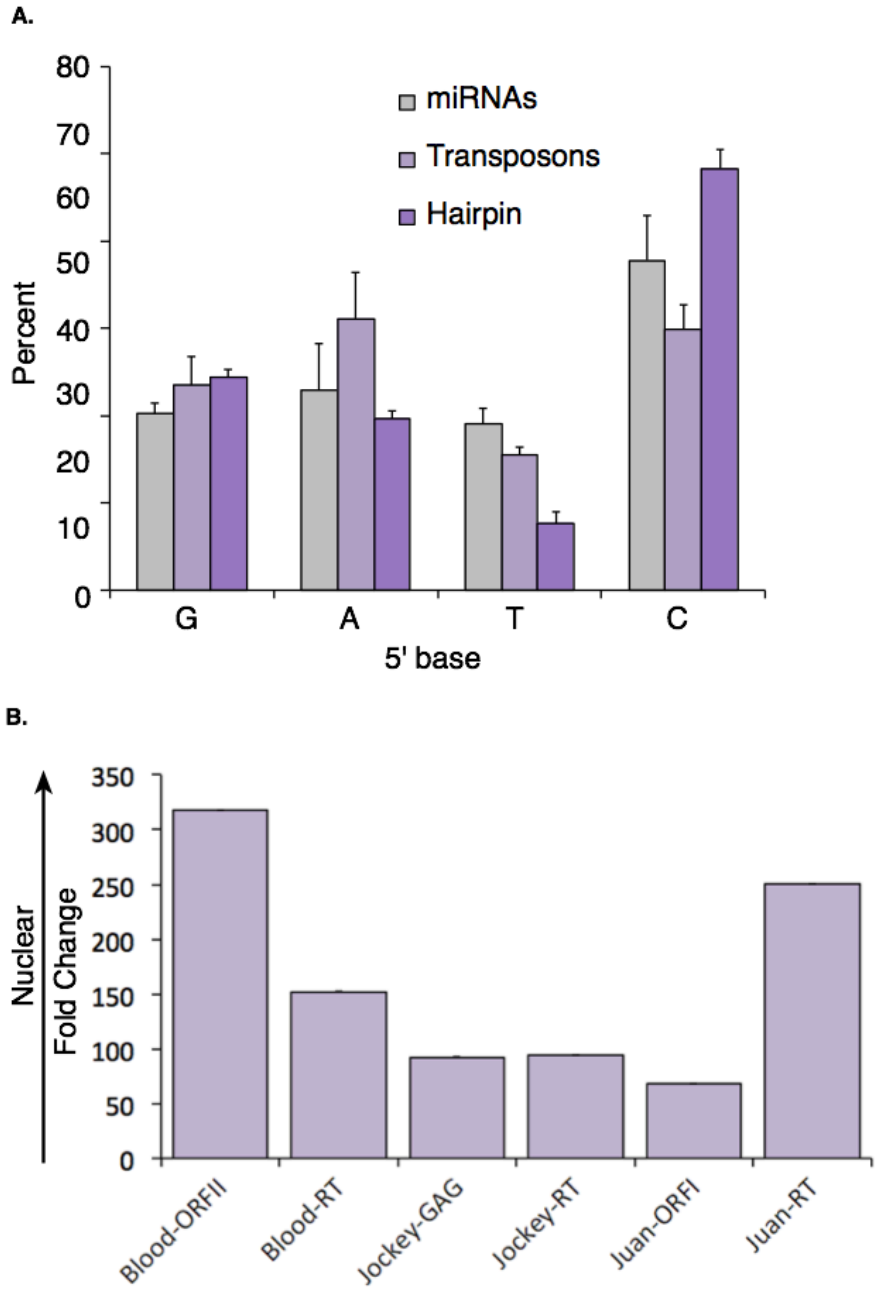


Figure 3.10 Distribution and location of miRNA, Tn- and hp-derived esiRNA 5' nucleotides and precursors. (A) siRNAs were mapped, filtered and percentages of 5' nucleotide were calculated as described in Figure S5. Only percentages in the control sample are plotted. (B) *Drosophila* cells were separated into nuclear and cytoplasmic fractions, total RNA was isolated and RT-qPCR was performed on blood, juan and jockey transcripts. Differences between RNA levels in the cytoplasm and nucleus are plotted for each RNA and showed that blood, juan and jockey transcripts are retained in the nucleus.

Figure 3.11 CCC knockdown results in increased levels of both hairpin and transposon precursors in the nuclear compartment

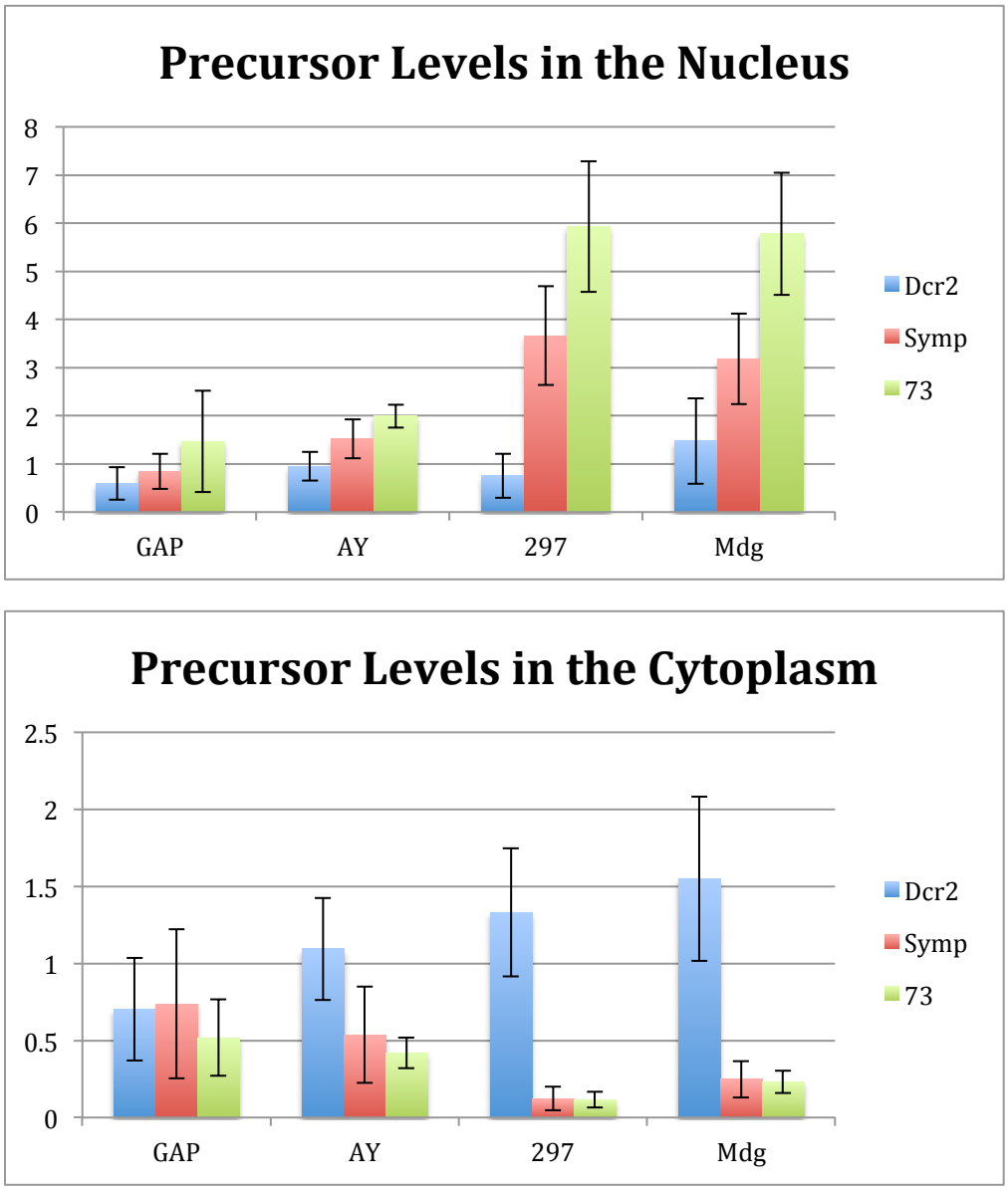


Figure 3.11 Top: Nuclear levels of precursors, Bottom: cytoplasmic levels of precursors in Dicer-2 (blue) Symplekin (red) and Cpsf73 (green) knock downs. Hairpin precursor tested is AY, retrotransposon precursors are Dm297 and Mdg1. Gap was used as a control, all ddCq was in reference to 18S rRNA to eliminate potential effects of CCC depletion.

DISCUSSION

Since the discovery of endogenous small interfering (esi)RNAs in *Drosophila*, very little progress in understanding their biogenesis and molecular mechanisms of action has been made. Here we provide evidence that components of two major RNA processing pathways, 3' end processing and esiRNA biogenesis, interact in *Drosophila* somatic cells, a connection not previously reported.

mRNA 3' end processing performed by the CCC is co-transcriptional and therefore occurs in the nucleus (Bentley, 2005; Greenleaf, 1993). The RNA pol II CTD phosphatase Ssu72 interacts with the N-terminal region of Symplekin to direct processing of mRNAs with a 3' poly(A) tail (Xiang et al., 2010) and with the stem loop binding protein for replication dependent histone mRNAs (Dan Michalski, data not shown). Here, we show that this N-terminal region of Symplekin can also interact directly with esiRNA processing factor Dcr2 (Figure 3.2B, 3.3) in the nuclear compartment (Figure 3.4B). The Symplekin C-terminal region binds CPSF73 and CPSF100 to form the CCC (Michalski and Steiniger, 2015), therefore leaving the N-terminal region free to bridge the CCC and other cellular factors. While previous work shows that regulation of Tns by piRNAs in the *Drosophila* germline is a nuclear process (X. A. Huang et al., 2013; Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012; Wang and Elgin, 2011) and researchers have documented a nuclear pool of Dcr2 that associates with heat shock loci and transcription machinery in *Drosophila* (Cernilogar et al., 2011), potential nuclear functions of Dcr2 in *Drosophila* somatic cells have not been extensively investigated (Fagegaltier et al., 2009). Our data support a model in which the N-terminal

region of Symplekin mediates Dcr2-CCC complex formation, but only when the CCC is not actively engaged in co-transcriptional mRNA 3' end processing (Figure 3.9).

