2015-07-24

Using Experimental and Computational Strategies to Understand the Biogenesis of microRNAs and piRNAs: A Dissertation

Bo W. Han
University of Massachusetts Medical School

USING EXPERIMENTAL AND COMPUTATIONAL STRATEGIES TO UNDERSTAND THE BIOGENESIS OF MICRORNAS AND PIRNAS

A Dissertation Presented

By

Bo Han

Submitted to the Faculty of the University of Massachusetts Medical School

Graduate School of Biomedical Sciences

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

July 24, 2015

RNA Therapeutics Institute

USING EXPERIMENTAL AND COMPUTATIONAL STRATEGIES TO

UNDERSTAND THE BIOGENESIS OF MICRORNAS AND PIRNAS

Dissertation Presented

By

Bo Han

The signatures of the Dissertation Committee signify completion and approval as to style and content of the Dissertation

_____

Phillip Zamore, Ph.D., Thesis Advisor

_____

Nikolaus Rajewsky, Ph.D., Member of Committee

_____

Sean Ryder, Ph.D., Member of Committee

_____

Erik Sontheimer, Ph.D., Member of Committee

The signature of the Chair of the Committee signifies that the written dissertation meets the requirements of the Dissertation Committee

_____

Zhiping Weng, Ph.D., Chair of Committee

The signature of the Dean of the Graduate School of Biomedical Sciences signifies that the student has met all graduation requirements of the school.

_____

Anthony Carruthers, Ph.D.,
Dean of the Graduate School of Biomedical

Biochemistry and Molecular Pharmacology

July 24, 2015

DEDICATION

To my supportive family

Thank You!

ACKNOWLEDGEMENTS

This thesis would not be possible without the support of my colleagues, friends, and family. First and foremost, my gratitude goes to my advisor Phillip Zamore, who granted me the freedom to pursue anything that interested me scientifically. Beyond that, his unmatched scientific insights, unconditional supports, and meticulous scholarship greatly boosted the level of science I did. What Phil taught me is more than just experimental or writing skills, more importantly, he trained me as an independent scientist. Additionally, Phil has created an incredible working environment by recruiting a group of wonderful people with different strengths. To that end, I am also indebted to my colleagues and former members in the Zamore Lab, Alicia Boucher, Amena Arif, Cansu Colpan, Carlos Fabián Flores-Jasso, Cha San Koh, Chengjian Li, Chris Roy, Christian Matranga, Cindy Tipping, Desiree Brady, Elif Sarinay, Gwen Farley, Hervè Seitz, Irena Pekker, Jennifer Broderick, Katharine Cecchini, Kaycee Quarles, Keith Boundy, Liang Meng Wee, Megha Ghildiyal, Paul Albosta, Pei-Hsuan Wu, Ryuya Fukunaga, Samson Jolly, Shengmei Ma, Stefan Ameres, Tianfang Ge, Timothy Chang, Tracey Lincoln, Wei Wang, William Salomon, Xin Li, Yongjin Lee, Yujing Yang, Zhao Zhang. I was fortunate to work with them and there are a few that I would like to acknowledge specifically. Stefan Ameres is my experiment mentor and collaborator on my first paper. He is very smart, restrict, and generous. Wei Wang collaborated with me on three incredible manuscripts. Without her help, I could never achieve any of those. Jui-Hung Hung taught me a lot of

programming skills and I really enjoy the time writing Tailor with him. Chengjian Li is like a big brother and he is always very supportive both in the lab and in my personal life. Jennifer Broderick helps me proofreading all my manuscripts and my writing improves a lot because of her. Gwen Farley is our lab manager and she is incredibly helpful. Tiffanie Covello is the most efficient lab administrator. She has helped me in every way to set up dates for my qualifying exam, TRAC meetings, defense, and numerous practice talks. I also have to thank my committee members, Zhiping Weng, Sean Ryder, Melissa Moore, Erik Sontheimer, and Nikolaus Rajewsky. Zhiping encouraged me to start my journal of bioinformatics and I learned a lot from her and colleagues in her lab. Sean and Melissa are very supportive for both my scientific projects and graduation. Many great ideas came from them during my TRAC meetings. I would also like to thank Erik and Nikolaus for agreeing to sacrifice their valuable time and serve in my defense committee.

Bo Han

05/24/2015

ABSTRACT

Small RNAs are single-stranded, 18–36 nucleotide RNAs that can be categorized as miRNA, siRNA, and piRNA. miRNA are expressed ubiquitously in tissues and at particular developmental stages. They fine-tune gene expression by regulating the stability and translation of mRNAs. piRNAs are mainly expressed in the animal gonads and their major function is repressing transposable elements to ensure the faithful transfer of genetic information from generation to generation. My thesis research focused on the biogenesis of miRNAs and piRNAs using both experimental and computational strategies.

The biogenesis of miRNAs involves sequential processing of their precursors by the RNase III enzymes Drosha and Dicer to generate miRNA/miRNA* duplexes, which are subsequently loaded into Argonaute proteins to form the RNA-induced silencing complex (RISC). We discovered that, after assembled into Ago1, more than a quarter of *Drosophila* miRNAs undergo 3′ end trimming by the 3′-to-5′ exoribonuclease Nibbler. Such trimming occurs after removal of the miRNA* strand from pre-RISC and may be the final step in RISC assembly, ultimately enhancing target messenger RNA repression. Moreover, by developing a specialized Burrow-Wheeler Transform based short reads aligner, we discovered that in the absence of Nibbler a subgroup of miRNAs undergoes increased tailing—non-templated nucleotide addition to their 3′ ends, which are usually associated with miRNA degradation. Therefore, the 3′

trimming by Nibbler might increase miRNA stability by protecting them from degradation.

In *Drosophila* germ line, piRNAs associate with three PIWI-clade Argonaute proteins, Piwi, Aub, and Ago3. piRNAs bound by Aub and Ago3 are generated by reciprocal cleavages of sense and antisense transposon transcripts (a.k.a., the “Ping-Pong” cycle), which amplifies piRNA abundance and degrades transposon transcripts in the cytoplasm. On the other hand, Piwi and its associated piRNA repress the transcription of transposons in the nucleus. We discovered that Aub- and Ago3-mediated transposon RNA cleavage not only generates piRNAs bound to each other, but also produces substrates for the endonuclease Zucchini, which processively cleaves those substrates in a periodicity of ~26 nt and generates piRNAs that predominantly load into Piwi. Without Aub or Ago3, the abundance of Piwi-bound piRNAs drops and transcriptional silencing is compromised. Our discovery revises the current model of piRNA biogenesis.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ago: Argonaute

A. thaliana: *Arabidopsis thaliana*

Aub: Aubergine

BWT: Burrows-Wheeler transform

CAGE: 7-methyl guanosine-cap analysis of gene expression

C. elegans: *Caenorhabditis elegans*

CHS: chalcone synthase

D. melanogaster: *Drosophila melanogaster* (fruit fly)

DGCR8: DiGeorge syndrome critical region 8

dpc: days post coitum

dpp: post partum

dsRNA: double-stranded RNA

endo-siRNAs: endogenous small interfering RNAs

flam: *flamenco*

FM-index: Full-text index in Minute space

H3K4me2: di-methylated histone H3 lysine 4

H3K9me: Methylated histone H3 at lysine 9

Hen1: Hua Enhancer 1

HESO1: HEN1 SUPPRESSOR1

Loqs: Loquacious

mRNAs: messenger RNA

miRNA: microRNA

Nbr: Nibbler

NMD: nonsense-mediated decay

nt: nucleotide

PACT: Protein activator of the interferon-induced protein kinase, PKR

Pasha: Partner of Drosha

PAZ: PIWI/Argonaute/Zwille

PIWI: P-element induced wimpy testes

piRNA: PIWI-interacting RNA

Pri-miRNA: Primary miRNA

Pre-miRNA: Precursor miRNA

R2D2: Two dsRNA binding domain (R2) and Dicer-2 associated (D2)

RISC: RNA-induced silencing complex

RNAi: RNA interference

Rr luc: *Renilla reniformis* luciferase

rRNA: ribosomal RNA

S2 cell: Schneider 2 cell line

SA: suffix array

ssRNA: single-stranded RNA

siRNA: small interfering RNA

TRBP: HIV Trans-activating response (TAR) RNA-binding protein

COPYRIGHT INFORMATION

Chapter I Introduction

History of small RNA

Small RNA-mediated silencing, also known as RNA interference (RNAi), was first described as “co-suppression” in early 1990s. In an effort to alter the color of the flowers, plant scientists introduced into petunia an extra copy of chalcone synthase (*CHS*)—a critical enzyme in the pigmentation process. Unexpectedly, the exogenous *CHS* turned out to suppress the endogenous allele and produced totally or partially white flower (van der Krol et al., 1990; Napoli et al., 1990). Detailed analysis revealed that the messenger RNA (mRNA) of the endogenous *CHS* was reduced dramatically. A similar phenomenon was later observed in fungus (Romano and Macino, 1992) and fruit fly (Pal-Bhadra et al., 1997). Nonetheless, the molecular mechanism of co-suppression remained elusive. In 1998, Fire and colleagues reported that double-stranded RNA (dsRNA), when injected into *C. elegans*, mediates the silencing of endogenous genes with homolog sequences (Fire et al., 1998). This phenomenon is called RNA interference and later found to be operational in many organisms as well as in human cells. In the past decade, scientists have made enormous progress elucidating the mechanism of RNAi—those exogenous dsRNAs are converted into ~21 nucleotide (nt) short RNAs, which guide a group of proteins to form the RNA-induced silencing complex (RISC) and repress messenger RNAs (mRNAs) through sequence complementarity. Interestingly, the endogenous counterpart of RNAi exists in many species and is first described in *C.elegans* in 1993 (Lee et al., 1993; Wightman et al., 1993). To date, three categories of endogenous small

RNAs have been well characterized. They include microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs), and PIWI-interacting RNAs (piRNAs).

microRNA Biogenesis

miRNAs are mainly transcribed by RNA Polymerase II either from independent transcriptional unit or from the introns of protein coding genes (Figure 1.1; Lee et al., 2004; Cai et al., 2004; Borchert et al., 2006). Those primary miRNAs (pri-miRNA) contain one or multiple stem-loops that harbor the future miRNAs. In the nucleus, pri-miRNAs are recognized and processed into precursor miRNAs (pre-miRNA) by the microprocessor complex, which consists of an RNase III enzyme Drosha and its partner protein DiGeorge syndrome critical region gene 8 (DGCR8; named Pasha in fly; Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). Alternatively, a subset of miRNAs is produced from introns, which undergo lariat de-branching to produce pre-miRNAs (Ruby et al., 2007; Okamura et al., 2007).

Figure 1.1

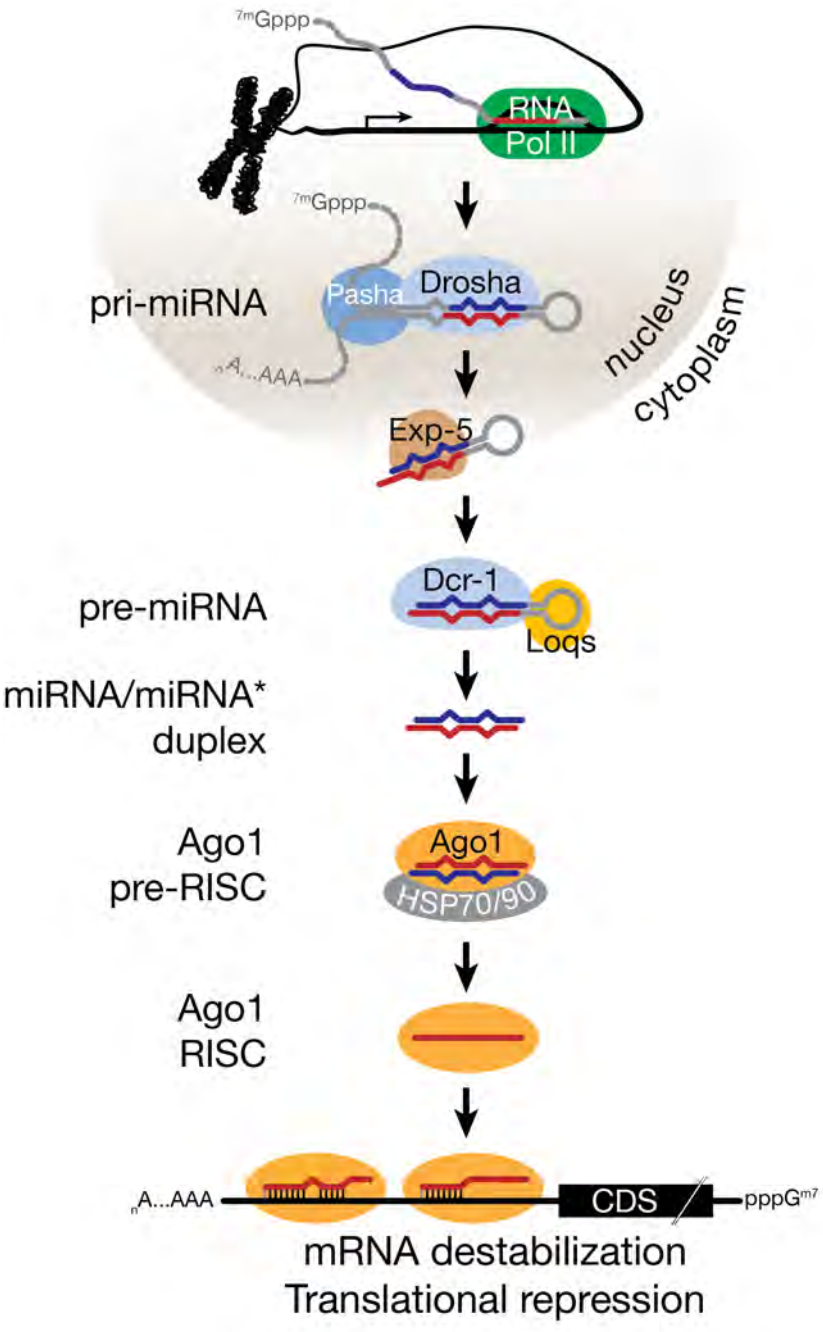

7mGppp
RNA
Pol II
7mGppp
pri-miRNA
Pasha
Drosha
nucleus
cytoplasm
nA...AAA
Exp-5
Dcr-1
pre-miRNA
Loqs
miRNA/miRNA*
duplex
Ago1
pre-RISC
Ago1
HSP70/90
Ago1
RISC
nA...AAA
CDS
pppGm7
mRNA destabilization
Translational repression