To understand the functional implications of CCC-Dcr2 interactions, esiRNA and precursor levels were measured in Symplekin and CPSF73 RNAi-depleted samples. We observed increased esiRNAs generated from Tns (Figure 3.7A), constant S retroTn precursor levels, and dramatically increased AS retroTn precursors in these samples (Figure 3.7B). We hypothesize that dsRNA precursor levels for mdg1 and jockey retroTns are determined by AS transcript levels as these are limiting (Russo et al., 2016). More AS transcript effectively leads to an increase in retroTn Dcr2 substrates. As Dcr2 activity seems unaffected by interaction with the CCC, higher substrate levels would lead to increased esiRNAs levels (Figure 3.7B). In contrast, RNAi-depletion of CCC components reduced the number of esiRNAs generated from hp substrates while the Esi2 precursor level was constant (Figure 3.7B). We predict that esiRNAs are generated from hps in the cytoplasm (Figure 3.9) and our data show that the Dcr2-CCC complex does not form in the cytoplasm (Figure 3.4). Therefore, a hypothesis for the observed molecular phenotypes is inefficient nuclear export of hp RNAs in Symplekin and CPSF73 RNAi-depleted samples as these Dcr2 substrates have mRNA 3' end processing defects (Figure 3.7D). Previous work shows that less polyadenylated RNAs are ineffectively exported from the nucleus (Y. Huang and Carmichael, 1996). Additionally, 3' end misprocessing of RNAs generated from the Esi2 locus (Figure 3.7C) might lead to changes in secondary structure that unpredictably affect nuclear export. Inefficient nuclear export of hp RNAs with modified 3' ends would not change total substrate levels, but could result in less

Esi2-derived esiRNAs as cytoplasmic hp substrate levels would be reduced. Taken together, these data support a model in which the CCC indirectly affects esiRNAs generated from both Tn and hp precursors (Figure 3.9).

Bioinformatic analyses of Tn and hp derived esiRNAs reveals physical distinctions between these groups (Figures 3.8abc, 3.9). Additionally, retroTn substrates and retroTn-derived esiRNAs are highly enriched in the nucleus while hp substrates and esiRNAs are cytoplasmic. We hypothesize that these observed disparities are directly related to distinct substrate secondary structures (Figure 3.7C) and compartmentalization of esiRNA biogenesis factors required to process each structure. dsRNAs derived from transcription of retroTns are comprised of both S and AS transcripts, (Russo et al., 2016) generally resulting in fully complementary, blunt-ended dsRNAs as many AS retroTn transcripts are poorly polyadenylated (Russo et al., 2016). The secondary structures of hps containing multiple inverted repeats are likely variable and complex with frayed ends. Previous *in vitro* assays suggest that Dcr2 alone can bind and processively cleave blunt dsRNAs. However, Dcr2 requires a co-factor, Loqs-PD, to process dsRNAs with frayed termini presumably because Loqs-PD allows Dcr2 to bind a substrate internally (Sinha et al., 2015); Loqs-PD is cytoplasmic in *Drosophila* culture cells (Miyoshi et al., 2010). Lastly, if effects observed from the CCC knockdown are due to 3' processing defects, and hairpins are processed in the cytoplasm, decreased export would result higher levels of precursor in the nucleus as the machinery to cleave the hairpin is not present. However, the same cannot be said for the retrotransposon precursors because even though the machinery is present for processing, it is most likely

increased retention of the Sense strand that contributes to the increased level. Since the antisense strand is the limiting factor in the amount of dsRNA that is formed, increased levels of the Sense strand would not be able to form greater amounts of substrate. Taken together, these data suggest a model in which nuclear retained blunt-ended, fully complementary retroTn precursors can be processed in the nucleus by Dcr2 alone while more complicated hp substrates requiring Loqs-PD are cleaved in the cytoplasm by Dcr2 (Figure 3.9). This model is supported by our observations that esiRNAs map the entire length of retroTns (Figure 3.7D). Additionally, previous work shows that R2D2 and Dcr2 aggregate in cytoplasmic D2 bodies together with hp substrates (Nishida et al., 2013).

This model predicts that depletion of Loqs-PD would only affect cleavage of hp substrates, but not esiRNAs generated from retroTns. Zhou *et al* previously reported that depletion of Loqs isoforms reduced the number of esiRNAs derived from both hps and Tns (Zhou et al., 2009); however, close examination of the data reveal that retroTn-mapping esiRNAs were unaffected by Loqs knockdown. The most notably affected Tn, Proto-P, is not regulated by the esiRNA pathway (Harrington and Steiniger, 2016).

In conclusion, our data support a novel model in which esiRNAs are differentially processed from retroTn and hp substrates; retroTn precursors are processed by Dcr2 in the nucleus, while biogenesis of esiRNAs from hp substrates occurs in the cytoplasm. Additionally, Dcr2 clearly interacts with the CCC in the nucleus, but not in the cytoplasm. These data contribute significantly to our understanding of Dcr2 dependent esiRNA production in *Drosophila* culture cells, but questions regarding Dcr2-CCC complex assembly and function remain. Future studies investigating the role of the Dcr2-CCC

complex in both mRNA 3’ end processing and retroTn substrate processing will further elucidate molecular details of how these proteins function in *Drosophila* culture cells.

REFERENCES

Agranat, L., Raitskin, O., Sperling, J., Sperling, R., 2008. The editing enzyme ADAR1 and the mRNA surveillance protein hUpf1 interact in the cell nucleus. Proceedings of the National Academy of Sciences 105, 5028–5033. doi:10.1073/pnas.0710576105

Barnard, D.C., Ryan, K., Manley, J.L., Richter, J.D., 2004. Symplekin and xGLD-2 are required for CPEB-mediated cytoplasmic polyadenylation. Cell 119, 641–651. doi:10.1016/j.cell.2004.10.029

Bentley, D.L., 2005. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. Curr Opin Cell Biol 17, 251–256. doi:10.1016/j.ceb.2005.04.006

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. doi:10.1093/bioinformatics/btu170

Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G.J., 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell 128, 1089–1103. doi:10.1016/j.cell.2007.01.043

Cernilogar, F.M., Onorati, M.C., Kothe, G.O., Burroughs, A.M., Parsi, K.M., Breiling, A., Sardo, Lo, F., Saxena, A., Miyoshi, K., Siomi, H., Siomi, M.C., Carninci, P., Gilmour, D.S., Corona, D.F.V., Orlando, V., 2011. Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. Nature 480, 391–395. doi:10.1038/nature10492

Chan, S.L., Huppertz, I., Yao, C., Weng, L., Moresco, J.J., Yates, J.R., Ule, J., Manley, J.L., Shi, Y., 2014. CPSF30 and Wdr33 directly bind to AAUAAA in mammalian mRNA 3' processing. Genes Dev 28, 2370–2380. doi:10.1101/gad.250993.114

Czech, B., Malone, C.D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J.A., Sachidanandam, R., Hannon, G.J., Brennecke, J., 2008. An endogenous small interfering RNA pathway in Drosophila. Nature 453, 798–802. doi:10.1038/nature07007

Fagegaltier, D., Bougé, A.-L., Berry, B., Poisot, E., Sismeiro, O., Coppée, J.-Y., Théodore, L., Voinnet, O., Antoniewski, C., 2009. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proceedings of the National Academy of Sciences 106, 21258–21263. doi:10.1073/pnas.0809208105

Ghildiyal, M., Seitz, H., Horwich, M.D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E.L.W., Zapp, M.L., Weng, Z., Zamore, P.D., 2008. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. Science 320, 1077–1081. doi:10.1126/science.1157396

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., Pierce, E.A., 2011. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). Bioinformatics 27, 2518–2528. doi:10.1093/bioinformatics/btr427

Greenleaf, A.L., 1993. Positive patches and negative noodles: linking RNA processing to transcription? Trends Biochem. Sci. 18, 117–119.

Grimaud, C., Bantignies, F., Pal-Bhadra, M., Ghana, P., Bhadra, U., Cavalli, G., 2006. RNAi components are required for nuclear clustering of Polycomb group response elements. Cell 124, 957–971. doi:10.1016/j.cell.2006.01.036

Gu, T., Elgin, S.C.R., 2013. Maternal Depletion of Piwi, a Component of the RNAi System, Impacts Heterochromatin Formation in Drosophila. PLoS Genet 9, e1003780. doi:10.1371/journal.pgen.1003780.s010

Harrington, A.W., Steiniger, M., 2016. Bioinformatic analyses of sense and antisense expression from terminal inverted repeat transposons in Drosophila somatic cells. Fly (Austin) 1–10. doi:10.1080/19336934.2016.1165372

Haynes, K.A., Caudy, A.A., Collins, L., Elgin, S.C.R., 2006. Element 1360 and RNAi components contribute to HP1-dependent silencing of a pericentric reporter. Curr. Biol. 16, 2222–2227. doi:10.1016/j.cub.2006.09.035

Huang, X.A., Yin, H., Sweeney, S., Raha, D., Snyder, M., Lin, H., 2013. A major epigenetic programming mechanism guided by piRNAs. Dev. Cell 24, 502–516. doi:10.1016/j.devcel.2013.01.023

Huang, Y., Carmichael, G.G., 1996. Role of polyadenylation in nucleocytoplasmic transport of mRNA. Molecular and Cellular Biology 16, 1534–1542.

Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T., Tomari, Y., 2010. Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. Molecular Cell 39, 292–299. doi:10.1016/j.molcel.2010.05.015

Iwasaki, S., Sasaki, H.M., Sakaguchi, Y., Suzuki, T., Tadakuma, H., Tomari, Y., 2015. Defining fundamental steps in the assembly of the Drosophila RNAi enzyme complex. Nature 521, 533–536. doi:10.1038/nature14254

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T.N., Siomi, M.C., Siomi, H., 2008. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. Nature 453, 793–797. doi:10.1038/nature06938

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler,

D., 2002. The human genome browser at UCSC. Genome Res. 12, 996–1006. doi:10.1101/gr.229102
Kim, J.H., Richter, J.D., 2006. Opposing polymerase-deadenylase activities regulate cytoplasmic polyadenylation. Molecular Cell 24, 173–183. doi:10.1016/j.molcel.2006.08.016

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. doi:10.1186/gb-2009-10-3-r25

Le Thomas, A., Rogers, A.K., Webster, A., Marinov, G.K., Liao, S.E., Perkins, E.M., Hur, J.K., Aravin, A.A., Tóth, K.F., 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev 27, 390–399. doi:10.1101/gad.209841.112

Liu, Q., Rand, T.A., Kalidas, S., Du, F., Kim, H.-E., Smith, D.P., Wang, X., 2003. R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. Science 301, 1921–1925. doi:10.1126/science.1088710

Maisonhaute, C., Ogereau, D., Hua-Van, A., Capy, P., 2007. Amplification of the 1731 LTR retrotransposon in Drosophila melanogaster cultured cells: origin of neocopies and impact on the genome. Gene 393, 116–126. doi:10.1016/j.gene.2007.02.001

Marques, J.T., Kim, K., Wu, P.-H., Alleyne, T.M., Jafari, N., Carthew, R.W., 2010. Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. Nat Struct Mol Biol 17, 24–30. doi:10.1038/nsmb.1735

Michalski, D., Steiniger, M., 2015. In vivo characterization of the Drosophila mRNA 3' end processing core cleavage complex. RNA 21, 1404–1418. doi:10.1261/rna.049551.115

Miyoshi, K., Miyoshi, T., Hartig, J.V., Siomi, H., Siomi, M.C., 2010. Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. RNA 16, 506–515. doi:10.1261/rna.1952110

Miyoshi, K., Okada, T.N., Siomi, H., Siomi, M.C., 2009. Characterization of the miRNA-RISC loading complex and miRNA-RISC formed in the Drosophila miRNA pathway. RNA 15, 1282–1291. doi:10.1261/rna.1541209

Nechaev, S., Fargo, D.C., Santos, dos, G., Liu, L., Gao, Y., Adelman, K., 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. Science 327, 335–338. doi:10.1126/science.1181421

Nishida, K.M., Miyoshi, K., Ogino, A., Miyoshi, T., Siomi, H., Siomi, M.C., 2013. Roles of

R2D2, a cytoplasmic D2 body component, in the endogenous siRNA pathway in Drosophila. Molecular Cell 49, 680–691. doi:10.1016/j.molcel.2012.12.024

Okamura, K., Balla, S., Martin, R., Liu, N., Lai, E.C., 2008a. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. Nat Struct Mol Biol 15, 581–590. doi:10.1038/nsmb.1438

Okamura, K., Chung, W.-J., Ruby, J.G., Guo, H., Bartel, D.P., Lai, E.C., 2008b. The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. Nature 453, 803–806. doi:10.1038/nature07015

Potter, S.S., Brorein, W.J., Dunsmuir, P., Rubin, G.M., 1979. Transposition of elements of the 412, copia and 297 dispersed repeated gene families in Drosophila. Cell 17, 415–427.

Rogers, S.L., Rogers, G.C., 2008. Culture of Drosophila S2 cells and their use for RNAi-mediated loss-of-function studies and immunofluorescence microscopy. Nat Protoc 3, 606–611. doi:10.1038/nprot.2008.18

Rozhkov, N.V., Hammell, M., Hannon, G.J., 2013. Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev 27, 400–412. doi:10.1101/gad.209767.112

Russo, J., Harrington, A.W., Steiniger, M., 2016. Antisense Transcription of Retrotransposons in Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. Genetics 202, 107–121. doi:10.1534/genetics.115.177196

Ryan, K., Calvo, O., Manley, J.L., 2004. Evidence that polyadenylation factor CPSF-73 is the mRNA 3' processing endonuclease. RNA 10, 565–573.

Sabath, I., Skrajna, A., Yang, X.-C., Dadlez, M., Marzluff, W.F., Dominski, Z., 2013. 3'-End processing of histone pre-mRNAs in Drosophila: U7 snRNP is associated with FLASH and polyadenylation factors. RNA 19, 1726–1744. doi:10.1261/rna.040360.113

Santos, dos, G., Schroeder, A.J., Goodman, J.L., Strelets, V.B., Crosby, M.A., Thurmond, J., Emmert, D.B., Gelbart, W.M., FlyBase Consortium, 2015. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res 43, D690–7. doi:10.1093/nar/gku1099

Savva, Y.A., Jepson, J.E.C., Chang, Y.-J., Whitaker, R., Jones, B.C., St Laurent, G., Tackett, M.R., Kapranov, P., Jiang, N., Du, G., Helfand, S.L., Reenan, R.A., 2013. RNA editing regulates transposon-mediated heterochromatic gene silencing. Nature Communications 4, 2745. doi:10.1038/ncomms3745

Schönemann, L., Kühn, U., Martin, G., Schäfer, P., Gruber, A.R., Keller, W., Zavolan, M.,

Wahle, E., 2014. Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. Genes Dev. doi:10.1101/gad.250985.114

Sentmanat, M.F., Elgin, S.C.R., 2012. Ectopic assembly of heterochromatin in Drosophila melanogaster triggered by transposable elements. Proceedings of the National Academy of Sciences 109, 14104–14109. doi:10.1073/pnas.1207036109

Sienski, G., Dönertas, D., Brennecke, J., 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell 151, 964–980. doi:10.1016/j.cell.2012.10.040

Sinha, N.K., Trettin, K.D., Aruscavage, P.J., Bass, B.L., 2015. Drosophila Dicer-2 Cleavage Is Mediated by Helicase- and dsRNA Termini-Dependent States that Are Modulated by Loquacious-PD. Molecular Cell 58, 406–417. doi:10.1016/j.molcel.2015.03.012

Sullivan, K.D., Steiniger, M., Marzluff, W.F., 2009. A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. Molecular Cell 34, 322–332. doi:10.1016/j.molcel.2009.04.024

Sun, F.L., Cuaycong, M.H., Elgin, S.C., 2001. Long-range nucleosome ordering is associated with gene silencing in Drosophila melanogaster pericentric heterochromatin. Molecular and Cellular Biology 21, 2867–2879. doi:10.1128/MCB.21.8.2867-2879.2001

Tatomer, D.C., Rizzardi, L.F., Curry, K.P., Witkowski, A.M., Marzluff, W.F., Duronio, R.J., 2014. Drosophila symplekin localizes dynamically to the histone locus body and tricellular junctions. Nucleus 0. doi:10.4161/19491034.2014.990860

Tchurikov, N.A., Ilyin, Y.V., Skryabin, K.G., Ananiev, E.V., Bayev, A.A., Krayev, A.S., Zelentsova, E.S., Kulguskin, V.V., Lyubomirskaya, N.V., Georgiev, G.P., 1981. General properties of mobile dispersed genetic elements in Drosophila melanogaster. Cold Spring Harb. Symp. Quant. Biol. 45 Pt 2, 655–665.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., Parker, R., 2006. Control of translation and mRNA degradation by miRNAs and siRNAs. Genes Dev 20, 515–524. doi:10.1101/gad.1399806

Wang, S.H., Elgin, S.C.R., 2011. Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proceedings of the National Academy of Sciences 108, 21164–21169. doi:10.1073/pnas.1107892109

Wen, J., Mohammed, J., Bortolamiol-Becet, D., Tsai, H., Robine, N., Westholm, J.O., Ladewig, E., Dai, Q., Okamura, K., Flynt, A.S., Zhang, D., Andrews, J., Cherbas, L., Kaufman, T.C., Cherbas, P., Siepel, A., Lai, E.C., 2014. Diversity of miRNAs, siRNAs, and

piRNAs across 25 Drosophila cell lines. … research.

Xiang, K., Nagaike, T., Xiang, S., Kilic, T., Beh, M.M., Manley, J.L., Tong, L., 2010. Crystal structure of the human symplekin-Ssu72-CTD phosphopeptide complex. Nature 467, 729–733. doi:10.1038/nature09391

Xie, W., Donohue, R.C., Birchler, J.A., 2013. Quantitatively increased somatic transposition of transposable elements in Drosophila strains compromised for RNAi. PLoS ONE 8, e72163. doi:10.1371/journal.pone.0072163

Yang, X.-C., Burch, B.D., Yan, Y., Marzluff, W.F., Dominski, Z., 2009. FLASH, a proapoptotic protein involved in activation of caspase-8, is essential for 3' end processing of histone pre-mRNAs. Molecular Cell 36, 267–278. doi:10.1016/j.molcel.2009.08.016

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J.A., Perrimon, N., Hannon, G.J., 2009. Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. RNA 15, 1886–1895. doi:10.1261/rna.1611309

CHAPTER 4: MATERIALS AND METHODS

Strand Specific RT-qPCR.

Reverse Transcription for Strand Specific qPCR. 50 ng of total Dmel-2 RNA isolated using the RNeasy MinElute Cleanup kit (Qiagen) were reverse transcribed with RevertAid reverse transcriptase (Thermo Scientific) and a strand specific, gene-specific reverse transcription primer (RT sense or antisense primer). This primer contains a unique nucleic acid tag 5' of the complementary sequence that does not map to the *Drosophila* transcriptome (Table 4.1). The RT reaction contained 5x reaction buffer (no random hexamers or oligo dT), 1 µl Ribolock (40U/µL), 1 mM dNTPs, 100 nM RT-primer and 2 µL of RevertAid (200U/µL) in a total volume of 20 µL. The reaction was incubated at 50°C for 1 hour, heat inactivated at 85°C for 5 minutes, and then diluted 1:10 with nuclease free water. Quantitative PCR (qPCR) was optimized and performed on a Bio-Rad CFX96 Real-Time system using SYBR Green detection chemistry (Bio-Rad SsoAdvanced Universal SYBR green). Briefly, 4 µl of diluted cDNA (10-fold) was mixed with 5 µl 2× SYBR green (Bio-Rad) and 0.5 µl of forward and 0.5 µl of reverse primer (500 nM final concentration). Initial denaturation was carried out at 95° for 3 min followed by a 30-sec denature step and a 30-sec annealing step (40×). Gene-specific primers for strand-specific qPCR are provided in the table below. RT-qPCR experiments were conducted in technical triplicates.