Figure Legend 1.1. MicroRNA Biogenesis in *Drosophila*

Pre-miRNAs, which are 60–70 nt in length and form a stem-loop structure with 2-nt overhang at its 3′ end—a signature of RNase III product—, are then exported into the cytoplasm by Exportin-5 (called Ranbp21 in flies; Yi et al., 2003; Bohnsack et al., 2004). In the cytoplasm, another RNase III enzyme Dicer, with the aid of its partner proteins—TAR RNA-binding protein 2 (TARBP2, also known as TRBP), protein kinase R-activating protein (PACT) in human, and Loquacious isoform PB (Loqs-PB) in flies— cleaves the pre-miRNA and liberates the miRNA/miRNA* duplex (Bernstein et al., 2001; Hutvagner, 2001; Grishok et al., 2001; Ketting et al., 2001; Jiang et al., 2005; Forstemann et al., 2005; Haase et al., 2005; Lee et al., 2006). With the aid of HSC70–HSP90 chaperone machinery, this duplex is loaded into an AGO protein (Iwasaki et al., 2010). Subsequent maturation steps expel the miRNA*, producing a mature RISC, which finds its target through sequence complementarity.

microRNA-mediated Gene Regulation

AGO divides the small RNA guide into functional domains: anchor, seed, central, 3′ supplementary, and tail (Wee et al., 2012). The seed region—2^{nd}–7^{th} nt of the small RNA—is pre-organized by AGO in a quasi-helical structure that pre-pays the entropic penalty in nucleic acid pairing (Parker et al., 2009). Such pre-organized structure initiates the binding of RISC to the mRNA target and determines the specificity (Bartel, 2004; Bartel, 2009). A single mutation in the seed region can abolish target regulation (Haley and Zamore, 2004). The seed pairing between miRNA and target induces a structural rearrangement of AGO

proteins, allowing subsequent pairing between the 3′ complementary region—13^{th}–16^{th} nt of the small RNA—and the target (Wang et al., 2008; Wang et al., 2008; Wang et al., 2009b; Schirle and MacRae, 2012). Such pairing reduces the off-rate of RISC from the target and enhances target repression (Grimson et al., 2007; Wee et al., 2012).

miRNA target sequences tend to occur in the 3′ untranslated regions (UTRs) of mRNAs (Wightman et al., 1993; Lee et al., 1993; Pasquinelli et al., 2000; Lai, 2002). Due to the short length of the seed region, more than half of genes in mammals are regulated by miRNAs (Farh et al., 2005; Stark et al., 2005; Friedman et al., 2009). Plant miRNAs often trigger the cleavage of their targets due to their extensive complementarity (Llave et al., 2002; Tang et al., 2003; German et al., 2008; Addo-Quaye et al., 2008). However, only a few animal miRNAs have sufficient complementarity to induce the cleavage by AGO proteins (Yekta, 2004). Therefore animal miRNAs deploy different strategies in target regulation. Mass spectrometry and ribosome profiling experiments suggest that the majority of miRNA-mediated protein decrease is caused by mRNA destabilization, instead of translational repression (Baek et al., 2008; Guo et al., 2010), although translational repression might also exist and precede mRNA degradation for a subset of targets (Bazzini et al., 2012).

microRNA Turnover

Binding to an mRNA target not only leads to the degradation of that mRNA, but also reduces the stability of the miRNA itself, especially when the miRNA pairs

with a target with a high complementarity (Krutzfeldt et al., 2005; Chatterjee and Grosshans, 2009). Two distinct classes of modification associate with miRNA degradation—“tailing” and “trimming” (Ameres et al., 2010). Tailing describes non-templated nucleotide addition to the 3′ ends of miRNAs and trimming refers to the 3′-to-5′ exonucleolytic resection of miRNA from its 3′ end. The detailed mechanism and function of tailing and trimming remain elusive, but it is hypothesized that target binding induces a structural rearrangement of Ago-miRNA complex that releases the 3′ end of miRNA from the PAZ domain and makes it more accessible to the tailing enzyme (nucleotidyl transferase).

Target-directed small RNA tailing is restricted to miRNAs in *Drosophila*. piRNAs in mice, siRNAs and piRNAs in flies, and miRNAs in plants are 2′-*O*-methylated at their 3′ end by methyltransferase Hen1. Such modification prevents tailing and protects miRNAs from target-induced degradation, because they often have extensive complementarity with their targets (Saito et al., 2007; Horwich et al., 2007; Ameres et al., 2011).

piRNA

piRNAs are single-stranded, 23–36 nucleotide RNAs that act as guides for an animal-specific class of Argonaute proteins, the PIWI proteins. The first piRNAs—derived from the *Suppressor of Stellate* locus in *Drosophila melanogaster* testes—were discovered in 2001 (Aravin et al., 2001). Although the authors noted the larger size of those “rasiRNAs” (repeat-associated siRNAs), piRNAs were not recognized until 2006 as a distinct class of small

interfering RNAs (siRNAs) that derive from single-stranded, rather than double-stranded RNA (dsRNA), precursors (Vagin et al., 2006; Saito et al., 2006; Girard et al., 2006; Aravin et al., 2006). It was their association with PIWI, but not AGO, proteins and their independence from Dicer that finally distinguished piRNAs from siRNAs. Moreover, while siRNAs are expressed ubiquitously, piRNAs are predominantly found in animal gonads and are thought to be indispensable for fertility.

piRNA Classification and Evolution

piRNAs can be divided into four classes according to their origins and functions. The best-studied class corresponds to repeat-associated piRNAs, which silence transposons in insects and mice (Figure 1.2). The second class of piRNAs originates from the 3′ UTRs of mRNAs. Examples of such mRNA-derived piRNAs include piRNAs in the ovarian, somatic, follicle cells of flies and in the pre-pachytene, meiotic spermatocytes of mice. How mRNA-derived piRNAs are made or what they do is unknown. The third class, exemplified by pachytene piRNAs in the mouse testis, derives from intergenic, long non-coding RNAs (lncRNAs). Their targets and function are also elusive. Nonetheless, these three classes of piRNAs are highly conserved and have been found in Cnidaria (*Nematostella vectensis*) and Porifera (*Amphimedon queenslandica*; Grimson et al., 2008). The final piRNA class is *Caenorhabditis*-specific. Worm piRNAs—called 21U RNAs—cooperate with 22G siRNAs to repress “non-self” RNA transcription via histone methylation (Ruby et al., 2006; Lee et al., 2012).

Figure 1.2

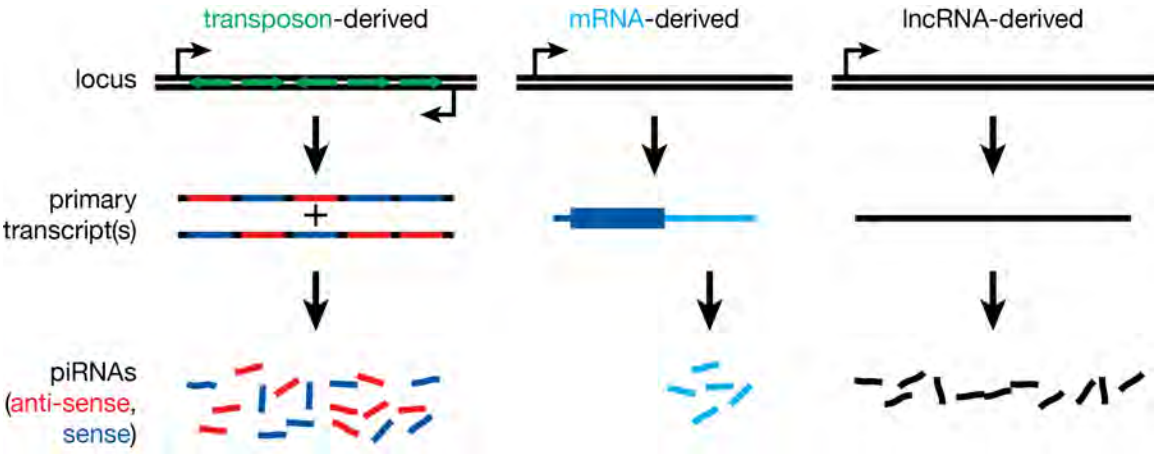

transposon-derived
mRNA-derived
lncRNA-derived
locus
primary
transcript(s)
piRNAs
(anti-sense,
sense)

Figure Legend 1.2. Classification of piRNAs

piRNAs can be classified into three groups based on their origins. The precursors of transposon-derived piRNAs are typically transcribed from both genomic strands and produce both sense and antisense piRNAs. In contrast, mRNA-derived piRNAs are always sense to the mRNA from which they are processed. Such piRNAs often come from 3′ UTRs. Long non-coding RNAs (lncRNAs) produce piRNAs from the entire transcript. piRNA function is only well understood for transposon-derived piRNAs.

PIWI proteins

Argonaute proteins lie at the center of all small RNA pathways. During animal evolution, Argonaute proteins have diverged into two clades: AGO proteins, which associate with miRNAs and siRNAs, and PIWI proteins, which bind piRNAs. Animals typically encode multiple PIWI-family members: most primates have four PIWI genes (*PIWIL1–4*), mice (*Piwil1*, *Piwil2*, *Piwil4*) and flies (*piwi, aub*, *ago3*) have three, and worms have two (*prg-1*, *prg-2*). Like most Argonaute proteins, even those in bacteria and archaea, both AGO and PIWI proteins possess highly conserved MID, PAZ and PIWI domains. The MID domain anchors the 5′ end and the PAZ anchors the 3′ end of the guide small RNA (Carmell et al., 2002; Hutvagner and Simard, 2008; Cenik and Zamore, 2011). Small RNAs guide Argonautes to their targets through Watson-Crick base pairing (Wee et al., 2012). With enough complementarity, the RNaseH-like structure of the PIWI domain cleaves the phosphodiester bond between the two nucleotides in the target RNA that base-pair with the 10^{th} and 11^{th} nucleotides of the small RNA guide (Liu et al., 2004; Song et al., 2004). The side-chains of conserved aspartic acid and histidine residues in the PIWI domain form a DDH catalytic triad that coordinates a magnesium ion proposed to activate a nucleophilic water molecule that breaks the phosphodiester bond. A subset of PIWI proteins, including *Drosophila* Piwi and mouse MIWI2 (officially PIWIL4) can repress transposon transcription by promoting histone or DNA methylation (Aravin et al., 2007; Carmell et al., 2007; Kuramochi-Miyagawa et al., 2008; Sienski et al.,

2012; Huang et al., 2013; Rozhkov et al., 2013; Le Thomas et al., 2013). Mutation of the catalytic triad of these proteins does not impair transposon silencing, yet their catalytic triad has been conserved in evolution, implying an undiscovered role for slicing by these PIWI proteins (Reuter et al., 2011; Saito et al., 2009; Sienski et al., 2012; Darricarrère et al., 2013). In worms, 21U piRNAs guide the PIWI protein PRG-1 to initiate silencing of “non-self” RNA transcription via 22G siRNAs, which enter the nucleus and direct methylation of histone H3 lysine 9 (H3K9) (Bagijn et al., 2012; Lee et al., 2012; Shirayama et al., 2012).

Tudor Proteins

The gene *tudor* emerged in a genetic screen for flies that were grandchildless, like its namesake, the English Tudor royal family (Boswell and Mahowald, 1985). Tudor contains 11 Tudor (Tud) domains, which promote protein-protein interactions by folding into a β-barrel structure that binds (6-*N*,6-*N*) methyl-lysine and (6-*N*,6-*N*) methyl-arginine. *Drosophila* and mouse PIWI proteins have di-methylated arginines on their N-termini, and Tudor proteins are believed to serve as scaffolds that organize the components of the piRNA pathway into functional units (Nishida et al., 2009; Kirino et al., 2009). More than ten fly and eight mouse Tudor proteins function in the piRNA pathway: loss of any Tudor protein leads to sterility as well as a distorted piRNA pool. For example, loss of *tejas* and *spindle-E* eliminates most piRNAs in the *Drosophila* germline nurse cells but leave the somatic follicle cell piRNA pathway unaltered (Patil and Kai, 2010; Gonzalez-Reyes et al., 1997). In contrast, *female sterile (1) Yb* (*fs(1)Yb*) acts only in the

somatic follicle cells (Szakmary et al., 2009). Other Tudor proteins are required in both somatic and germline cells: loss of *vreteno* causes a large decline in piRNAs in both tissues (Handler et al., 2011; Zamparini et al., 2011). While Tudor proteins are essential to the piRNA pathway, the detailed mechanisms by which these proteins function remain unknown.