Table 4.1 Primers used in Strand Specific RT-qPCR

Description	5' to 3' Sequence	Position (bp)
Dm297-RT RT sense primer*	caagactcagctggttcctgcgacttcttcttcttcaagc	4674-4693
Dm297-RT sense qPCR-F*	ggcagacagagacggag	4629-4645
Dm297-RT sense qPCR-R (tag)*	caagactcagctggttcctg	unique tag
Dm297-RT RT antisense primer*	gagaagctcatagtacctcgggcagacagagacggag	4629-4693
Dm297-RT antisense qPCR-F (tag)*	gagaagctcatagtacctcg	unique tag
Dm297-RT antisense qPCR-R*	cgacttcttcttcttcaagc	4673-4693
Dm297-env RT sense primer	gcctgtcccgatatatgaacctcaataatgtcgttgtg	6317-6335
Dm297-env sense qPCR-F	gacaccactatacacaccac	6269-6289
Dm297-env sense qPCR-R (tag)	gcctgtcccgatatatgaac	unique tag
Dm297-env RT antisense primer	gttattaatcgtaataacgggacaccactatacacaccac	6269-6288
Dm297-env antisense qPCR-F (tag)	gttattaatcgtaataacgg	unique tag
Dm297-env antisense qPCR-R	ctcaataatgtcgttgtg	6317-6335
blood-ORFII RT sense primer	ccagaaaaccgcgtgtctacgctgcttacgcatactgtc	2624-2643
blood-ORFII sense qPCR-F	cgtaaaaggcgaatcgcctg	2534-2554
blood-ORFII sense qPCR-R (tag)	ccagaaaaccgcgtgtctac	unique tag
blood-ORFII RT antisense primer	cccatacgcgagatacactgcgtaaaaggcgaatcgcctg	2534-2554
blood-ORFII antisense qPCR-F (tag)	cccatacgcgagatacactg	unique tag
blood-ORFII antisense qPCR-R	gctgcttacgcatactgtc	2624-2643
blood-RT RT sense primer*	ctcgtcgctttcggatttgccaaagcctcgttaagtggcg	4726-4746
blood-RT sense qPCR-F*	cctataccaacagatgccgac	4647-4668
blood-RT sense qPCR-R (tag*)	ctcgtcgctttcggatttgc	unique tag
blood-RT RT antisense primer*	gactgcagacatcagatcggcctataccaacagatgccgac	4647-4668
blood-RT antisense qPCR-F (tag)*	gactgcagacatcagatcgg	unique tag
blood-RT antisense qPCR-R*	caaagcctcgttaagtggcg	4726-4746
mdg1-ORFII RT sense primer	cgtttaaaccagaccgacaccgggtaatgttattaccgctg	2133-2154
mdg1-ORFII sense qPCR-F	ctgagatcggtgaggatatcg	2053-2074
mdg1-ORFII sense qPCR-R (tag)	cgtttaaaccagaccgacac	unique tag
mdg1-ORFII RT antisense primer	ggcacacttatgctcagcacctgagatcggtgaggatatcg	2053-2074
mdg1-ORFII antisense qPCR-F (tag)	ggcacacttatgctcagcac	unique tag
mdg1-ORFII antisense qPCR-R	cgggtaatgttattaccgctg	2133-2154
mdg1-RT RT sense primer*	ctacgatgccgctaagaaccctcctgctctgtagtggac	4923-4942
mdg1-RT sense qPCR-F*	gtaaacaagcatgtggagcg	4824-4844
mdg1-RT sense qPCR-R (tag)*	ctacgatgccgctaagaacc	unique tag
mdg1-RT RT antisense primer*	gatcggcgaccatttgtgaggtaaacaagcatgtggagcg	4824-4844
mdg1-RT antisense qPCR-F (tag)*	gatcggcgaccatttgtgag	unique tag
mdg1-RT antisense qPCR-R*	ctcctgctctgtagtggac	4923-4942
jockey-gag RT sense primer	gcctagaattacctaccgcgtgctccatattctccgtttcag	919-897
jockey-gag sense qPCR-F	acctatcctcacccttctc	776-795
jockey-gag sense qPCR-R (tag)	gcctagaattacctaccgcg	unique tag
jockey-gag RT antisense primer	ctacgttacagcgtgcatagacctatcctcacccttctc	776-795
jockey-gag antisense qPCR-F (tag)	ctacgttacagcgtgcatag	unique tag
jockey-gag antisense qPCR-R	tgctccatattctccgtttcag	919-897
jockey-RT RT sense primer*	gcctagaattacctaccgcgggaagttgaagtggtctgaag	2922-2943
jockey-RT sense qPCR-F*	gtggacattgataatgccacaag	2841-2864
jockey-RT sense qPCR-R (tag)*	gcctagaattacctaccgcg	unique tag
jockey-RT RT antisense primer*	ctacgttacagcgtgcataggtggacattgataatgccacaag	2841-2864
jockey-RT antisense qPCR-F (tag)*	ctacgttacagcgtgcatag	unique tag
jockey-RT antisense qPCR-R*	ggaagttgaagtggtctgaag	2922-2943
juan-ORFI RT sense primer	gctgctcttatcacatttgccctaggttttgtagcatggatttg	586-609
juan-ORFI sense qPCR-F	ctgtgagttctacacgtacgatac	499-522
juan-ORFI sense qPCR-R (tag)	gctgctcttatcacatttgc	unique tag
juan-ORFI RT antisense primer	gccagtcgtattccttctcgctgtgagttctacacgtacgatac	499-522
juan-ORFI antisense qPCR-F (tag)	gccagtcgtattccttctcg	unique tag
juan-ORFI antisense qPCR-R	cctaggttttgtagcatggatttg	586-609
juan-RT RT sense primer*	gctgctcttatcacatttgcctgtgagcagttgacaaccac	2168-2189
juan-RT sense qPCR-F*	gcgcaatgtaaaaacatatccg	2082-2104
juan-RT sense qPCR-R (tag)*	gctgctcttatcacatttgc	unique tag
juan-RT RT antisense primer*	gccagtcgtattccttctcggcgcaatgtaaaaacatatccg	2082-2104
juan-RT antisense qPCR-F (tag)*	gccagtcgtattccttctcg	unique tag
juan-RT antisense qPCR-R*	ctgtgagcagttgacaaccac	2168-2189
Actin5C RT sense primer*	gtgcgtacaccttaatcaccgggtgccacacgcagctcat	280-298
Actin5C sense qPCR-F*	ggcgcagagcaagcgtggta	175-195
Actin5C sense qPCR-R (tag)*	gtgcgtacaccttaatcacc	unique tag
18s RT sense primer*	ctctcctcctcagcatgctgaccagacttgccctccaat	553-571
18s sense qPCR-F*	ctgagaaacggctaccacatc	400-422
18s sense qPCR-R (tag)*	ctctcctcctcagcatgctg	unique tag

'*' indicates that primers were used for detection of strand-specific transcripts in poly A+/- fractions.

Northern Blotting.

A total of 50 µg of total RNA was separated on a 1% agarose denaturing gel at 100 V for ~4.5 hr. RNA was transferred to hybond+ nitrocellulose membrane, UV cross-linked, and rRNA was stained with methylene blue. Following prehybridization, blots were probed with 32P-end labeled ~50-nt probes (table 4.2). Blots were grouped based on predicted transcript levels (group no. 1-S Dm297, S blood and S mdg1; group no. 2-AS Dm297, AS blood and AS mdg1; group no. 3-S and -AS juan, and S and AS jockey) and exposed to film. This treatment ensured that qualitative levels of transcripts within a group could be assessed.

Table 4.2 Northern Blot Probe Sequences

Description	Oligonucleotide sequence (5' to 3')
Dm297 sense probe	gatgagtcttgctttaagggtaggccaatcttcgatgttcggaagtcccaaa
Dm297 antisense probe	ttgatttttagtcttaagctgagatccaaagaataaagtcgtgaaactatt
blood sense probe	aattcccaaatcaaatcggcaatattagcagcatttcctcagtagtcctcaga
blood antisense probe	gacactctgtagaggttaagcgggcagaaccgtttctgctactcgaagagat
mdg1 sense probe	tcccatcacactgacactctactcactcagatcgctttttttcttcataattgcc
mdg1 antisense probe	acaccctaatactaaatatgcgaattcagcatgtacgcctttaggggtcgcac
jockey sense probe	gcaaccttggtcctgaacgcttgctgaatatttgtgatgtgcctgctgaag
jockey antisense probe	cttcagcaggcacatcacaaatattcagcaagcgttcaggaccaaggttgc
juan sense probe	gtaggcaatgagatctggggttgatttccaaagagagcagatggagcgatg
juan antisense probe	catcgctccatctgctctctttggaaatcaaccccagatctcattgcctac

PolyA+/- Selection

Total RNA was fractionated into polyA+ and polyA− fractions using the MicroPoly(A) Purist Kit (Ambion AM1919). Fractionation was verified by RT-qPCR of known polyadenylated transcripts (Actin) and known transcripts that lack polyadenylation (18S ribosomal RNA).

Library Preparation, Sequencing, and Analysis

Ribosomal RNA Depletion.

The 28S, 18S, and 5S ribosomal RNAs (rRNAs) were depleted from 5 µg of each large RNA fraction using the Ribo-Zero Magnetic Kit (Epicentre). While this kit was designed for human/mouse/rat, it performs adequately for *Drosophila*. rRNA depletion was confirmed by RT-qPCR and validated using RNA 6000 Pico Bioanalyzer chips (Agilent). The 2S rRNA was depleted from the small RNA fraction according to Seitz *et al.* (2008) with the following modifications: 0.1 nM 2S rRNA complementary oligo was bound to 500 µg streptavidin beads in 1 ml 0.5× SSC for 1 hr. at 4°. The beads were then washed five times in 0.5× SSC followed by a 5-min incubation at 65° to remove secondary structure. A total of 2 µg of the small RNA fraction was diluted to 12.5 ng/µl and 160 µl was added to the bead slurry. The remaining steps of the protocol were as described (Seitz *et al.* 2008). Following rRNA depletion from both small and large RNA fractions, RNA integrity, and rRNA depletion were validated on a Bioanalyzer.

Large and Small RNA Fractionation.

Total RNA from 8 × 10^6 *Drosophila* Dmel-2 tissue culture cells was isolated using QIAzol Lysis Reagent (Qiagen). Total RNA was fractionated into large (>200 nt) and small (<200 nt) fractions using RNeasy Mini spin columns and RNeasy MinElute spin columns, respectively (Qiagen). DNA was removed from the large fraction by on-column DNase digestion (Qiagen). Fractionation and DNA removal were verified by RT-qPCR. RNA integrity and size fractionation were confirmed using small RNA and RNA 6000 Pico Bioanalyzer chips (Agilent).

RNA-Seq and Small RNA-Seq Library Preparation.

RNA-seq libraries were prepared in triplicate from 35 ng of the rRNA-depleted large RNA fraction using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB). Small RNA sequencing (smRNA-seq) libraries were prepared in triplicate from ~475 ng of the 2S rRNA-depleted small RNA fraction using the NEBNext Small RNA Library Prep Set for Illumina (NEB). Each small interfering RNA (siRNA)- and RNA-seq library was amplified with a primer having a unique barcode. The appropriate size of each library was validated on a Bioanalyzer using a high sensitivity DNA chip (Agilent) and quantitated using the Qubit dsDNA BR Assay Kit (Molecular Probes) according to the manufacturer's instructions. All siRNA-seq libraries were multiplexed and sequenced in one flow cell using a MiSeq and MiSeq Reagent Kit v2 (50-cycle) (Illumina). RNA-seq libraries were multiplexed and sequenced in two HiSeq lanes by the Genome Access Technology Center (GATC) at Washington University.

RNA Seq Library Analysis

All adapter sequences were trimmed and the libraries cleaned using Cutadapt (Martin 2011). We aggressively trimmed the siRNA reads to 25 nt from the 5' end following adapter removal to filter remaining rRNAs, small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNAs), and transfer RNAs (tRNAs) out of the dataset before mapping. All datasets were mapped to the *Drosophila melanogaster* genome and transcriptome using the RNA-seq Unified Mapper (RUM) (Grant *et al.* 2011). The NEB kit used to prepare the RNA-seq samples produces libraries with high directionality and RUM utilized this feature to strand specifically map the RNA-seq reads. RUM separated unique and nonuniquely mapping sequences into separate output files that could be

further analyzed.

The University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu, Dm6 assembly, August 2014) was used to visualize nonunique and unique bedgraph output files (Kent *et al.* 2002; dos Santos *et al.* 2015). The genome browser displays a peak normalized read count (reads per million, RPM) on the *y*-axis for the visualized genomic location.

Small RNA-Seq Library Analysis.

siRNAs were analyzed using a newly developed pipeline called SMACR (**S**equence **M**apping, **A**nnotation, and **C**ounting for s**R**NAs; https://github.com/mrmckain/SMACR). Raw reads were first trimmed using Trimmomatic v.0.33 (Bolger et al., 2014), with parameters optimized for siRNA data: Adapter trimming using TruSeq3-SE adapters, seed mismatch of 1, palindrome clip threshold of 20, and simple clip threshold of 7; a quality sliding window of 3 basepairs (bp) with a minimal average score of 20; and a minimum length of 19. Trimmed reads were then filtered to remove any longer than 30 bp. Relative abundances were then calculated for all unique trimmed reads. Unique is a read that is different from all others. The unique reads were then mapped to the *Drosophila melanogaster* genome (Dmel v.6.01; (Santos et al., 2015)) using bowtie v.1.1.1 (Langmead et al., 2009) allowing for either 0 or 1 mismatches. The mapping and read abundance information were then merged, estimating reads per million (rpm) for each mapped unique sequence. SMACR can simultaneously read in multiple experimental datasets, including replicates, and maintains each dataset as uniquely identified to the particular experiment and replicate. Annotation coordinates from Dmel