Nuage

Animal germline cells feature a perinuclear structure termed nuage, French for "cloud," reflecting its amorphous, electron-dense appearance under the electron microscope. In the fruit fly ovary, most germline piRNA pathway components localize to nuage, including the RNA helicases Vasa and Armitage, the PIWI proteins Aub and Ago3, as well as the Tudor proteins (Brennecke et al., 2007; Lim and Kai, 2007). The localization of piRNA pathway proteins within nuage builds on a foundation of Vasa: in *vasa* mutants, all known piRNA pathway proteins fail to localize to nuage. In *tejas* and *spindle-E* mutants, only Vasa remains in nuage, suggesting that the binding of Tejas and Spindle-E immediately follows the deposition of Vasa (Patil and Kai, 2010; Lim and Kai, 2007). Topping off the structure are the PIWI proteins Aub and Ago3: no known piRNA pathway proteins depend on *aub* and *ago3* to localize to nuage.

In the mammalian testis—where piRNAs play an essential role in spermatogenesis—two different granular structures, "intermitochondrial cement" (also called the "pi-body") and "chromatoid body" (also called "piP-body") take the place of nuage (Aravin et al., 2009). In the embryonic testis, DDX4 (mammalian

Vasa), MILI (officially PIWIL2), TDRD1, and GASZ localize to the intermitochondrial cement. MIWI2, a PIWI protein found only in the embryonic and newborn testis, is found in a distinct granule containing HMG box protein MAELSTROM, TDRD9 (the homolog of Spindle-E), and many proteins that are components of P-bodies, somatic cytoplasmic granules involved in RNA-degrading pathways such as nonsense-mediated decay (NMD) and miRNA-guided mRNA repression. The chromatoid body first appears in spermatocytes during Meiosis I and contains MIWI (officially PIWIL1), MILI, DDX4, MAELSTROM, TDRD1, TDRD6, and some P-body components. The co-localization of P-body components and PIWI proteins suggests that piRNAs are involved in mRNA turnover, but this idea remains to be tested.

piRNA clusters

The loci that give rise to piRNAs are not dispersed throughout the genome but rather are concentrated in specific piRNA “clusters” (Brennecke et al., 2007). The best-studied piRNA cluster is *flamenco* (*flam*), which was genetically defined as a locus required for transposon silencing long before the fly genome was sequenced or piRNAs were discovered (Pelisson et al., 1994; Prud'homme et al., 1995; Robert et al., 2001). *flamenco* is predominantly, if not exclusively, active in the fly somatic follicle cells (Mohn et al., 2014). Disruption of *flamenco* leads to derepression of somatic transposons, including *gypsy*, which forms viral particles that invade the adjacent germline and, ultimately, cause sterility (Malone et al., 2009). Early hypotheses assumed *flamenco* encoded a protein-coding gene, but

efforts to find protein-coding genes at the locus were fruitless. High-throughput sequencing of 18–30 nt RNAs demystified the identity of *flamenco* as a piRNA cluster that generates piRNAs from embedded transposon fragments, most of which are inserted in the antisense orientation. Therefore *flamenco* piRNAs are predominately antisense to transposon RNA. A P-element insertion in the *flamenco* promoter abolishes piRNA production from this ~180 kb locus, suggesting that a single transcript spans this region (Malone et al., 2009; Goriaux et al., 2014). This observation, together with the finding that piRNA production in flies needs neither Dicer-1 nor Dicer-2, led to the belief that piRNAs originate from single-stranded RNA precursors (Vagin et al., 2006). While *flamenco* produces piRNAs only from one strand (a “uni-strand” cluster), most germline piRNA clusters produce piRNAs from both strands (“dual-strand” clusters). Unlike uni-strand clusters, which exhibit canonical polymerase II transcriptional signatures including enrichment of di-methylated histone H3 lysine 4 (H3K4me2) at their promoters and 7-methylguanylate caps on the 5′ ends of their transcripts, dual-strand clusters lack those features and are enriched in the repressive H3K9me3 mark (Zhang et al., 2014b; Mohn et al., 2014). Interestingly, transcription of dual-strand clusters, but not uni-strand clusters, requires Rhino (a paralog of Heterochromatin Protein 1, HP1), Cutoff (a yeast Rai1-like protein), and Deadlock (Zhang et al., 2012; Zhang et al., 2014b; Mohn et al., 2014). Current evidence suggests that Rhino, Deadlock, and Cutoff form a complex that licenses dual-strand clusters to produce piRNAs by suppressing splicing or

polyadenylation and cleavage of their transcripts. The DEAD box protein UAP56 then binds dual-strand-cluster transcripts and escorts them to nuclear periphery opposite nuage, where, it is proposed, they are transferred to the piRNA biogenesis machinery to be processed into mature piRNAs (Zhang et al., 2012).

The biogenesis of 3′ UTR-derived piRNAs is also poorly understood. How the piRNA biogenesis machinery identifies certain mRNAs as its substrates remains one of the most mysterious questions in the field. Similarly, little is known about the biogenesis of intergenic piRNAs, which exist in mice but not in flies. The transcription factor A-MYB (MYBL1) regulates the transcription of many intergenic piRNA precursors in the mouse testis (Bolcun-Filas et al., 2011; Li et al., 2013). Of note, A-MYB also drives the transcription of many genes encoding piRNA pathway components, such as MIWI and TDRD1. But how these long non-coding transcripts, but not protein-coding transcripts, are processed to become mature piRNAs is largely unknown.

The biogenesis of *C. elegans* piRNAs differs from that in flies and mice. The 21U piRNAs are each independently transcribed as capped, 26–29 nt small RNAs (Ruby et al., 2006). An 8-nt motif located ~40 nt upstream promotes their coordinated transcription. The maturation of 21U piRNAs begins with the removal of the cap and two nucleotides from the 5′ end of the pre-21U piRNA transcript and ends with the trimming and 2′-*O*-methylation of their 3′ ends.

piRNA Maturation and Function

After transcription, piRNA primary transcripts (pri-piRNAs) are processed to mature piRNAs. One hypothesis suggests that a single-stranded endoribonuclease, Zucchini, cleaves pri-piRNAs to create piRNA intermediates bearing a 5′-monophosphate and 3′-hydroxyl ends. Without Zucchini, pri-piRNA transcripts accumulate and mature piRNAs decline. A UPF1-like helicase, Armitage, participates in piRNA biogenesis at this step but its function still remains mysterious (Olivieri et al., 2010; Haase et al., 2010). It is believed, with some experimental support, that piRNA intermediates are loaded into PIWI proteins, then trimmed by a 3′-to-5′ exonuclease to a length determined by the footprint of the PIWI protein (Kawaoka et al., 2011).

Once bound to a piRNA, Piwi can enter the nucleus, where it is presumed to bind complementary nascent transposon transcripts. Piwi can then recruit histone methyltransferases that deposit the repressive H3K9me3 mark to establish heterochromatin and restrict transposon transcription (Klenov et al., 2011; Sienski et al., 2012; Rozhkov et al., 2013; Le Thomas et al., 2013; Klenov et al., 2014). The mechanism employed by the germline piRNA pathway is more complex: the germline contains Piwi, Aub, and Ago3 and germline piRNAs mainly derive from dual-strand piRNA clusters and correspond to both sense and antisense transposon sequences. Curiously, Aub and Piwi seem to preferentially bind piRNAs in the antisense orientation while Ago3 tends to associate with sense piRNAs (Brennecke et al., 2007). Aub-bound antisense piRNAs typically

start with a 5′ uridine while Ago3-bound sense piRNAs often have adenosine as their tenth nucleotide. Complementarity between the first ten nucleotides of Aub- and Ago3-bound piRNAs led to the hypothesis that sense piRNAs are generated by target cleavage directed by antisense piRNAs, and vice versa (Figure 1.3A). Based on this, the Hannon and Siomi laboratories proposed the “Ping-Pong” model to explain the biogenesis of germline piRNAs in insects (Figure 1.3B; Brennecke et al., 2007; Gunawardane et al., 2007). In this model, maternally deposited or newly synthesized Aub-bound, antisense piRNAs initiate the Ping-Pong cycle by cleaving transposon mRNA transcripts and generating sense-oriented piRNA intermediates. The sense intermediates are loaded into Ago3 and trimmed to mature sense piRNAs. These Ago3-bound, sense piRNAs can then bind and cleave the antisense transposon sequences present in the transcripts of the original piRNA cluster, producing piRNA intermediates that begin with uridines—a substrate that reinitiates the cycle. The model suggests that the piRNA biogenesis pathway is an adaptive system that silences active transposons with sequences complementary to piRNA cluster transcripts. Experiments in cultured, immortalized silkworm germline cells suggest that the DEAD-box RNA helicase Vasa assembles a complex with transposon transcripts, Aub, Ago3, and the Tudor protein Qin (Xiol et al., 2014). After Aub cleaves the transposon, Vasa facilitates the transfer of the 3′ cleavage product to Ago3. Without Qin, the interaction between Aub and Ago3 is weakened, and piRNA amplification occurs mainly between Aub proteins (Zhang et al., 2011; Zhang et

al., 2014a). For unknown reasons, piRNAs generated from such homotypic Ping-Pong cannot silence transposons.

Figure 1.3

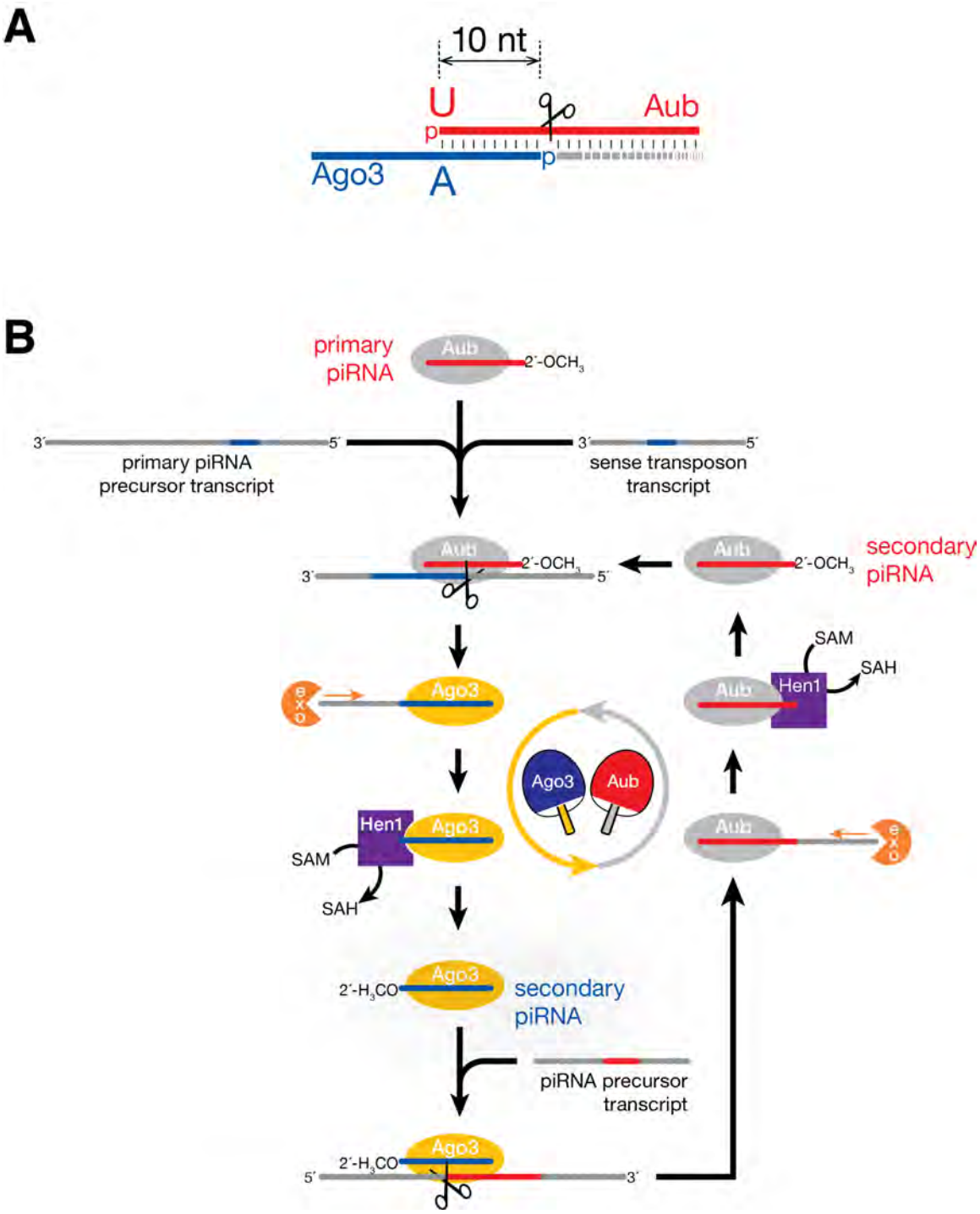

A
10 nt
U
Aub
p
p
Ago3
A
B
primary piRNA
Aub
2´-OCH3
3´
5´
primary piRNA precursor transcript
sense transposon transcript
secondary piRNA
SAM
SAH
Hen1
Ago3
Aub
exo
2´-H3CO
piRNA precursor transcript

Figure Legend 1.3. Ping-Pong Cycle Amplifies piRNAs in *Drosophila*

(A) In flies, piRNAs bound by Aubergine (Aub) and Ago3 typically overlap by ten nucleotides, suggesting that one piRNA was made by cleavage of its precursor by a PIWI protein guided by the other piRNA.