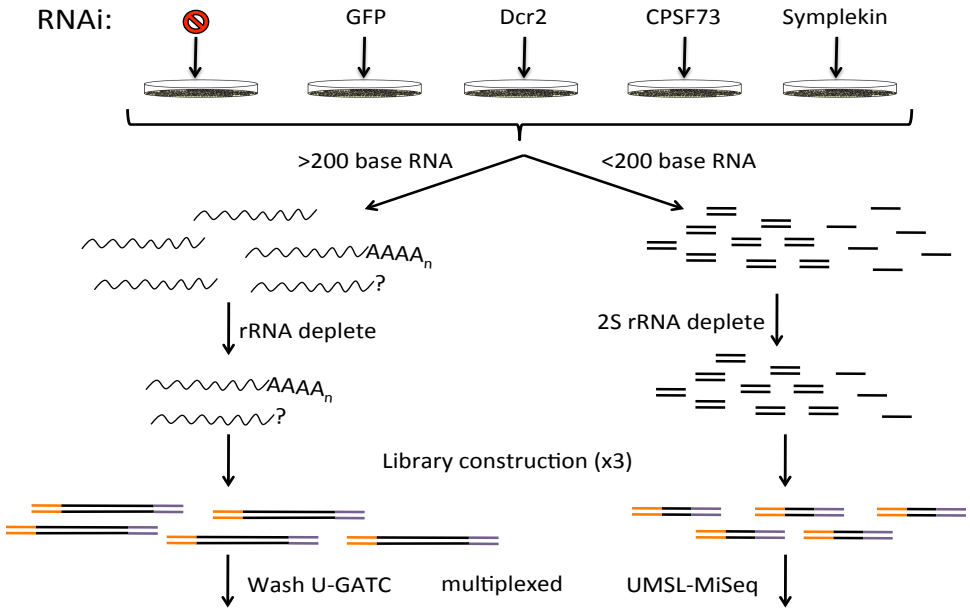
v.6.01 for miRNAs, noncoding RNAs, transposons, and two hairpin structures were used to link mapped siRNAs to annotation features. If a siRNA was found to map to more than one feature type, it was disregarded. Abundance (normalized read counts) of siRNAs mapping to a particular feature were totaled and percentages of siRNAs mapping to each feature were calculated for each replicate. 5' and 3' nucleotide abundance, siRNA abundance, and relative phasing to the core siRNA for a given mapping site were then analyzed in the final set of siRNAs. Averages include technical triplicates and the biological replicate and standard deviations reflect the standard error of the mean for all four samples. RNA-seq reads from Symplekin and CPSF73 depleted samples were strand specifically mapped using the RNA-seq Unified Mapper (RUM) (Grant et al., 2011) and visualized with the University of California Santa Cruz (UCSC) genome browser (http://genome.ucsc.edu, Dm6 assembly, August 2014) (Kent et al., 2002; Santos et al., 2015).

Small Capped RNA Data Analysis.

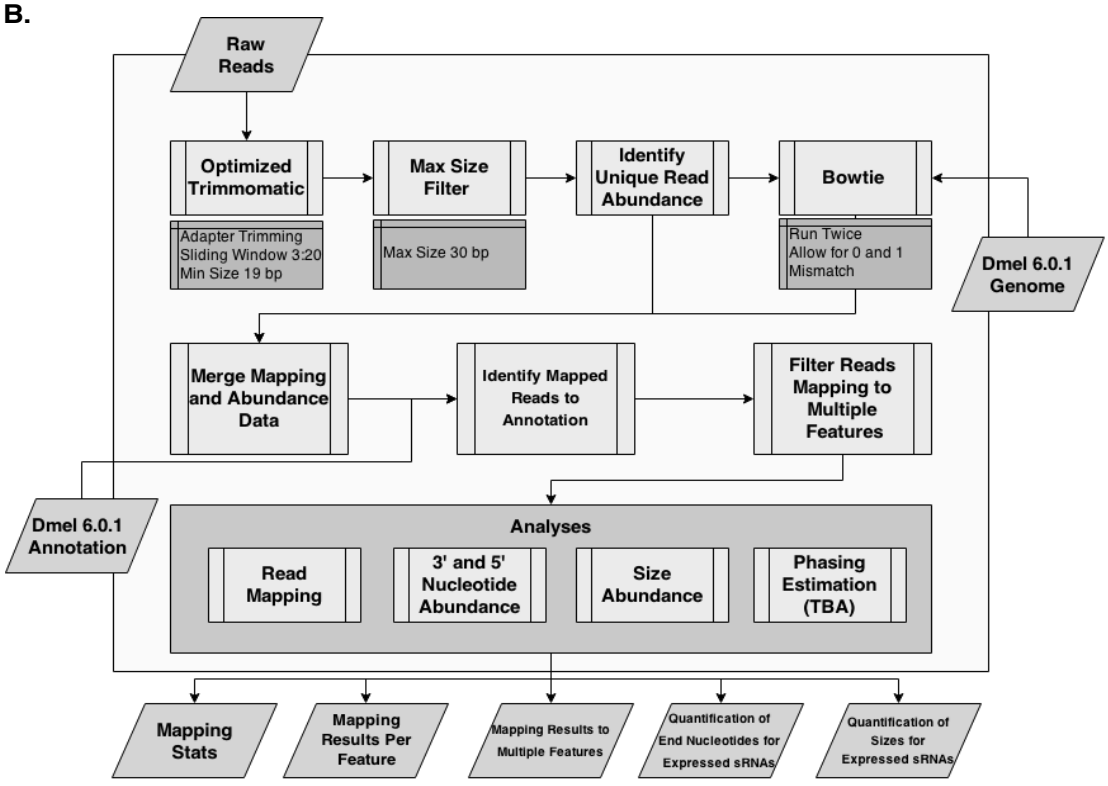
Small-capped RNA-seq datasets (SRA: SRP001584, SRR032457, and SRR032458) were obtained from the Gene Expression Omnibus (GEO) accession number GSE18643 (Nechaev *et al.* 2010). FASTQ files were mapped strand specifically using RUM to obtain nonuniquely mapping reads. The UCSC genome browser (www.genome.ucsc.edu) was used to visualize the bedgraph output files. For presentation in Figure 1 and Figure 2, screen captures of nonunique S and AS reads and tss mapping to full-length, representative (for Dm297, blood, juan, and jockey) or individual (for mdg1{}1720) Tns were taken and overlaid to scale.

Figure 5.1 Sequencing and SMACR Workflow

A.



B.



Construction of Stable S2 Cell Lines.

Gene Cloning and Plasmid Construction

Full-length CPSF73, CPSF100, and Symplekin were PCR amplified from Drosophila gene collection (DGC) clones (Open Biosystems) using specific primers. Dicer-2 full length gene was cloned from pGAD vector. The amplified genes were directionally cloned into pENTR D-TOPO (Invitrogen) to create Dcr-2, CPSF73, CPSF100, and Symplekin::pENTRD-TOPO. Proper gene insertion was confirmed via restriction digest and Sanger sequencing. Dcr-2, CPSF73, CPSF100, and Symplekin::pENTRD-TOPO were each re- combined with pAHW destination vectors (Drosophila Gateway vector collection, Carnegie Institution for Science) using Clonase II (Life Technologies). Proper gene recombination was confirmed via restriction digest and Sanger sequencing.

Creation of Stable Drosophila Dmel-2 Tissue Culture Lines

Full-length Dcr-2 and CCC factor::pAHW were transfected into Drosophila Dmel-2 cells with Effectene Transfection Reagent (Qiagen) according to the manufacturer's protocol. The pCoBlast vector (Invitrogen) containing a Blasticidin resistance gene was cotransfected to enable selection of successfully transfected cells. Cells were grown in SF-900 II SFM (Gibco) and maintained at 27°C under normal atmospheric conditions. Forty-eight hours post-transfection, Blasticidin (25 μg/mL) was added to the media to select for stably transfected cells. Cells were split and passaged into fresh selective media every 5 days to a centration of 1×10^6/mL. Stable transfection was confirmed via Western blot with anti-HA antibody (Covance).

Transient Knock Down of Target Proteins via RNAi

RNAi was performed as described (Sullivan et al., 2009). Briefly, DNA from a sequence of the target gene was amplified by PCR using primers that contain a T7 recognition sequence. This tag allows the DNA to be transcribed to RNA. With tags on both ends of the dsDNA of interest, the T7 polymerase is able to make complementary strands of RNA. After the transcription reaction, the DNA is degraded by addition of RQ1 DNAse (Promega) at a concentration of 1 Unit/1ug of original DNA, incubated for 30 minutes at 37 degrees, then heated to 95 degrees to denature the remaining RNA. The transcription reaction is then allowed to slowly come to room temperature, ensuring the correct Tm will at some point be reached for the complementary RNA strands to hybridize. To induce knock down of target proteins, 10 ug of dsRNA was added to 1 million cells in a 6 well plate. 10ug dsRNA was administered to the cells again on day two. On day three, cells were split, equal volumes of fresh media was added to each well, and 10 ug dsRNA was added to each well. The cells are left to recover for one day, and then on Day 5 they are harvested. To perform knockdown on 1 million cells, it requires 40ug dsRNA.

Crude Nuclear Extract

Crude nuclear extracts were prepared as described (Sullivan et al., 2009) and used for Immunoprecipitation assays. Briefly, approximately 10 million cells were pelleted and washed with cold 1X PBS. Cells were then resuspended in 500 ul Hypotonic Buffer A, protease inhibitor, and incubated on ice for 30 minutes. Cells were then lysed by passage through a 27 gauge needle 20X. Cells were spun, supernatant was removed and discarded, and remaining cell pellet was resuspended in 50% cell pellet volume of Low

Salt Buffer C. Equal volume of High Salt Buffer C was added drop by drop while cells were being mixed with a stirbar on ice. Cells let spin for 30 minutes on ice then spun full speed, 4 degrees for 5 minutes. Supernatent was removed and dialized in Buffer D over night. Buffers can be found in Table 4.3**.**

Refined Nuclear and Cytoplasmic Extract

Marzluff, Adelman, Lamonde (MAL) Nuclear and Cytoplasmic Fractionation Protocols.
The following protocols are designed to fractionate S2 cells into extremely clean nuclear and cytoplasmic fractions that can be usd for both protein and RNA/DNA applications. It is a combination of our labs "crude" nuclear extract protocol from the Marzluff lab in conjunction with elements of protocols gleaned from protocols out of the Adelman and Lamonde labs. It employs the use of a sucrose cushion to separate the nuclei from the rest of the lysed cells. The first protocol is designed for IP and protein expression analysis while the second is designed for RNA and DNA isolation. This method was developed by myself, and as such, the full protocol is listed in step by step fashion as it does not exist in other publications or sources.

MAL Protocol for IP and Protein Expression Analysis.