(B) The Ping-Pong model for secondary piRNA biogenesis seeks to explain the unique relationship of Aub- and Ago3-bound piRNAs. Antisense piRNAs guide Aub to cleave transposon mRNA and cluster transcripts that contain transposon fragments. The resulting 3′ cleavage products bear a 5′ monophosphate, allowing them to load into Ago3. An unidentified, 3′ to 5′ exonuclease trims the Ago3-bound RNA to the proper length before Hen1 methylates its 3′ end. This process produces new piRNAs that are in the same orientation as the transposon. The piRNAs can now guide Ago3 to cleave piRNA precursor transcripts harboring complementary transposon fragments. These new 3′ cleavage products can then be loaded into Aub, trimmed, and methylated, generating mature piRNAs that can guide Aub to initiate the next cycle.

In the mouse testis, piRNAs are even more complex, with three types of piRNAs appearing during spermatogenesis: prenatal piRNAs from repetitive sequences, mRNA-derived piRNAs from 3′ UTRs, and pachytene piRNAs from long non-coding transcripts (Kuramochi-Miyagawa et al., 2001; Deng and Lin, 2002; Girard et al., 2006; Aravin et al., 2006; Grivna et al., 2006a; Grivna et al., 2006b; Aravin et al., 2007; Carmell et al., 2007; Li et al., 2013). Prenatal piRNAs bind both MIWI2 and MILI. MIWI2 begins to accumulate around 14.5–15.5 days post coitum (dpc), declines starting at birth, and becomes undetectable ~4 days post partum (dpp), when prospermatogonial cells re-enter the mitotic cell cycle and initiate the first wave of spermatogenesis. Like fly Piwi, MIWI2 silences transposons transcriptionally; its endonuclease activity is dispensable for transposon repression (De Fazio et al., 2011). MILI expression begins in the embryonic testis (12.5 dpc) and lasts until the round spermatid stage, near the end of spermatogenesis. It binds all three groups of piRNAs. In the embryonic testis, MILI performs “Ping-Pong” with itself, amplifying the piRNA pool before presenting mature piRNAs to MIWI2, which translocates into the nucleus and recruits the DNA methyltransferase DNMT3L to methylate and repress transposon loci (Aravin et al., 2008). Loss of MIWI2 or MILI leads to loss of DNA methylation, derepression of transposons, defects in spermatogenesis, and sterility. Most of the prenatal piRNA clusters are insertions of individual transposons or transposon fragments. Consequently, ~40 times more clusters

are required in mice than in flies to account for the sources of the same percentage of piRNAs.

Sequencing of small RNAs after birth and before the pachytene stage (i.e., “pre-pachytene”) identified the second group of mouse piRNAs: mRNA-derived piRNAs. While some transposon-mapping piRNAs remain in the pre-pachytene spermatocytes, these could have been made during the earlier, mitotic stages of spermatogenesis. Most pre-pachytene piRNAs are mRNA-derived; they typically map to the 3′ UTRs of protein-coding genes. Because small RNAs are used as guides to identify targets by nucleotide complementarity, it is a great mystery what function such mRNA-derived—i.e., sense—piRNAs could serve.

In response to a transcriptional program orchestrated by the transcription factor A-MYB, MIWI, additional piRNA pathway proteins, and the pachytene piRNAs emerge beginning at 12.5 dpp. The pachytene piRNAs rapidly grow to numbers that dwarf the abundance of the other two piRNA types, and they remain the most abundant piRNA population in the adult mouse testis. Mouse piRNA production requires most of the genes encoding homologs that function in the fly piRNA pathway, including the Zucchini homolog MitoPLD, the Armitage homolog Mov10l1, and several Tudor proteins (Watanabe et al., 2011; Zheng et al., 2010; Frost et al., 2010; Vourekas et al., 2015; Pan et al., 2005; Chuma et al., 2006; Arkov et al., 2006; Hosokawa et al., 2007; van der Heijden and Bortvin, 2009; Wang et al., 2009a; Chen et al., 2009a; Vasileva et al., 2009; Shoji et al., 2009; Huang et al., 2011; Yabuta et al., 2011; Tanaka et al., 2011; Mathioudakis

et al., 2012; Saxe et al., 2013; Pandey et al., 2013; Patil et al., 2014). Loss of A-MYB, MIWI, or other members of the pachytene piRNA pathway blocks spermatogenesis and causes male sterility. Nonetheless, how pachytene piRNAs are made or what they do remains unknown.

Unlike the miRNA and siRNA pathways, the piRNA pathway lacks good biochemical and cell culture tools to analyze its molecular details, especially for mRNA-derived and intergenic piRNAs. Moreover, both the sources and targets of transposon-silencing piRNAs often lie in parts of the genome that remain incompletely sequenced and incompletely assembled. Therefore, a new methodology is in demand for making a breakthrough in the piRNA research.

Chapter II The 3′-to-5′ Exonuclease Nibbler Shapes the 3′ Ends of MicroRNA

Disclaimer

This chapter was a product of a collaborative effort among the authors: Bo W Han (BWH), Stefan L Ameres (SLA), Jui-Hung Hung (JHH), Zhiping Weng (ZW), and Phillip D. Zamore (PDZ). SLA and BWH performed the biochemical experiments. BWH and JHH performed the computational analyses. ZW and PDZ supervised the project.

Summary

We show that after loading into Ago1, more than a quarter of all *Drosophila* miRNAs undergo 3′ end trimming by the 3′-to-5′ exonuclease Nibbler (CG9247). Depletion of Nibbler by RNAi reveal that miRNAs are frequently produced by Dicer-1 as intermediate that are longer than ~22 nt. Trimming of miRNA 3′ ends occurs after removal of the miRNA* strand from pre-RISC and may be the final step of RISC assembly, ultimately enhancing target mRNA repression. Additionally, we discovered that, in the absence of Nibbler, longer isoforms of miRNA were subjected to increased un-templated nucleotide addition to their 3′ ends—a molecular phenomenon known as “tailing”. Since tailing is associated with miRNA degradation, we conclude that 3′ end trimming by Nibbler improves miRNA stability. Our data provide a molecular explanation for the previously reported heterogeneity of miRNA 3′ ends and propose a model in which Nibbler converts miRNAs into isoforms that are compatible with the preferred length of Ago1-bound small RNAs.

Introduction

MicroRNAs (miRNAs) are ~22 nucleotide (nt) small RNAs that control development, physiology, and pathology in animals and plants by regulating messenger RNA (mRNA) stability and translation in plants, green algae, and animals. Loss of proteins required for the production or function of miRNAs typically result in severe developmental defects or lethality (Bartel, 2004; Bartel, 2009).

miRNA genes are generally transcribed by RNA polymerase II to generate 5′ capped and 3′ polyadenylated primary miRNAs (pri-miRNAs) that are then sequentially processed into mature miRNA duplexes (Lee et al., 2004; Cai et al., 2004; Borchert et al., 2006). Pri-miRNAs contain one or more characteristic stem-loops that are recognized and cleaved by the nuclear RNase III enzyme Drosha to generate ~70 nt long precursor miRNAs (pre-miRNAs; Lee et al., 2003; Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004). Pre-miRNAs comprised a single-stranded loop and a partially base-paired stem whose termini bear the hallmarks of RNase III processing: a two-nucleotide 3′ overhang, a 5′ phosphate, and a 3′ hydroxyl group. Nuclear pre-miRNAs are exported by Exportin-5 to the cytoplasm, where the RNase III enzyme Dicer liberates ~22 nt mature miRNA/miRNA* duplexes from the pre-miRNA stem (Yi et al., 2003; Bohnsack et al., 2004; Bernstein et al., 2001; Hutvagner, 2001; Grishok et al., 2001; Ketting et al., 2001; Jiang et al., 2005; Forstemann et al., 2005; Haase et al., 2005; Lee et al., 2006). Like all Dicer products, miRNA

duplexes contain two-nucleotide 3′ overhangs, 5′ phosphate, and 3′ hydroxyl groups. In flies, Dicer-1 cleaves pre-miRNAs to miRNAs, whereas Dicer-2 converts long double-stranded RNA into small interfering RNAs (siRNAs), which direct RNAi for host defense against viral infection and somatic transposon mobilization (Ghildiyal et al., 2008; Czech et al., 2008).

miRNA duplexes assemble into Argonaute proteins to form the precursor RNA-induced silencing complex (pre-RISC), a process uncoupled from small RNA production. In flies, miRNAs typically bind to Argonaute1 (Ago1) and siRNAs to Argonaute2 (Ago2) (Tomari et al., 2007; Czech and Hannon, 2011). During RISC assembly, one of the two strands of a miRNA duplex is selectively retained to form an active silencing complex. Strand selection is determined by the relative thermodynamic stability of the duplex ends, the identity of the 5′ nucleotides, as well as the structure and length of the miRNA duplex (Czech and Hannon, 2011). In mature RISC, a single-stranded miRNA directs Ago1 to bind partially complementary sequences, typically within the 3′ untranslated region (3′ UTR) of miRNAs. RISC binding represses mRNA expression by accelerating its decay or inhibiting its translation.

Results

The 3′ end of miRNA-34 is trimmed after its production by Dicer-1

miRBase annotates miR-34 (miR-34-5p) as 24 nt long, pairing to a 23 nt miR-34* strand (miR-34-3p, Figure 2.1A), but high resolution northern hybridization revealed additional, abundant 23, 22, and 21 nt miR-34 isoforms (Figure 2.1B). We analyzed small RNA sequencing data from fly heads for reads mapping to the miR-34 genomic locus. Of those reads, 98.5% began at the annotated 5′ end of miR-34; for miR-34-mapping reads bound to Ago1, 99.0% shared this same, unique 5′ end (data not shown). Similarly, 98.8% of all miR-34 reads in total RNA data sets from S2 cells shared this 5′ end (Figure 2.1C). On the contrary, the 3′ ends displayed various degree of heterogeneity. Since the 3′ ends of miR-34 are generated by Dicer-1, thus the shorter isoforms of miR-34 could reflect the inaccurate processing of Dicer-1.

Figure 2.1

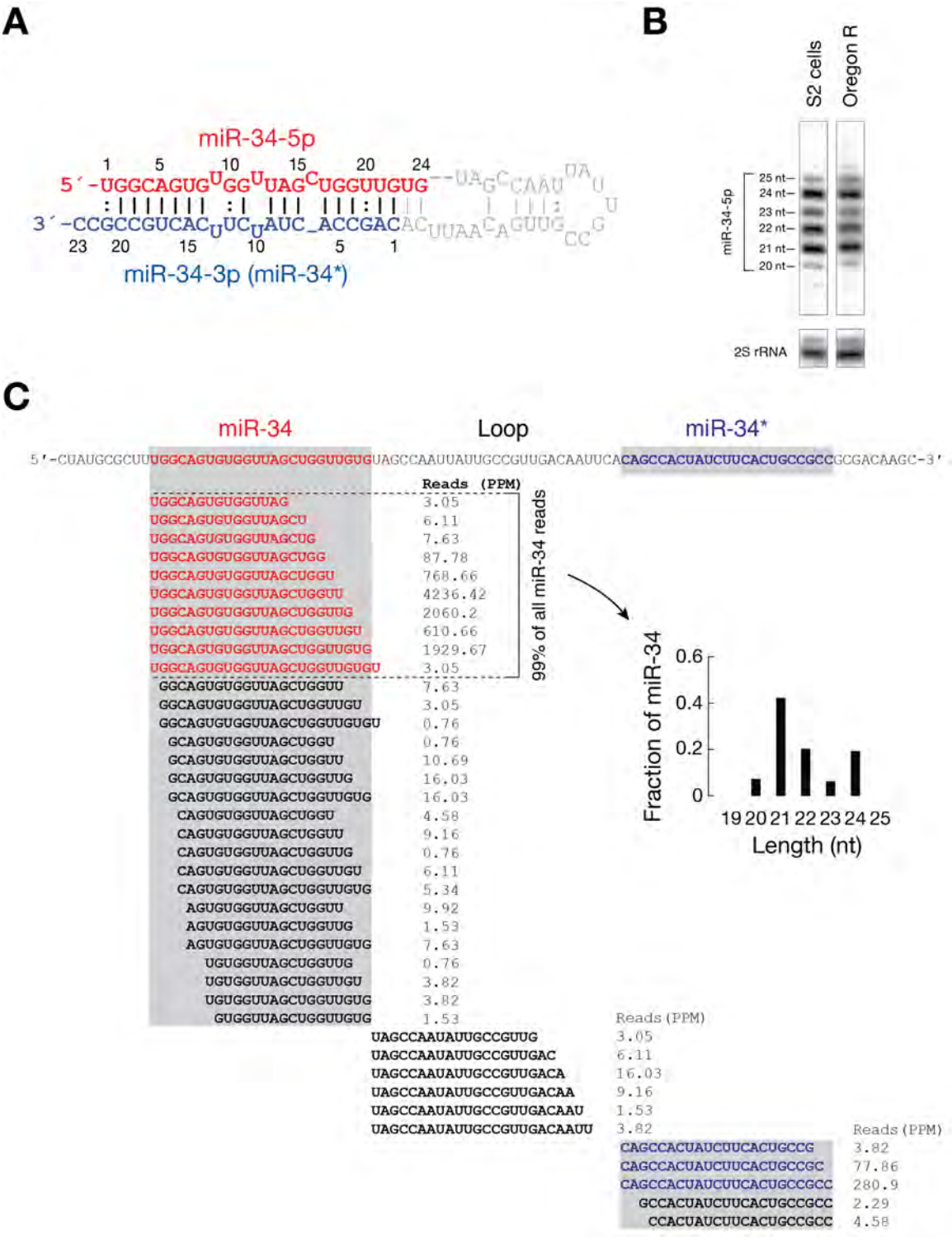

A
miR-34-5p
1 5 10 15 20 24
5′-UGGCAGUGUGGUUAGCUGGUUGUG
3′-CCGCCGUCACUUCUAUC_ACCGAC
23 20 15 10 5 1
miR-34-3p (miR-34*)
B
S2 cells
Oregon R
miR-34-5p
25 nt
24 nt
23 nt
22 nt
21 nt
20 nt
2S rRNA
C
miR-34
Loop
miR-34*
5′-CUAUGCGCUUUGGCAGUGUGGUUAGCUGGUUGUGUAGCCAAUUAUUGCCGUUGACAAUUCACAGCCACUAUCUUCACUGCCGCCGCGACAAGC-3′
Reads (PPM)
UGGCAGUGUGGUUAG 3.05
UGGCAGUGUGGUUAGCU 6.11
UGGCAGUGUGGUUAGCUG 7.63
UGGCAGUGUGGUUAGCUGG 87.78
UGGCAGUGUGGUUAGCUGGU 768.66
UGGCAGUGUGGUUAGCUGGUU 4236.42
UGGCAGUGUGGUUAGCUGGUUG 2060.2
UGGCAGUGUGGUUAGCUGGUUGU 610.66
UGGCAGUGUGGUUAGCUGGUUGUG 1929.67
UGGCAGUGUGGUUAGCUGGUUGUGU 3.05
99% of all miR-34 reads
GGCAGUGUGGUUAGCUGGUU 7.63
GGCAGUGUGGUUAGCUGGUUGU 3.05
GGCAGUGUGGUUAGCUGGUUGUGU 0.76
GCAGUGUGGUUAGCUGGU 0.76
GCAGUGUGGUUAGCUGGUU 10.69
GCAGUGUGGUUAGCUGGUUG 16.03
GCAGUGUGGUUAGCUGGUUGUG 16.03
CAGUGUGGUUAGCUGGU 4.58
CAGUGUGGUUAGCUGGUU 9.16
CAGUGUGGUUAGCUGGUUG 0.76
CAGUGUGGUUAGCUGGUUGU 6.11
CAGUGUGGUUAGCUGGUUGUG 5.34
AGUGUGGUUAGCUGGUU 9.92
AGUGUGGUUAGCUGGUUG 1.53
AGUGUGGUUAGCUGGUUGUG 7.63
UGUGGUUAGCUGGUUG 0.76
UGUGGUUAGCUGGUUGU 3.82
UGUGGUUAGCUGGUUGUG 3.82
GUGGUUAGCUGGUUGUG 1.53
Fraction of miR-34
0.6
0.4
0.2
0
19 20 21 22 23 24 25
Length (nt)
Reads (PPM)
UAGCCAAUAUUGCCGUUG 3.05
UAGCCAAUAUUGCCGUUGAC 6.11
UAGCCAAUAUUGCCGUUGACA 16.03
UAGCCAAUAUUGCCGUUGACAA 9.16
UAGCCAAUAUUGCCGUUGACAAU 1.53
UAGCCAAUAUUGCCGUUGACAAUU 3.82
Reads (PPM)
CAGCCACUAUCUUCACUGCCG 3.82
CAGCCACUAUCUUCACUGCCGC 77.86
CAGCCACUAUCUUCACUGCCGCC 280.9
GCCACUAUCUUCACUGCCGCC 2.29
CCACUAUCUUCACUGCCGCC 4.58

Figure Legend 2.1. Dme-miR-34 Displays 3′ End Heterogeneity

(A) Structure of pre-miR-34. miR-34 (24 nt) is shown in red, and miR-34* (23 nt) is shown in blue.

(B) miR-34 isoforms detected in total RNA from S2 cells and Oregon R flies by northern hybridization. 2S rRNA serves as loading control.

(C) Reads mapping to the pre-miR-34 hairpin in high-throughput sequencing datasets of total small RNAs from S2 cells. Red, miR-34 reads that share the most abundant 5′ end; blue, miR-34*. Read abundance is reported as parts per million (ppm). The length distribution of miR-34 reads sharing the most abundant 5′ end is shown.

To test whether inaccurate processing of pre-miR-34 by Dicer-1 explains miR-34 heterogeneity, we incubated 5′ ^{32}P-radiolabeled pre-miR-34 with purified, recombinant Dicer-1/Loquacious PB, S2 cell lysate or 0–2 hr *Drosophila* embryo lysate for 15 min (Figure 2.2A). In all three conditions, pre-miR-34 was rapidly converted to 24 nt (Dcr-1/Loqs-PB: 61%; S2 cell lysate: 63%; embryo lysate: 60%), 25 nt (Dcr-1/Loqs-PB: 25%; S2 cell lysate: 26%; embryo lysate: 29%), and 23 nt (Dcr-1/Loqs-PB: 13%; S2 cell lysate: 11%; embryo lysate: 11%) products; we observed no isoforms shorter than 23 nt. Thus, the shorter isoforms of miR-34 are unlikely to reflect inaccurate processing of pre-miR-34 by Dicer-1, but a likely consequence of 3′ end trimming after dicing (Figure 2.1C). Supporting this idea, incubation of 5′ ^{32}P-labeled pre-miR-34 or a mature miR-34/miR-34* duplex in 0–2 hr embryo lysate produced 21 to 22 nt isoforms (Figure 2.2B). In contrast, *let-7* was not shortened when incubated in embryo lysate.

Figure 2.2

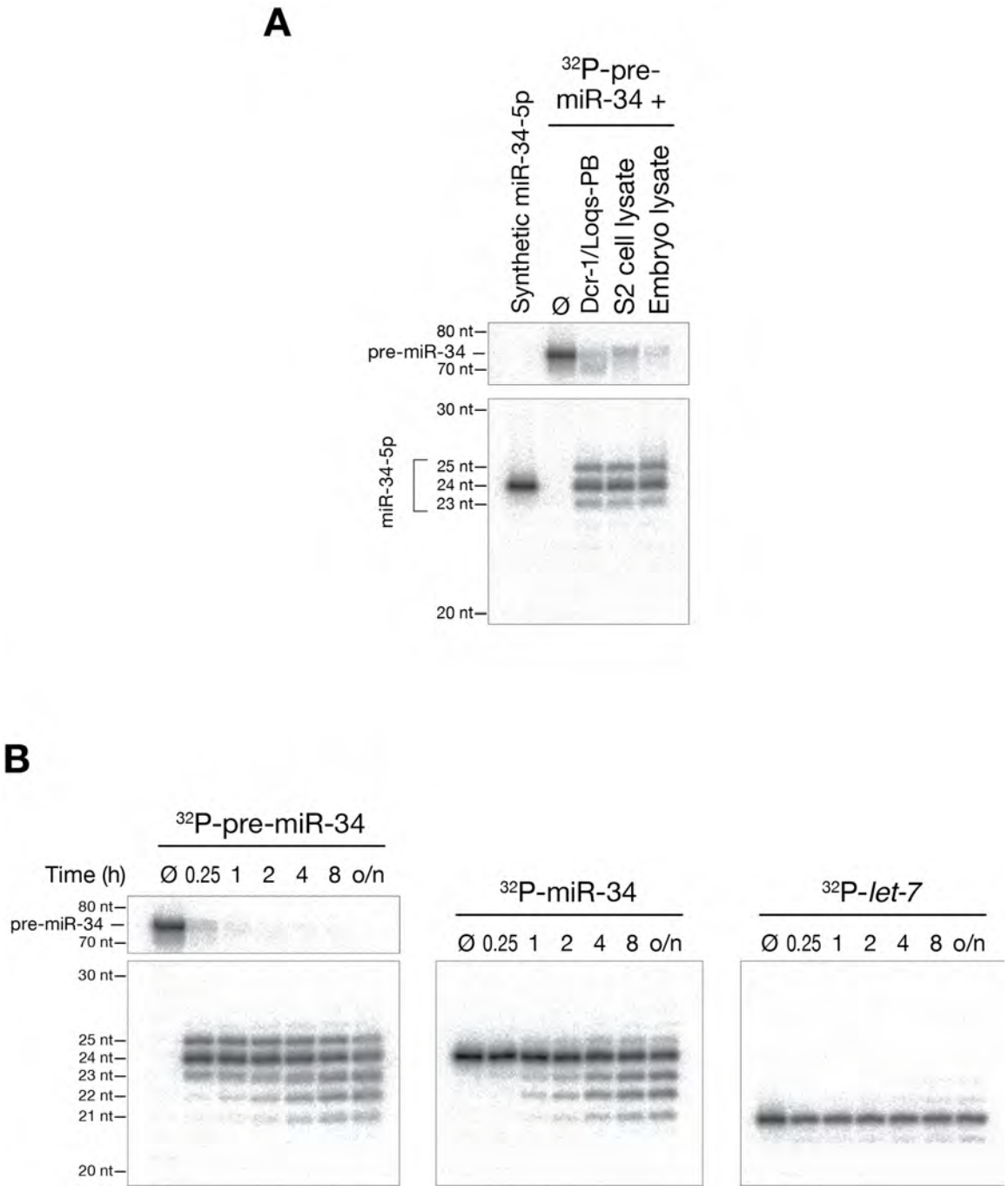

A
Synthetic miR-34-5p
32P-pre-miR-34 +
Ø
Dcr-1/Loqs-PB
S2 cell lysate
Embryo lysate
80 nt
pre-miR-34
70 nt
30 nt
miR-34-5p
25 nt
24 nt
23 nt
20 nt
B
32P-pre-miR-34
Time (h)
Ø 0.25 1 2 4 8 o/n
80 nt
pre-miR-34
70 nt
30 nt
25 nt
24 nt
23 nt
22 nt
21 nt
20 nt
32P-miR-34
Ø 0.25 1 2 4 8 o/n
32P-let-7
Ø 0.25 1 2 4 8 o/n

Figure Legend 2.2. miR-34 is Trimmed After its Production by Dicer-1

(A) 5′ 32P-radiolabeled pre-miR-34 was incubated with purified, recombinant Dicer-1/Loquacious-PB heterodimer (Dcr-1/Loqs-PB), S2 cell lysate, or 0–2 hr embryo lysate. Products were resolved by denaturing polyacrylamide gel electrophoresis.

(B) 5′ 32P-radiolabeled pre-miR-34, 24 nt miR-34, or 21 nt let-7 RNA was incubated in 0–2 hr embryo lysate. Products were resolved by denaturing polyacrylamide gel electrophoresis.

miRNA Trimming Requires Ago1

Trimming of miR-34 might occur immediately after its production by Dicer-1 when miR-34 is still bound to miRNA*, after loading of the miR-34/miR-34* duplex into Ago1 to generate pre-RISC, or following the eviction of miR-34* from pre-RISC to create miR-34-guided Ago1-RISC. To distinguish among these possibilities, we monitored pre-miR-34 processing and miRNA trimming in 0–2 h embryo lysate immuno-depleted of Ago1 (Figure 2.3A). Although pre-miR-34 was efficiently converted into miR-34 in the absence of Ago1, the resulting 23–25 nt Dcr-1 products were not trimmed. In contrast, the miR-34 cleaved from pre-miR-34 was trimmed in lysate containing Ago1 (Figure 2.3A). Similarly, the fraction of trimmed miR-34 decreased in S2 cells depleted of Ago1 by RNAi, compared to the control, when measured by both Northern hybridization (Figure 2.3B) and high throughput sequencing (Figure 2.3C). RNAi depletion of Ago2—the Argonaute protein that binds small interfering RNAs in the RNA interference pathway—had no effect on the amount of trimmed miR-34. We conclude that trimming of miR-34 requires Ago1, presumably because miR-34 trimming occurs after loading into Ago1.