1. 1x10^9 cells collected, pelleted and washed 2X with cold 1X PBS. Resulted in 750ul pellet.
2. Cells were resuspended in 3.75 mls Hypotonic Buffer A, (5X cell pellet volume) and 37.5 ul 100X protease inhibitor cocktail. Let swell on ice for 30 minutes
3. Cells lysed by passing through 15ml 'tight' dounce homogenizer 20X (this is pestle B)

4. Transfer cell lysate into a 15 ml conical tube and spun at 3000 rpm at 4 degrees 10 minutes to pellet “crude” nuclei. Remove crude cytoplasmic fraction (this is the clearish top stuff, not the cell pellet) it should be about 2.5 mls. Try not to get any of the pellet. Transfer the crude cytoplasmic fraction to a 1.7 ml eppendorf tube and spin 13K 4 degrees for 10 minutes to pellet any cellular debris. Remove the supernatant from the tube being careful not to touch any of the pellet. This is the “clean” cytoplasmic fraction. Put this in the -20 until ready for dialysis. You should recover about 2 mls of Cytoplasmic fraction. When nuclear prep is finished, remove cytoplasmic fraction and dialize in 50 ml conical slide-a-lyzer overnight, then with an additional change of buffer in the morning for 4 hours.
5. Take the crude nuclear from step 4 and resuspend in 2 mls of Buffer S1. Buffer S1 = 20 mM HEPES, .88 M Sucrose, 5mM MgCl2, .5DTT)
6. Layer this on top of 15 mls room temperature Buffer S2 (20 mM HEPES, 2M sucrose, 5mM MgCL2, .5mM DTT added just prior to use.) Buffer S2 must be allowed to come to room temperature or the nuclei will be unable to pellet through the cushion. This is done in the 30 ml sorval S34 tubes. You should have a tube with 15mls clear S2 and about 3 mls of cloudy S1/cell pellet
7. Carefully transport tubes without mixing the layers to the high speed centrifuge. Spin 12.5K RPM in a 34S Sorval centrifuge for 30 minutes at 4 degrees.
8. Remove all the sucrose supernatent (both layers) being careful to make sure all the un-pelleted cellular debris does not touch the nuclear pellet. Nuclear pellet

will be very small, and look like a smear of brown debris along the bottom/inside of the tube wall. Resuspend nuclear pellet in 500 ul buffer A. Spin this at 13K rpm for 5 minutes at 4 degrees. Remove buffer A, resulting in a nuclear pellet of approximately 200 ul. Resuspend the nuclei in 100 ul Low Salt Buffer C. 100 ul High Salt Buffer C was then added drop by drop to cells on stir plate. This was left on stir plate for 30 minutes on ice to lyse cells.

9. Spin this at 13K rpm at 4 degrees for 10 minutes to pellet nuclei debris and remove the supernatant. This is the "clean" nuclear pellet. Remove approximately 175 ul nuclear lysate. Dialize this over night in Buffer D, with a change of buffer in the morning followed by 4 more hours of dialysis.

MAL Protocol for RNA Extraction

1. Grow cells to density of 4 million/ml. 800 million cells collected, spun at 2000 g for 2 minutes to pellet and washed 2X with 50 mls cold 1X PBS.
2. Cell pellet was approximately 600 ul, cells were resuspended in 3 mls Hypotonic Buffer A, (5X cell pellet volume), 30 ul 100X protease inhibitor cocktail and 40 ul Ribolock . Let swell on ice for 30 minutes.
3. Cells lysed by passing through 15ml 'tight' dounce homogenizer 20X. This is pestle "B"
4. Cells were spun at 3000 rpm at 4 degrees 10 minutes to pellet "crude" nuclei. This results in a "crude" nuclear pellet consisting of cellular debirs, nuclei, and unlysed cells and a "crude cytoplasmic" fraction consisting mostly of cytoplasm.

"Crude Cytoplasmic" fraction was removed (~1 ml per tube) and spun 13K 4 degrees for 10 minutes to remove any residual cellular debris.

5. Approximately 750 ul of pure cytoplasmic fraction was removed from each tube and combined together. 500 ul of the pure cytoplasmic fraction was removed immediately and frozen for future protein (western) analysis. The remaining 1000 ul was used immediately in a liquid trizol extraction.
    a. Pure cytoplasmic fraction was transferred to a 15 ml falcon tube and 3 volumes trizol was added to (3000ul) and mixed thoroughly by inversion.
    b. Incubate at room temperature for 5 minutes.
    c. Add .2 volume chloroform and vortexed (600 ul). Then incubate for 2 minutes at room temperature. This was split into 4, 1.7ml eppis, spun full speed at 4 degrees for 10 minutes. Aqueous phase was removed and saved as the cytoplasmic RNA fraction. Organic phase was discarded.
6. While the crude cytoplasmic lysate was spinning (Step 4) to pellet residual cellular debris, each crude nuclear pellet was resuspended in 2 mls of Buffer S1.
7. Each pellet/S1 resuspension was layered on top of 15 mls Buffer S2 (20 mM HEPES, 2M sucrose, 5mM MgCL2, .5mM DTT added just prior to use.) in a S34 plastic sorval centrifuge tube. To layer the S1 suspension, take a 1000ul pipette and drop by drop add the S1 solution on top of the S2 cushion. It should form a visible layer on top. As is the case in the protein extraction protocol, buffers S1 and S2 must be allowed to come to room temperature.

8. Sucrose cushion/nuclear lysate was carefully transferred to a room temperature sorval S34 rotor and spun at 12.5K RPM, 4degree centrifuge for 30 minutes.
9. Sucrose cushions were removed with an auto-pipette, and nuclear pellet was resuspended in 500 ul 1X PBS. The pure nuclear pellet will be very small and stuck firmly to the side of the sorval tube. The 500 ul PBS and pure nuclear pellet from both tubes were combined, then split in half again, then tubes spun at 4 degrees, 12K rpm for 5 minutes. This resulted in two nuclear pellets that were effectively exactly the same. One nuclear pellet was subjected to RIPA protein extraction and the other was subjected to Trizol RNA prep. The RIPA protein extraction is necessary in order to validate the fractions by western blot.
    a. **RIPA Protein extraction**. PBS was removed from the nuclear pellet and 100 ul of RIPA, 1 ul HALT protease inhibitor cocktail was added. This was mixed thoroughly by pipette.
    b. This was rotated at 4 degrees for 10 minutes.
    c. Spin full speed for 10 min, 4 degrees. Remove and save the supernatant, discard residual pellet.
    d. Resulted in ~90ul of lysate with a protein concentration of .65ug/ul total protein.
    e. **Trizol RNA prep**: 1 ml tri-reagent was added to the pellet to extract nuclear RNA.
    f. Let sit on bench 5 minutes, 200 ul chloroform was added, vortexed, incubated 2 minutes

g. Spun full speed at 4 degrees for 10 minutes

h. Aqueous phase removed and saved, organic phase was discarded.

10. Store samples a -80 until ready for downstream applications. Figure 4.2 details the procedure.

2.5 Molar Sucrose Stock Preparation

Weigh out appropriate amount of sucrose for a 2.5 molar concentration. Heat 75% of the water you will use on a stirplate until almost boiling. Slowly add the sucrose to the hot water. This much sucrose will not go into solution unless it is boiling. Qs to the final volume – sucrose takes up a lot of space so you need to start with only 60-70 percent of the total volume of water. Or if you overshoot, back calculate the sucrose concentration. Make approximately 200 mls at a time.

Validation of Fractions

Fractions were validated a number of ways. First, protein expression of equal amounts of total protein in the nuclear and cytoplasmic fractions were analyzed via western blot for the presence of Histone H3 protein and MEK1/2. H3 is exclusively nuclear while MEK1/2 is a cytoplasmic transcription factor. An example of Whole cell lysate and the refined nuclear and cytoplasmic fractions can be seen below. In addition to western blot, the fractionation can be assessed by PCR/qPCR of the nuclear and cytoplasmic fractions. The cytoplasmic fraction should contain no DNA.

Figure 4.2 General Workflow of the MAL Prep.

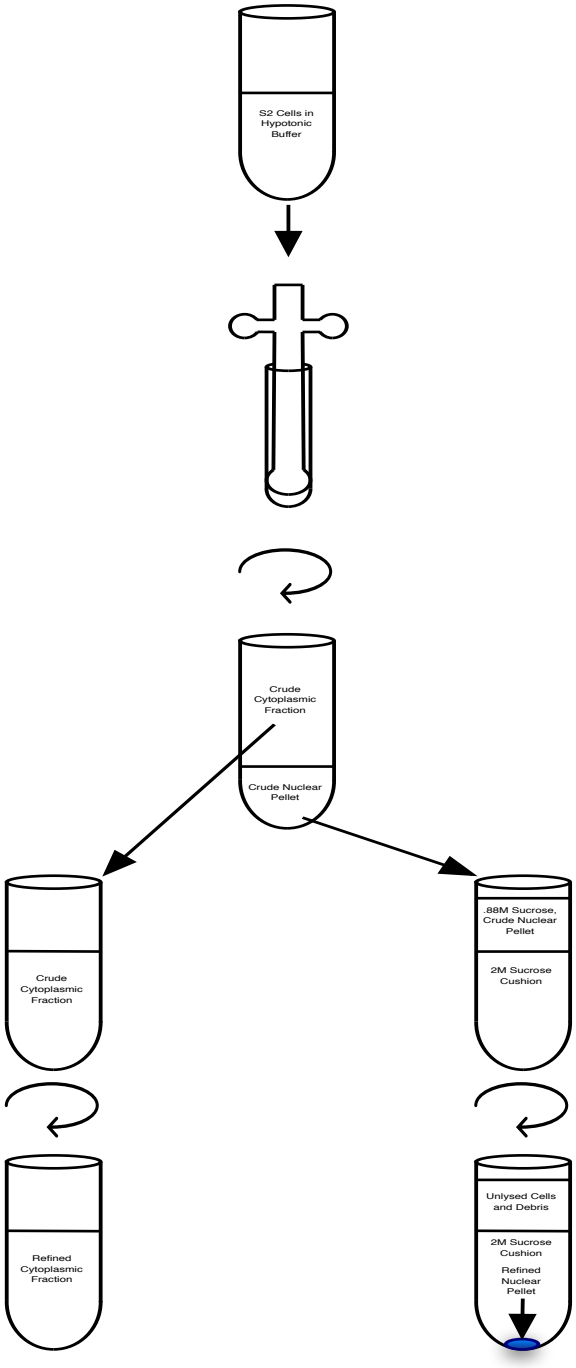


Table 4.3 Buffers Used in Crude and Refined Nuclear and Cytoplasmic Extracts.

Buffer A (Hypotonic)
10 mM Hepes/KOH, Ph7.9
1.5 mM MgCl2
10mM KCL
.5mM DTT (fresh)

Buffer S1
20 mM HEPES pH 7.9
.88 M sucrose
5mM MgCl2
.5mM DTT (fresh)

Buffer C (Low Salt):
20 mM Hepes/KOH Ph 7.9
25% glycerol
1.5 mM MgCl2
.2mM EDTA
.5mM DTT (Fresh)
.02M NaCl

Buffer S2
20 mM HEPES pH 7.9
2M Sucrose
5 mM MgCl2
.5 mM DTT (fresh)

Buffer C (High Salt):
20 mM Hepes/KOH Ph 7.9
25% glycerol
1.5 mM MgCl2
.2mM EDTA
.5mM DTT (Fresh)
1.2M NaCl

Buffer D:
20mM Hepes/KOH, pH 7.9
20% glycerol
.1M KCL
.2mM EDTA
.5mM DTT (Fresh)

Immunoprecipitation, Western Blotting, S1 Nuclease Assay

Immunoprecipitation of HA-tagged and endogenous proteins was performed as described (Sullivan et al., 2009) using 100 mg of crude nuclear or 175 mg refined cytoplasmic or nuclear extracts. S1 nuclease protection assay was performed as described (Michalski and Steiniger, 2015). Monoclonal and polyclonal HA antibodies (Cat#s MMS-101R and PRB-101C, respectively, Covance) were used for both IP (3 μL) and WB (1:1000). Anti-CPSF73, anti-Symplekin, and anti-CPSF100 antibodies (1:1000) were described previously (Sullivan et al., 2009; Yang et al., 2009). Commercial anti-Dcr-2 (Abcam ab4732), anti-Actin (Abcam ab8227), anti-H3 (Cell Signaling 4499), and anti-MEK1/2 (Cell Signaling 8727) were used at manufacturer recommended concentrations. The anti-R2D2 antibody was a generous gift from the Siomi lab (Nishida et al., 2013).