Figure 2.3

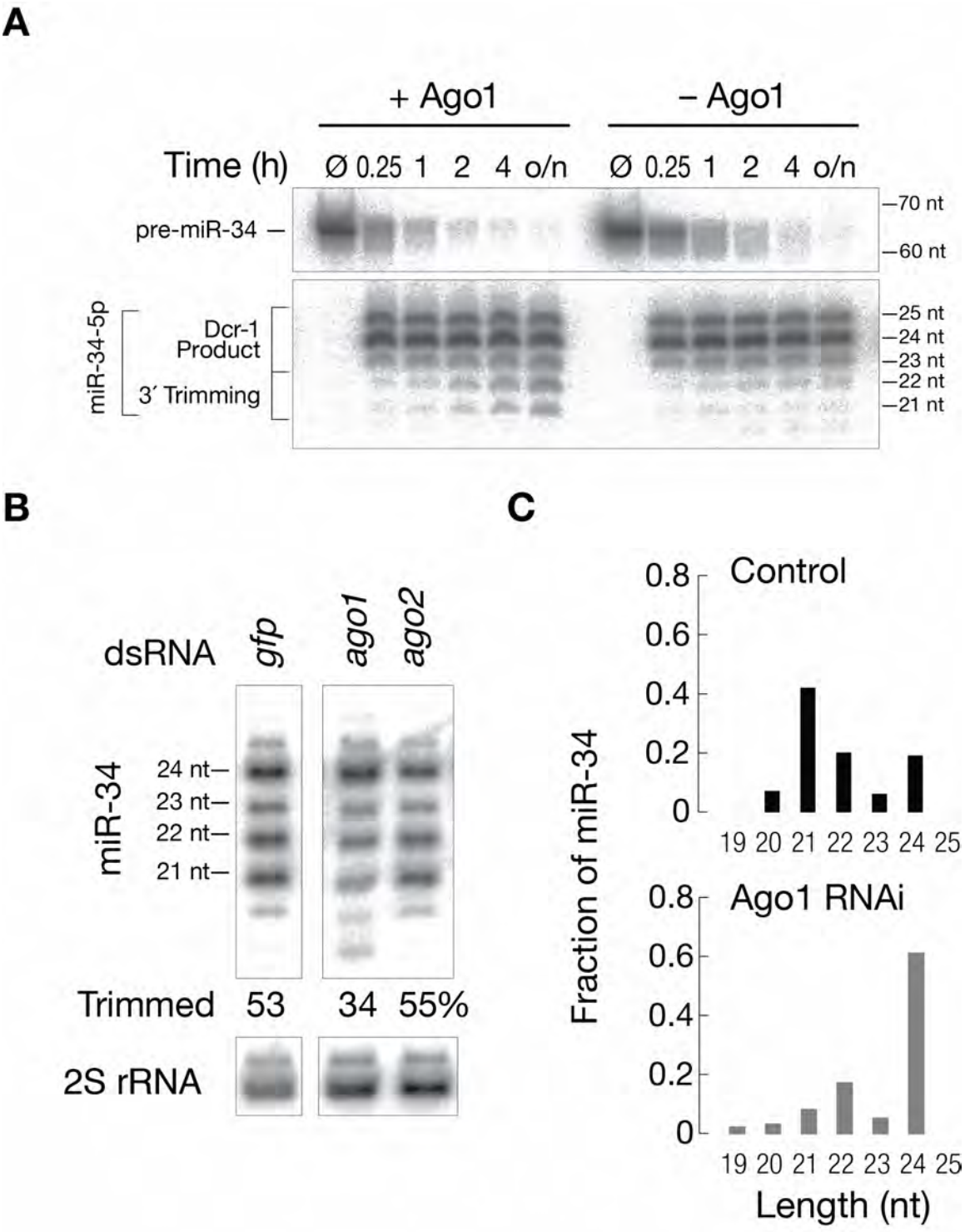

A
+ Ago1
– Ago1
Time (h)
Ø 0.25 1 2 4 o/n
Ø 0.25 1 2 4 o/n
pre-miR-34
70 nt
60 nt
miR-34-5p
Dcr-1 Product
3′ Trimming
25 nt
24 nt
23 nt
22 nt
21 nt
B
dsRNA
gfp
ago1
ago2
miR-34
24 nt
23 nt
22 nt
21 nt
Trimmed
53
34
55%
2S rRNA
C
Control
Ago1 RNAi
Fraction of miR-34
0.8
0.6
0.4
0.2
0
19 20 21 22 23 24 25
Length (nt)

Figure Legend 2.3. Trimming of miR-34 Requires Ago1

(A) 5′ 32P-radiolabeled pre-miR-34 was incubated in 0–2 hr embryo lysate or lysate immune-depleted of Ago1. Products were resolved by denaturing polyacrylamide gel electrophoresis.

(B) Double-stranded RNA (dsRNA)-triggered RNA interference (RNAi) targeting Ago1, but not Ago2, decreased trimming of miR-34, compared to treatment with a control dsRNA targeting GFP. “Trimmed” indicates the fraction of all miR-34 corresponding to 21 and 22 nt isoforms. The 2S ribosomal RNA (rRNA) serves as control.

(C) The fraction of long miR-34 isoforms, measured by high throughput sequencing, increased when S2 cells were depleted of Ago1 by RNAi. Only isoforms with the annotated miR-34 5′ end were analyzed. The abundance of miR-34 in the two libraries was 3,499 ppm (control) and 4,506 ppm (*ago1* RNAi).

miRNA* Strand Dissociation Limits the Rate of miRNA trimming

A key step in the assembly of mature Ago1-RISC is the removal of the miRNA* strand from the Ago1-bound, miRNA/miRNA* duplex, a process that converts pre-RISC to RISC. Mismatches between the miRNA seed sequence and the corresponding nucleotides in the miRNA* promote maturation of pre-Ago1-RISC. We performed in vitro trimming assays using three miR-34/miR-34* duplexes that differ in the strength of pairing of the miR-34 seed sequence to the seed match in miR-34* (Figure 2.4A). One duplex contained a mismatch within the miR-34 seed sequence. A second duplex, included two locked nucleic acid (LNA) ribose modifications within the seed match of miR-34*; LNA modifications increase the strength of base pairing by favoring the C3′ endo ribose conformation found in RNA helices. None of the modifications within the miR-34* strand are predicted to alter the relative thermodynamic stability of the miR-34 versus miR-34* 5′ ends, and therefore preserve the preference to load miR-34 rather than miR-34* into Ago1. The mismatch miR-34* more than doubled the rate of miR-34 trimming (k_{obs} = 5.8 × 10^{-5} nM/sec), compared to the canonical miR-34* (k_{obs} = 2.7 × 10^{-5} nM/sec). In contrast, the miR-34* containing LNA modifications more than halved the rate of trimming (k_{obs} = 1.1 × 10^{-5} nM/sec; Figures 2.4B and 2.4C).

Figure 2.4

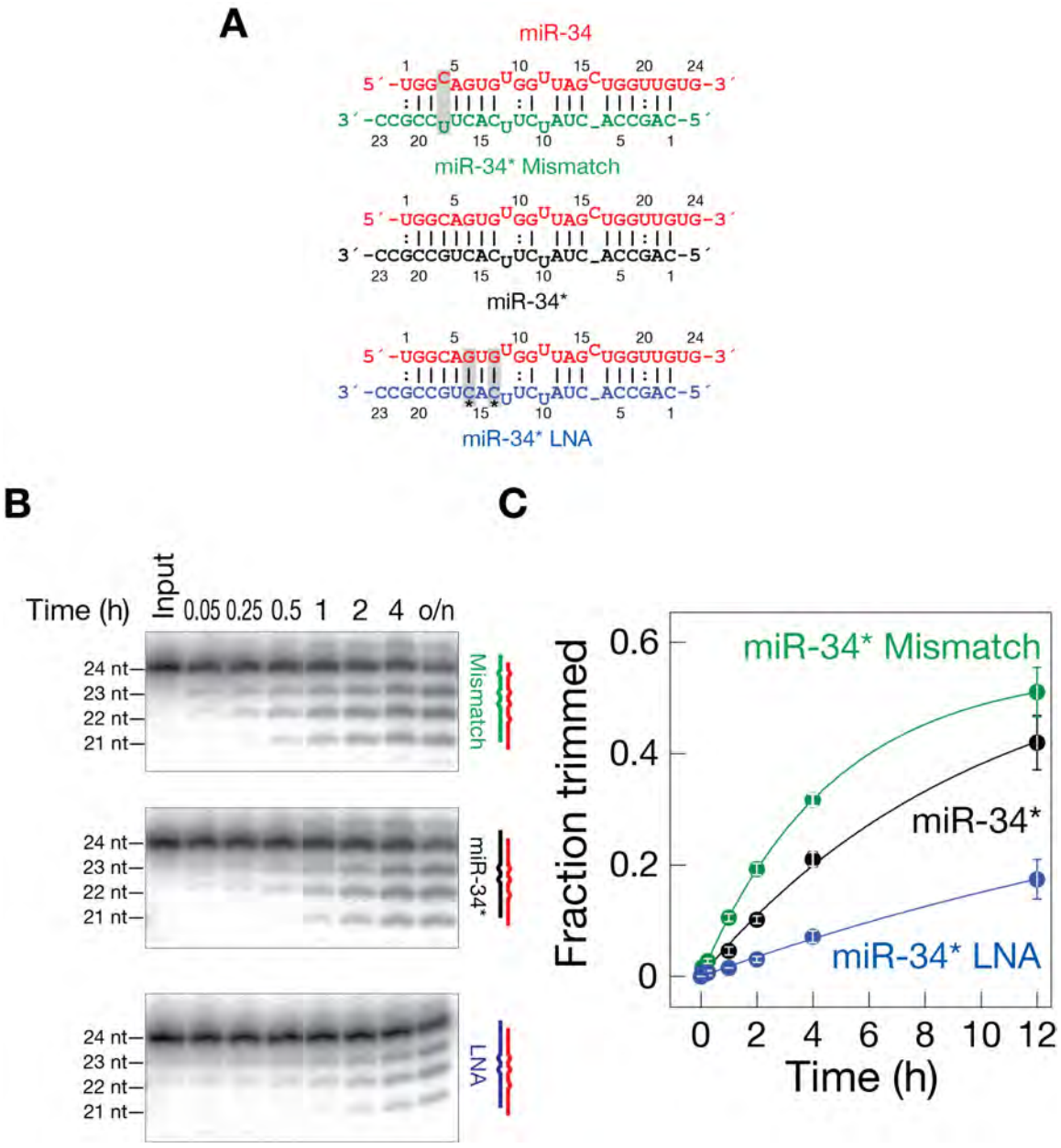

A
miR-34
5′-UGGCAGUGUGGUUAGCUGGUUGUG-3′
3′-CCGCCUUCACUUCUAUC_ACCGAC-5′
miR-34* Mismatch
5′-UGGCAGUGUGGUUAGCUGGUUGUG-3′
3′-CCGCCGUCACUUCUAUC_ACCGAC-5′
miR-34*
5′-UGGCAGUGUGGUUAGCUGGUUGUG-3′
3′-CCGCCGUCACUUCUAUC_ACCGAC-5′
miR-34* LNA
B
Time (h)
Input
0.05 0.25 0.5 1 2 4 o/n
24 nt
23 nt
22 nt
21 nt
Mismatch
miR-34*
LNA
C
Fraction trimmed
0.6
0.4
0.2
0
miR-34* Mismatch
miR-34*
miR-34* LNA
0 2 4 6 8 10 12
Time (h)

Figure Legend 2.4. Trimming of miR-34 is Limited by the miRNA* Removal

(A) Synthetic duplexes of 5′ 32P-radiolabeled miR-34 (red) paired to variants of miR-34*.

(B) miRNA/miRNA* duplexes in (A) were incubated in 0–2 hr embryo lysate, and the products were analyzed by denaturing polyacrylamide gel electrophoresis.

(C) Mean ± standard deviation for three independent replicates of the experiment in (B).

Mismatches between miR-34 and miR-34* also accelerated the rate of destruction of miR-34*, whereas the addition of LNA modifications to miR-34* slowed the decay of miRNA*, compared to an unmodified miR-34* RNA (Figure 2.5). Thus, miR-34* modifications that accelerate RISC assembly also accelerated trimming, whereas modifications that slow RISC assembly also slowed trimming. Our results suggest that miR-34 is first loaded into Ago1 as a 24 nt RNA and is only converted into shorter isoforms after miR-34* is removed from pre-RISC. The majority of 24 nt miR-34 likely corresponds to miR-34 bound to miR-34* in pre-RISC, since the 24 nt isoform, unlike the 21–23 nt isoforms, is not susceptible to target RNA-directed destruction, a process that requires extensive base pairing between the small RNA and its RNA target.

Figure 2.5

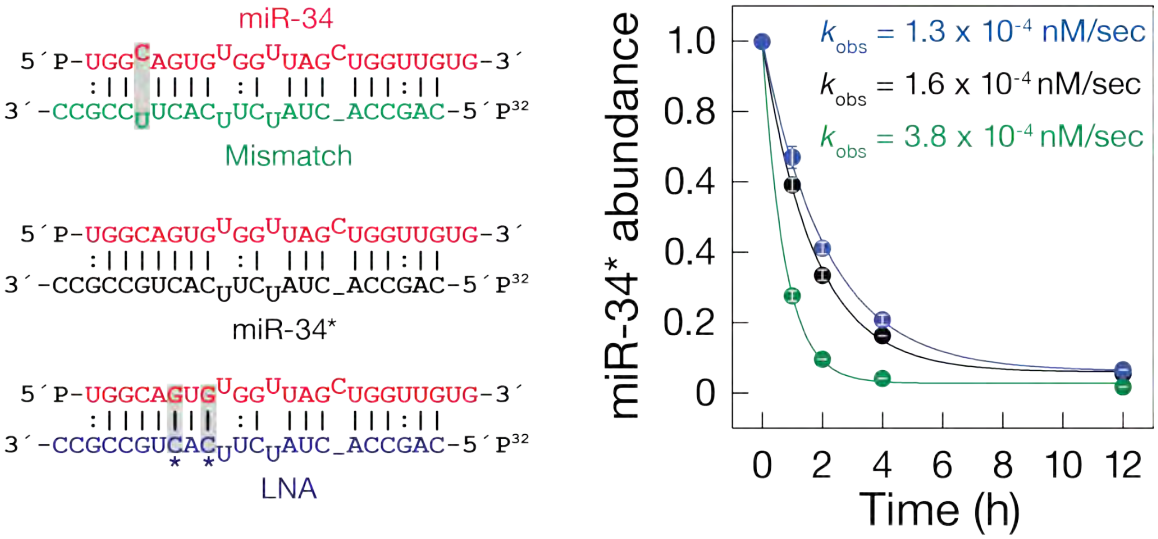

miR-34
5´P-UGGCAGUGUGGUUAGCUGGUUGUG-3´
3´-CCGCCUUCACUCUCUAUC_ACCGAC-5´P32
Mismatch
5´P-UGGCAGUGUGGUUAGCUGGUUGUG-3´
3´-CCGCCGUCACUCUCUAUC_ACCGAC-5´P32
miR-34*
5´P-UGGCAGUGUGGUUAGCUGGUUGUG-3´
3´-CCGCCGUCACUCUCUAUC_ACCGAC-5´P32
LNA
miR-34* abundance
1.0
0.8
0.4
0.4
0.2
0
0 2 4 6 8 10 12
Time (h)
kobs = 1.3 x 10-4 nM/sec
kobs = 1.6 x 10-4 nM/sec
kobs = 3.8 x 10-4 nM/sec

Figure Legend 2.5. Mismatch in the Seed Accelerate miRNA* Destruction

Synthetic duplexes of 5′ 32P-radiolabeled miR-34* variants (green, black, and blue) paired to non-radioactive, phosphorylated miR-34 (red) were incubated in 0–2 h embryo lysate, and then the decrease in abundance of miR-34* was analyzed by denaturing polyacrylamide gel electrophoresis. Mean ± standard deviation for three independent replicates. Decay rates (k_{obs}) were calculated by fitting the data to a single exponential.