RT-qPCR from Nuclear and Cytoplasmic Fractions

Nuclear/cytoplasmic enrichment analysis of precursors and esiRNAs by RT-qPCR used Trizol prepped total RNA from the refined fractionation. All samples were column cleaned using the Qiagen miRNeasy Mini Kit (217004) and DNAse treated (Ambion Turbo DNAse # AM 1907) prior to RT. Equal cellular volumes were used in the RT step. RT-qPCR of precursors utilized iScript Reverse Transcription Supermix and SsoAdvanced Universal SYBR Green (Biorad #170884, #1725271, respectively). siRNA RT-qPCR was performed using Taqman Micro RNA RT Kit and Taqman Universal Master Mix (AB #4366596, #4440040, respectively.) Custom small RNA assay numbers and PCR primers are listed in table 4.4. All qPCR experiments were performed in triplicate.

Table 4.4 Primer List for Small RNA, Transposons and Controls

Taqman Assays

Small RNA Target	Assay ID number
2S	001766
esi2.1	CSN1KEF
esi1.2	CSPACQN
Dm297	CSAAYTW
mdg1	CSBJWZ4
Mir2a	000261

Hairpin and Transposon Precursor Primers

Target Sequence	5' to 3' Sequence	Position (bp)	Reference
Dm297-RT Forward	ggcagacagagacggag	4629:4645	Russo et. al 2016
Dm297-RT Reverse	cgacttcttcttcttcaagc	4673:4693	Russo et. al 2016
Dm297-env Forward	gacaccactatacacaccac	6269:6289	Russo et. al 2016
Dm297-env Reverse	ctcaataatgtcgttgtg	6317:6335	Russo et. al 2016
Blood-ORFII Forward	cgtaaaaggcgaatcgcctg	2534:2554	Russo et. al 2016
Blood-ORFII Reverse	gctgcttacgcatactgtc	2624:2643	Russo et. al 2016
Blood-RT Forward	cctataccaacagatgccgac	4647:4668	Russo et. al 2016
Blood-RT Reverse	caaagcctcgttaagtggcg	4726:4746	Russo et. al 2016
Mdg-ORFII Forward	ctgagatcggtgaggatatcg	2053:2074	Russo et. al 2016
Mdg-ORFII Reverse	cgggtaatgttattaccgctg	2133:2154	Russo et. al 2016
Mdg-RT Forward	gtaaacaagcatgtggagcg	4824:4844	Russo et. al 2016
Mdg-RT Reverse	ctcctgctctgtagtggac	4923:4942	Russo et. al 2016
Jockey-gag Forward	acctatcctcacccttctc	776:795	Russo et. al 2016
Jockey-gag Reverse	tgctccatattctccgtttcag	919:897	Russo et. al 2016
Jockey-RT Forward	gtggacattgataatgccacaag	2841:2864	Russo et. al 2016
Jockey-RT Reverse	ggaagttgaagtggtctgaag	2922:2943	Russo et. al 2016
Juan-ORFI Forward	ctgtgagttctacacgtacgatac	499:522	Russo et. al 2016
Juan-ORFI Reverse	cctaggttttgtagcatggatttg	586:609	Russo et. al 2016
Juan-RT Forward	gcgcaatgtaaaaacatatccg	2082:2104	Russo et. al 2016
Juan-RT Reverse	ctgtgagcagttgacaaccac	2168:2189	Russo et. al 2016
AY119029 (esi2.1) F	ccagggcgctacattcaata	multiple	Marques et. al 2010
AY119029 (esi2.1) R	caaacacccacacacatacaca	multiple	Marques et. al 2010
CG18854 (esi1.2) F	caaggctagggctcgtca	multiple	Marques et. al 2010
CG18854 (esi1.2) R	ggtgctgcgcataccttt	multiple	Marques et. al 2010

Additional Primers

Target Sequence	5' to 3' Sequence	Reference
GAPDH F	CGTTCATGCCACCACCGCTA	Russo et. al 2016
GAPDH R	CCACGTCCATCACGCCACAA	Russo et. al 2016
sop '3 UTR F	GGATTGCTACACCTCGGCCCG	Tatomer et. al 2014
sop '3 UTR R	CTACAACAGAATCTCCAAATCGACC	Tatomer et. al 2014

Immunofluorescence

Immunofluorescence was performed essentially as described (S. L. Rogers and G. C. Rogers, 2008). Nuclei were stained with DAPI. Anti-Symplekin (Sullivan et al., 2009) was used at 1:500 and anti-Dcr-2 (Miyoshi et al., 2009) was used at 1:200. Secondary antibodies were used at 1:1000. Images were obtained on a Zeiss LSM 700, maintaining equal laser strength, gain and 1 AU. The images were processed with ImageJ.

RNA Extraction from Fly Heads for RT-qPCR

Collect approximately 10 flies. Flies should be kept on a CO2 plate at all times to keep them unconscious. Place 100 ul of 1X PBS on the dissection polymer and place two unconscious flies into the PBS droplet. Remove head by holding body with tweezers and snipping off head with scissors. Place severed head into eppendorf tube filled with 40 ul cold PBS. After all heads have been collected, extract RNA using standard Trizol protocol for homogenized tissue.

CHAPTER 5: DISCUSSION

Transposable elements comprise approximately 44% of the human, and 30% of the *Drosophila* genome (Goodier, 2016). As such, examining their regulation and the role they play is critical to our understanding of cellular development and viability. There are no known active DNA transposons in humans and very few active retrotransposons. However, there has been recent evidence that a small handful of LTR retrotransposons are active and suppressed by naturally occurring small RNA in humans (Yang & Kazazian, 2006). *Drosophila* is an intriguing model system in which to study the regulation of transposons as many of the *Drosophila* transposons are active (Kofler, Nolte, & Schlötterer, 2015). By studying the mechanism of their regulation in flies, it is hoped that a greater understanding of how these elements are regulated, and more importantly how they become de-regulated, will be achieved. The main function of esiRNAs in somatic cells is to repress retrotransposons (Fagegaltier et al., 2009). Previous research by Dr. Steiniger found a connection between the esiRNA associated protein Dcr-2, and the 3' end-processing factor Symplekin. It was this connection that was the initial impetus behind the work discussed in this dissertation.

Regulation of transposable elements in *Drosophila*

In 2008, multiple labs published papers detailing the discovery of a new class of small RNA, esiRNA (Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Okamura & Lai, 2008). These small RNAs were found to be Ago2 dependent and mapped back to both retrotransposons and structured loci (hairpins). The production of these small RNAs requires a dsRNA substrate that is cleaved by Dcr-2. Hairpins are made

up of many repeat elements, allowing for the formation of dsRNA secondary structure from a single transcript. However, retrotransposons do not possess the same inverted repeats, and thus the mechanism by which the dsRNA substrate could be formed remained a mystery. From the analysis of RNA-seq libraries that I created, we were able to begin to understand how retrotransposons could form the necessary dsRNA substrate. We identified convergent transcription of both sense and antisense transcripts for many of the retrotransposons we examined. Furthermore, analysis of short-capped RNA-seq data allowed us to identify transcriptional start sites that would allow this convergent transcription to occur. We then sought to understand the polyadenylation status of these transcripts. By fractionating RNA into PolyA+ and PolyA- fractions, we were was able to characterize the PolyA status of both the sense and antisense transcripts via strand specific RT-qPCR. In most cases the antisense transcript was poorly polyadenylated, but the sense transcript was heavily polyadenylated. Polyadenylation is necessary for proper nuclear export and since at least one transcript from the retrotransposon lacks proper polyadenylation, they would most likely be retained in the nucleus (Dower, Kuperwasser, Merrikh, & Rosbash, 2004). Additionally, the antisense transcript was lower in abundance, suggesting it would be the limiting factor in dsRNA formation. In later experiments, I showed that depletion of Symplekin results in greater nuclear levels of retrotransposon precursors, supporting the connection between polyadenylation and nuclear retention. If esiRNAs mapping to this locus were being cleaved from convergently derived dsRNA precursors, they would have to be Dcr-2 dependent. To address this, I created more libraries from S2 cells in which

Dcr-2 had been depleted. It was observed that in these cells, levels of both sense and antisense precursors go up, while the levels of small RNAs derived from these elements go down. One of the most striking things about this data, and one that supports our view of convergent transcriptional events giving rise to the dsRNA substrate is that small RNAs could be mapped back to the entirety of the retrotransposon, not just the LTRs. Additionally, small RNAs that map back to the non-LTR retrotransposons were also observed. Though this provides an exciting glimpse into the mechanism by which retrotransposons may regulate themselves, future experiments will focus on elucidating these details.

The work described herein details the mechanism by which retrotransposons can form the dsRNA substrate necessary for their regulation, however it does nothing to address the regulatory mechanisms of the other class of transposable elements, TIR transposons. Through bioinformatic analysis, I found that TIR transposons are regulated differently than their retrotransposon counterparts. I found that the ratio of sense to antisense transcription is very high, suggesting these elements may not be regulated by small RNA. As suspected, the amount of small RNAs that mapped to TIR transposons was very low. Another difference between these two classes of elements was the inability of TIR transposons to produce the necessary protein for their movement. Only one *Pogo* TIR transposon was identified that could conceivably produce a functional transposase. Thus, I hypothesize that unlike retrotransposons, regulation of TIR transposons is not mediated by small RNA, but more likely the lack of a functional transposase.

Dicer-2-CCC interaction

3’ end processing is a co-transcriptional event occurring in the nucleus (Sullivan, Steiniger, & Marzluff, 2009). This processing is accomplished, in part, by a trio of proteins: Symplekin, Cpsf73, and Cpsf100. These three proteins form the Core Cleavage Complex (CCC) and are integral for proper processing of both canonical and histone mRNAs (Sullivan et al., 2009). After confirming the interaction of Symplekin and Dcr-2 by reciprocal IP experiments, we sought to further characterize this interaction. Through a number of IP experiments using stably expressing Symplekin constructs and RNAi depletion of other CCC factors, we confirmed the interaction with Dcr-2 was direct, and that it was occurring on the N-terminus of Symplekin. Dcr-2 has been shown to associate with transcriptional machinery (Cernilogar et al., 2011) in the nucleus, and since the CCC is known to be a nuclear complex, we sought to further investigate the interaction of Dcr-2 with the CCC in terms of subcellular location. To this end, I designed a novel nuclear fractionation technique and was able to confirm a nuclear, but not cytoplasmic, interaction between Symplekin and Dcr-2.

Role of the CCC in esiRNA biogenesis

Due to the nuclear localization of the Dcr-2-CCC interaction, we first sought to determine if Dcr-2 had any role in 3’ end processing. RT-qPCR, S1 nuclease protection assays, and co-depletion observations revealed Dcr-2 to have no role in 3’ end processing. To determine what, if any, role Symplekin plays in esiRNA biogenesis and the regulation of their targets, I performed HTS on both large and small pools of RNA from cells in which either Dcr-2, or CCC members had been RNAi depleted. My data shows that depletion of Dcr-2 decreases the amount of small RNAs produced from both

retrotransposons and hairpins. However, depletion of CCC components increased the levels of retrotransposon-derived esiRNAs (resiRNA) while decreasing the levels of hairpin-derived esiRNA (hesiRNA). Further analysis of the RNA-Seq data revealed that for the retrotransposons, transcription of the antisense strand was dramatically increased. From my previous work, we show that antisense transcription is the limiting factor in the formation of resiRNA substrate. Thus, increases in antisense transcription would lead to more substrate, which in turn could lead to a greater abundance of small RNA. Hairpin precursors did not show a marked change with depletion of the CCC, however they did show 3' end misprocessing. 3' end processing has been shown to cause deficiencies in nuclear export, and if the hairpin precursors are being processed in the cytoplasm, their nuclear retention would explain the decreased levels of hesiRNAs without the corresponding decrease in the precursor.

Subcellular location of precursors

If retrotransposons are processed in the nucleus and hairpins are being processed in the cytoplasm, it would be expected that differential levels of these transcripts could be observed if a sufficiently pure nuclear and cytoplasmic fractionation of S2 cells could be achieved. To address this, I employed a the MAL fractionation technique and performed RT-qPCR on retrotransposons and hairpins, as well as their small RNAs. I found that the retrotransposon precursors were overwhelmingly nuclear, while the hairpin precursors were more cytoplasmic (similar to our GAP controls). This finding was mirrored in the small RNA RT-qPCR data. If the CCC effects observed on hesiRNAs were in fact due to 3' end misprocessing, then CCC knockdown should

increase the levels of these transcripts in the nucleus. Results from this experiment show that in fact, CCC depletion causes an increase in nuclear retention of the hairpin precursor. Interestingly, depletion of CCC components also causes an increase in the nuclear levels of the retrotransposon precursor as well. However, since the RT-qPCR used in this experiment is not strand specific, it is difficult to tell whether this is due to increased antisense transcription as previously observed or accumulation of the sense strand due to improper nuclear export.

Physical differences between resiRNAs and hesiRNAs

Due to the observed differences between retrotransposon and hairpin precursors regarding CCC depletion effects and localization, we sought to find any other differences between the small RNAs derived from them. Through bioinformatic analysis it was discovered that hesiRNAs are more evenly distributed over 21-23 nucleotides whereas the resiRNAs were predominantly 21 nucleotides. This observation suggests that processing is more stringent for the resiRNAs, however this has yet to be investigated. Further analysis also revealed preferences in the 3' and 5' base depending on the nature of the precursor. ResiRNAs preferred adenosine at the 3' end and adenosine or cytosine at the 5' end while hesiRNAs preferred a 3' guanine and a 5' cytosine. Taken together, this suggests that differences in the precursors due to either structure or location may have implications regarding the nature of the small RNAs they produce.

Conclusion

I hypothesize that the observed differences between retrotransposon and hairpin substrates and the small RNAs derived from them could be attributed to their secondary structure. In the first project detailed in this dissertation, I show convergent sense and antisense transcription of retrotransposons, but only sense transcription of hairpins. This would allow the retrotransposons to form near perfect dsRNA, while the hairpins would have a more variable and imperfect secondary structure. Furthermore, since the antisense retrotransposon transcript is poorly polyadenylated, one side of the resiRNA substrate would be blunt while the hairpin would most likely have bulges or frayed ends. In vitro data by the Bass lab has recently shown that Dcr-2 is able to cleave blunt ended dsRNA substrates without the addition of cofactors, but requires Loqs-PD to cleave substrates internally or substrates with frayed ends (Sinha, Trettin, Aruscavage, & Bass, 2015). Loqs-PD is one of four isoforms of the loquacious protein and is reported to have a cytoplasmic localization (K. Miyoshi, Miyoshi, Hartig, Siomi, & Siomi, 2010). Knockdown of Loqs-PD has been reported to decrease global levels of esiRNAs (Zhou et al., 2009). However, closer examination of the Zhou et al. data revealed resiRNAs were unaffected. Lastly, overexpression of Loqs-PD isoform is shown to increase levels of the hairpin derived esiRNA CG4068 (Marques et al., 2010). Taken together, this work suggests that retrotransposon dsRNA substrate is retained in the nucleus, potentially because of its secondary structure. Being blunt ended and perfectly complimentary, it could be cleaved by Dcr-2 without accessory proteins. In contrast, the hairpin precursors, resembling a more canonical mRNA, are exported to the cytoplasm where they are processed with the help of Loqs-PD.

Small RNA pathways are traditionally defined by the proteins associated with them. Here, I show that subdivision of the esiRNA pathway into resiRNA and hesiRNAs may be warranted based on substrate structure, location, and potentially different proteins involved. Future work will center on teasing out these differences in greater detail, especially with respect to differential protein involvement.

References

Cernilogar, F. M., Onorati, M. C., Kothe, G. O., Burroughs, A. M., Parsi, K. M., Breiling, A., et al. (2011). Chromatin-associated RNA interference components contribute to transcriptional regulation in Drosophila. *Nature*, *480*(7377), 391–395. http://doi.org/10.1038/nature10492

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., et al. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature*, *453*(7196), 798–802. http://doi.org/10.1038/nature07007

Dower, K., Kuperwasser, N., Merrikh, H., & Rosbash, M. (2004). A synthetic A tail rescues yeast nuclear accumulation of a ribozyme-terminated transcript. *RNA (New York, N.Y.)*, *10*(12), 1888–1899. http://doi.org/10.1261/rna.7166704

Fagegaltier, D., Bougé, A.-L., Berry, B., Poisot, E., Sismeiro, O., Coppée, J.-Y., et al. (2009). The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(50), 21258–21263. http://doi.org/10.1073/pnas.0809208105

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., et al. (2008). Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science (New York, N.Y.)*, *320*(5879), 1077–1081. http://doi.org/10.1126/science.1157396

Goodier, J. L. (2016). Restricting retrotransposons: a review. *Mobile DNA*, *7*(1), 16. http://doi.org/10.1186/s13100-016-0070-z

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., et al. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, *453*(7196), 793–797. http://doi.org/10.1038/nature06938

Kofler, R., Nolte, V., & Schlötterer, C. (2015). Tempo and Mode of Transposable Element Activity in Drosophila. *PLoS Genetics*, *11*(7), e1005406. http://doi.org/10.1371/journal.pgen.1005406

Marques, J. T., Kim, K., Wu, P.-H., Alleyne, T. M., Jafari, N., & Carthew, R. W. (2010). Loqs and R2D2 act sequentially in the siRNA pathway in Drosophila. *Nature Structural & Molecular Biology*, *17*(1), 24–30. http://doi.org/10.1038/nsmb.1735

Miyoshi, K., Miyoshi, T., Hartig, J. V., Siomi, H., & Siomi, M. C. (2010). Molecular mechanisms that funnel RNA precursors into endogenous small-interfering RNA and microRNA biogenesis pathways in Drosophila. *RNA (New York, N.Y.)*, *16*(3), 506–515. http://doi.org/10.1261/rna.1952110

Okamura, K., & Lai, E. C. (2008). Endogenous small interfering RNAs in animals. *Nature Reviews. Molecular Cell Biology*, *9*(9), 673–678. http://doi.org/10.1038/nrm2479

Sinha, N. K., Trettin, K. D., Aruscavage, P. J., & Bass, B. L. (2015). Drosophila Dicer-2 Cleavage Is Mediated by Helicase- and dsRNA Termini-Dependent States that Are Modulated by Loquacious-PD. *Molecular Cell*, *58*(3), 406–417. http://doi.org/10.1016/j.molcel.2015.03.012

Sullivan, K. D., Steiniger, M., & Marzluff, W. F. (2009). A core complex of CPSF73, CPSF100, and Symplekin may form two different cleavage factors for processing of poly(A) and histone mRNAs. *Molecular Cell*, *34*(3), 322–332. http://doi.org/10.1016/j.molcel.2009.04.024

Yang, N., & Kazazian, H. H. (2006). L1 retrotransposition is suppressed by endogenously

encoded small interfering RNAs in human cultured cells. *Nature Structural & Molecular Biology*, *13*(9), 763–771. http://doi.org/10.1038/nsmb1141

Zhou, R., Czech, B., Brennecke, J., Sachidanandam, R., Wohlschlegel, J. A., Perrimon, N., & Hannon, G. J. (2009). Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. *RNA (New York, N.Y.)*, *15*(10), 1886–1895. http://doi.org/10.1261/rna.1611309

PERMISSIONS

Genetics Society of America LICENSE TERMS AND CONDITIONS

Nov 03, 2016

This is a License Agreement between Andrew W Harrington ("You") and Genetics Society of America ("Genetics Society of America") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Genetics Society of America, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3981490570410
License date	Nov 02, 2016
Licensed content publisher	Genetics Society of America
Licensed content title	Genetics
Licensed content date	Dec 31, 1969
Type of Use	Thesis/Dissertation
Requestor type	Academic institution
Format	Electronic
Portion	chapter/article
Title or numeric reference of the portion(s)	All text, figures and supplemental information from the requested article
Title of the article or chapter the portion is from	Antisense Transcription of Retrotransposons in Drosophila: An Origin of Endogenous Small Interfering RNA Precursors.
Editor of portion(s)	N/A
Author of portion(s)	Andrew Harrington
Volume of serial or monograph.	N/A

Page range of the portion	
Publication date of portion	January 1, 2016
Rights for	Main product and any product related to main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	yes
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	yes
The lifetime unit quantity of new product	Up to 499
Made available in the following markets	Online
Specified additional information	Require minor editing privileges
The requesting person/organization is:	Andrew Harrington/University Missouri Saint Louis
Order reference number	
Author/Editor	Andrew Harrington
The standard identifier of New Work	Thesis
Title of New Work	Endogenous Small Interfering RNA: Insights into esiRNA biogenesis and their precursors
Publisher of New Work	Proquest
Expected publication date	Dec 2016
Estimated size (pages)	170
Total (may include CCC user fee)	0.00 USD

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Nov 03, 2016

This Agreement between Andrew W Harrington ("You") and Nature Publishing Group ("Nature Publishing Group") consists of your license details and the terms and conditions provided by Nature Publishing Group and Copyright Clearance Center.

License Number	3980300868766
License date	Nov 01, 2016
Licensed Content Publisher	Nature Publishing Group
Licensed Content Publication	Nature Reviews Genetics
Licensed Content Title	Small silencing RNAs: an expanding universe
Licensed Content Author	Megha Ghildiyal and Phillip D. Zamore
Licensed Content Date	Feb 1, 2009
Licensed Content Volume Number	10
Licensed Content Issue Number	2
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Figures	FIGURE 1 | Small RNA silencing pathways in Drosophila.
Author of this NPG article	no
Your reference number	

Title of your thesis / dissertation	Endogenous Small Interfering RNA: Insights into esiRNA biogenesis and their precursors
Expected completion date	Dec 2016
Estimated size (number of pages)	170
Requestor Location	Andrew W Harrington 1 University Blvd SAINT LOUIS, MO 63121 United States Attn: Andrew W Harrington
Billing Type	Invoice
Billing Address	Andrew W Harrington 1 University Blvd SAINT LOUIS, MO 63121 United States Attn: Andrew W Harrington
Total	0.00 USD
Terms and Conditions	