The 3′-to-5′ exoribonuclease Nibbler trims miR-34

To identify the exoribonuclease that trims miR-34, we performed a candidate RNAi screen in S2 cells using long double-stranded RNA targeting genes with sequence similarity to known or suspected exoribonucleases. Our screen included *Drosophila* homologs of exonucleases previously implicated in small RNA silencing pathways, such as the small RNA degrading nucleases (SDN) of plants (Ramachandran and Chen, 2008), Enhancer of RNAi-1 (Eri-1; Kennedy et al., 2004) and Mut-7 in *C. elegans* (Ketting et al., 1999), as well as components of the general cellular RNA decay machinery such as RRP4, a core component of the exosome, the SKI-2 ortholog Twister, and the general 5′-to-3′ exonuclease Pacman (XRN1; LaCava et al., 2005). RRP4, Twister and Pacman were previously proposed to degrade the mRNA products generated by RNAi and Xrn-1 was implicated in miRNA turnover (Orban and Izaurralde, 2005). The miRNA *bantam*, which does not undergo detectable trimming served as a control for general destabilization of miRNAs.

Among the exonucleases we tested, only depletion of CG9247 decreased the fraction of trimmed miR-34 (fraction trimmed = 20%), compared to control RNAi (fraction of miR-34 trimmed = 56%; Figure 2.6). We observed a similar loss of miR-34 trimming for two additional, non-overlapping dsRNAs targeting different regions within the second exon and the 3′ untranslated region of CG9247 (Figure 2.7A and 2.7B). In all cases trimming of miR-34 was reduced by

more than half. To reflect its role in 3′ shortening of miRNAs, we named CG9247 *Nibbler* (*nbr*).

Figure 2.6

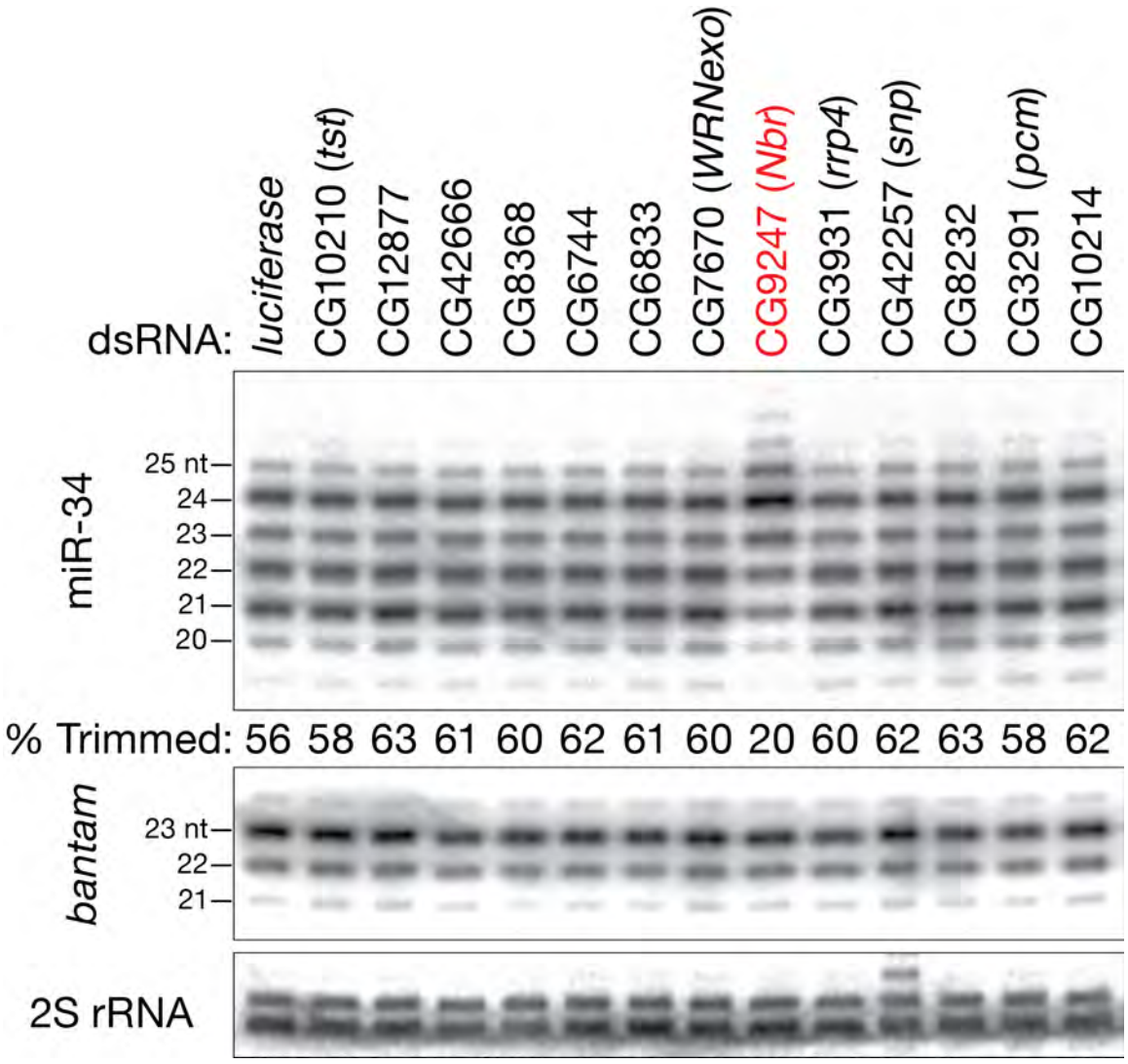

dsRNA:
luciferase
CG10210 (tst)
CG12877
CG42666
CG8368
CG6744
CG6833
CG7670 (WRNexo)
CG9247 (Nbr)
CG3931 (rrp4)
CG42257 (snp)
CG8232
CG3291 (pcm)
CG10214
miR-34
25 nt
24
23
22
21
20
% Trimmed: 56 58 63 61 60 62 61 60 20 60 62 63 58 62
bantam
23 nt
22
21
2S rRNA

Figure Legend 2.6. The 3′-to-5′ Exonuclease CG9247 Trims miR-34

S2 cells were transfected with dsRNA against a panel of predicted exonucleases and the effect on miR-34 length analyzed by high resolution northern hybridization. bantam and 2S rRNA served as controls. The fraction of miR-34 trimmed to 21 to 22 nt is indicated below each lane.

Figure 2.7

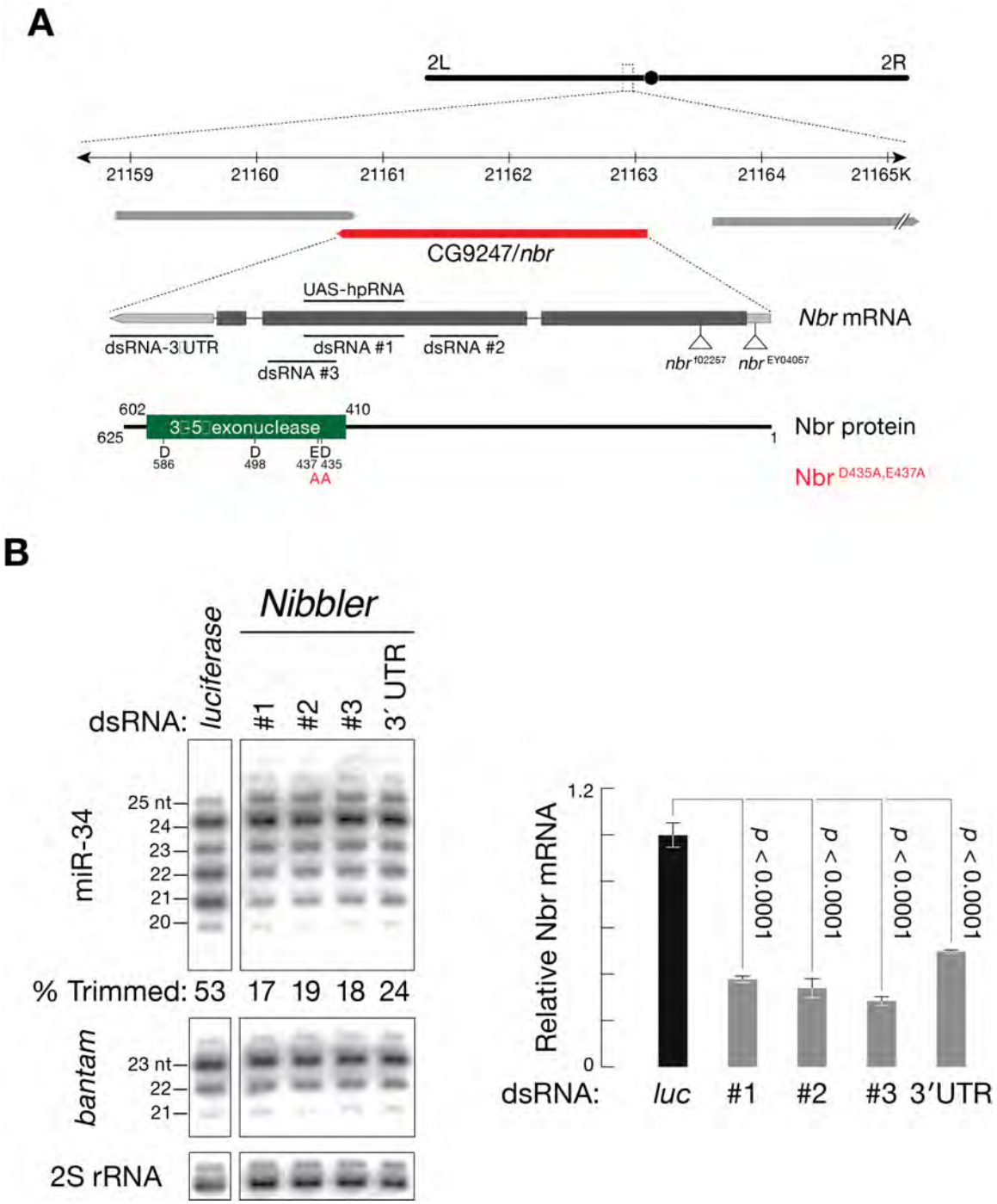

A
2L
2R
21159
21160
21161
21162
21163
21164
21165K
CG9247/nbr
UAS-hpRNA
Nbr mRNA
dsRNA-3′UTR
dsRNA #1
dsRNA #2
dsRNA #3
nbr f02257
nbr EY04057
602
410
625
1
3′-5′ exonuclease
Nbr protein
D 586
D 498
ED 437 435 AA
Nbr D435A,E437A
B
Nibbler
luciferase
dsRNA:
#1
#2
#3
3′ UTR
miR-34
25 nt
24
23
22
21
20
% Trimmed: 53 17 19 18 24
bantam
23 nt
22
21
2S rRNA
Relative Nbr mRNA
1.2
0
p < 0.0001
p < 0.0001
p < 0.0001
p < 0.0001
dsRNA:
luc
#1
#2
#3
3′UTR

Figure Legend 2.7. Nibbler Trims miR-34, Enhancing its Silencing

(A) The predicted structure of the *nibbler* (CG9247) gene, messenger RNA (mRNA), and protein.

(B) S2 cells were transfected with three dsRNAs targeting the second exon or the 3′ UTR of nibbler as indicated in (A). All four dsRNAs decreased miR-34 trimming, relative to a control dsRNA targeting firefly luciferase. bantam and 2S rRNA served as controls. The efficiency of those dsRNAs are determined by qPCR (right panel).

We note that RNAi depletion of *snipper* (*snp*; CG42257) decreased full length 2S rRNA and caused the accumulation of higher molecular weight isoforms of 2S rRNA (Figure 2.6), suggesting that Snipper plays a previously unknown role in the maturation of 2S rRNA, which is generated by the processing of 5.8S rRNA in flies.

Nibbler homologs include Mut-7 in *C. elegans* and EXD3 in humans (Figure 2.8A). *mut-7*, which was one of the very first genes discovered to act in the RNAi pathway, is required for transposon silencing, RNAi, and co-suppression in worms, but no role for *mut-7* in miRNA biogenesis has been reported. Like Mut-7 and EXD3, Nibbler belongs to the DEDD family of exoribonucleases, which are part of a larger superfamily that includes DNA exonucleases as well as the proof-reading domains of many DNA polymerases. DEDD exonucleases contain three characteristic sequence motifs (Figures 2.8A), which include four invariant acidic amino acids (DEDD). The structure of DNA polymerase suggests that these four amino acids organize two divalent metal ions at the catalytic center. Consistent with the view that Nibbler is a metal-dependent DEDD exoribonuclease, miR-34 trimming in fly lysate was inhibited by EDTA; adding additional Mg^{2+} rescued the inhibition (Figure 2.8B).

Figure 2.8

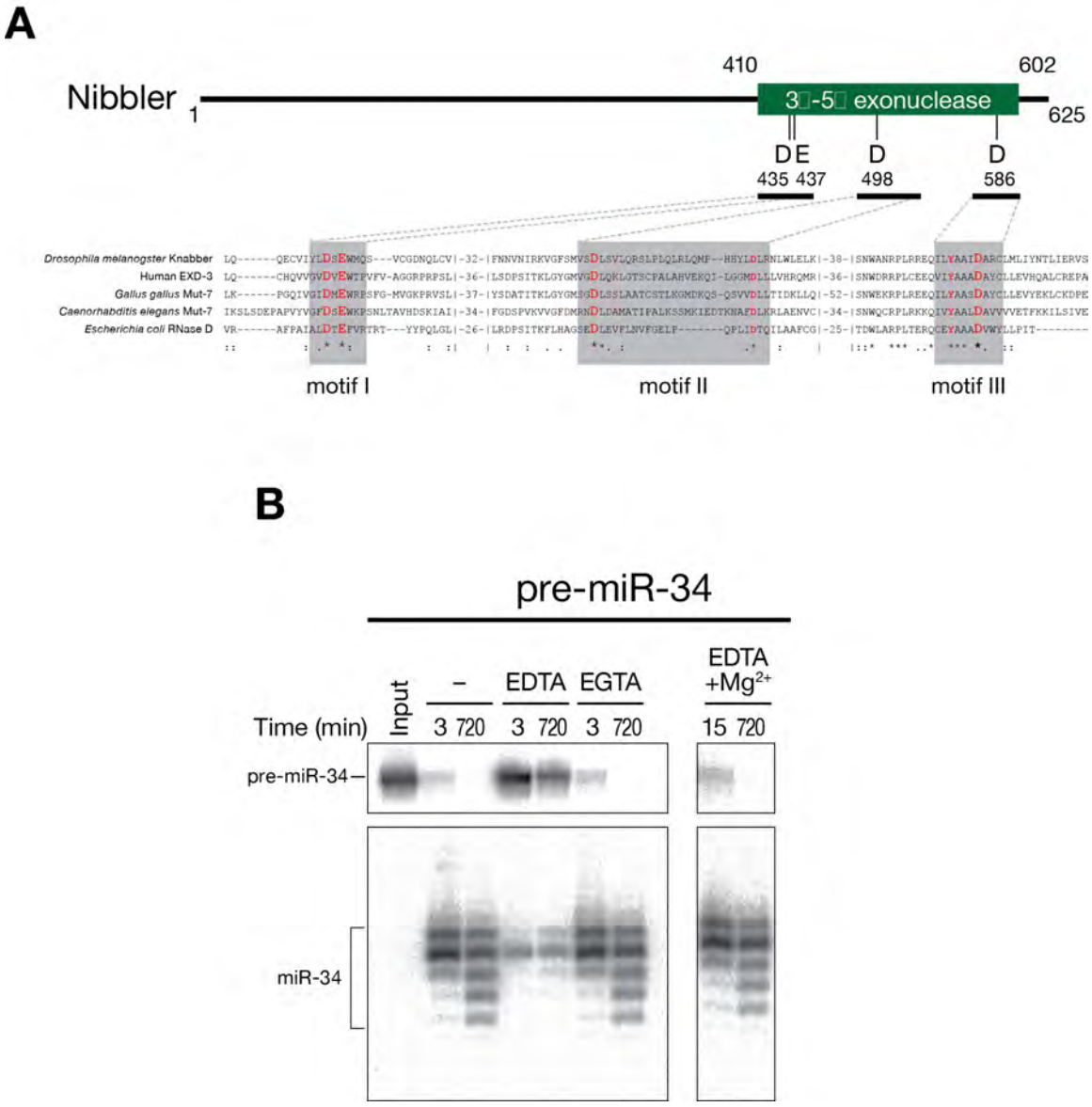

A
Nibbler
1
410
602
625
3′-5′ exonuclease
D E
435 437
D
498
D
586
Drosophila melanogster Knabber
Human EXD-3
Gallus gallus Mut-7
Caenorhabditis elegans Mut-7
Escherichia coli RNase D
motif I
motif II
motif III
B
pre-miR-34
Input
–
EDTA
EGTA
EDTA
+Mg2+
Time (min)
3 720 3 720 3 720
15 720
pre-miR-34
miR-34

Figure Legend 2.8. MicroRNA Biogenesis in Drosophila

(A) Nibbler belongs to the DEDD superfamily of exonucleases. The four invariant exonuclease domain amino acids (DEDD) required for catalysis are indicated. Multiple sequence alignment of three conserved regions of five DEDD family members is shown below. Red, conserved amino acids required for catalysis. (B) Trimming of miR-34 requires Mg^{2+}. 5′ 32P-radiolabeled pre-miR-34 was incubated in 0–2 h embryo lysate containing additional water (control), 5 mM (f.c.) EDTA, 5 mM EGTA, or 5 mM EDTA plus 5 mM Mg^{2+}. Samples were analyzed by denaturing polyacrylamide gel electrophoresis.

We changed two of the four invariant amino acids of the Nibbler DEDD motif to alanine (D435A and E437A; Figure 2.8A), mutations predicted to block exonuclease activity, then reintroduced wild-type or mutant Nibbler open reading frame into S2 cells depleted of endogenous *nibbler* using dsRNA targeting its 3′ UTR (dsRNA 3′ UTR; Figure 2.9A). Nibbler cDNA expression was driven by the constitutive Actin5C promoter. In these experiments, depletion of endogenous *Nibbler* in control S2 cells decreased the fraction of trimmed miR-34 from 54% to 32% (Figure 2.9A); the presence of a stable, wild-type Nibbler transgene enhanced miR-34 trimming (73% trimmed), even after depletion of endogenous *Nibbler* (78% trimmed miR-34). Enhanced miR-34 trimming likely reflects the greater abundance of Nibbler protein in the stable transgenic cell line, since *Nibbler* mRNA levels were ~100-times higher than in control S2 cells (data not shown). In contrast, expression of the D435A, E437A mutant Nibbler protein reduced miR-34 trimming. The fraction of trimmed miR-34 decreased to 16% when transgenic, D435A, E437A mutant Nibbler was expressed along with endogenous Nibbler. The fraction of trimmed miR-34 decreased to 7% when D435A, E437A mutant Nibbler was expressed and endogenous Nibbler was depleted by RNAi.

In cultured *Drosophila* S2 cells, trimming of miR-34 by Nibbler enhanced its target mRNA silencing activity. We compared the repression of a miR-34-regulated *Renilla reniformis* luciferase reporter in S2 cells stably expressing wild-type Nibbler to cell expressing D435A,E437A mutant Nibbler (Figure 2.9B). S2

cells expressing transgenic wild-type Nibbler produced mostly the 21 nt miR-34 isoform, whereas S2 cells stably expressing mutant Nibbler produce predominantly the 24 nt miR-34 isoform (Figure 2.9A). For each cell line, we compared the level of reporter expression when the cells were transfected with a control anti-miRNA 2′-*O*-methyl oligonucleotide to the reporter expression when the cells were transfected with an anti-miR-34 2′-*O*-methyl oligonucleotide. The ratio of anti-miR-34 to control indicated the extent of repression. We observed significantly (*p*-value = 0.003, *n* = 6; Figure 2.9C) greater repression of the miR-34 reporter in the cells expressing wild-type Nibbler, compared to those expressing the mutant protein, indicating that trimming of miR-34 to shorter isoforms enhances its activity. We conclude that trimming of long miRNAs by the Mg^{2+}-dependent, 3′-to-5′ exoribonuclease Nibbler enhances miRNA function.

Figure 2.9

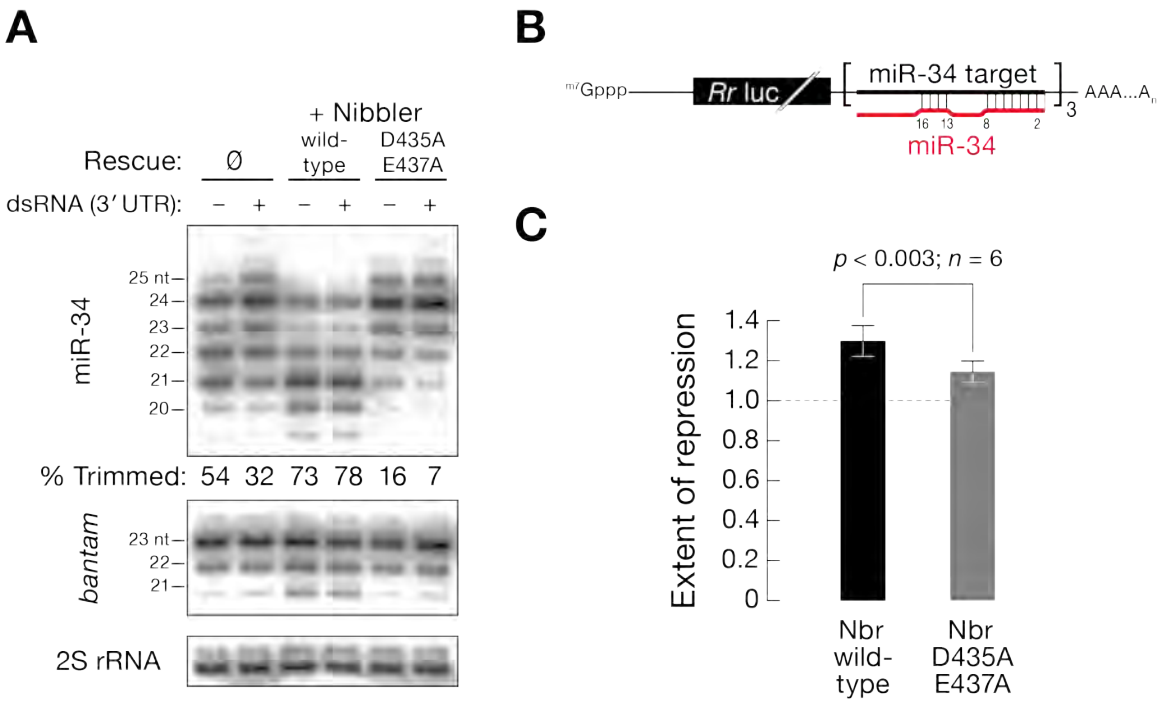

A
+ Nibbler
Rescue:
Ø
wild-type
D435A E437A
dsRNA (3′ UTR):
– + – + – +
miR-34
25 nt
24
23
22
21
20
% Trimmed: 54 32 73 78 16 7
bantam
23 nt
22
21
2S rRNA
B
m7Gppp
Rr luc
miR-34 target
AAA...An
3
16 13 8 2
miR-34
C
p < 0.003; n = 6
Extent of repression
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
Nbr wild-type
Nbr D435A E437A

Figure Legend 2.9. Nibbler Trimming of miR-34 Enhances miRNA Function

(A) S2 cells stably expressing wild-type or D435A,E437A mutant Nibbler CDS were transfected with dsRNA targeting the 3′ UTR of endogenous *nibbler* and the effect on miR-34 trimming measured. *bantam* and 2S rRNA served as controls.

(B) Reporter construct used in (C). The three miR-34 binding sites pair with miR-34 nucleotides 2–8 and 13–15, mimicking typical animal miRNA binding sites. The following abbreviation is used: *Rr* luc, *Renilla reniformis* luciferase.

(C) Nibbler trimming of miR-34 enhances miRNA function. Repression by miR-34 in S2 cells expressing wild-type or D435A,E437A mutant Nibbler was measured by blocking miR-34 using a 2′-*O*-methyl-modified anti-miRNA oligonucleotide and measuring the increase in *Rr* luciferase expression compared to a control oligonucleotide targeting *let-7*, a miRNA not normally expressed in S2 cells.

Nibbler Trims Many miRNAs

To assess the role of Nibbler in the production of other miRNAs, we sequenced 18–30 nt small RNAs from S2 cells treated with *Nibbler* dsRNA and from S2 cells treated with a control dsRNA. S2 cells produce 36 distinct miRNAs that were detected at >200 parts per million (ppm) in our high throughput sequencing. Among the isoforms of these 36 miRNA, we detected a small but statistically significant increase in the overall mean length of miRNAs when Nibbler was depleted: 21.96 nt in the control versus 22.11 nt in *Nibbler (RNAi)* (*p*-value = 3.9×10^{-5}, Wilcoxon signed rank test). If all of miR-34 were 22 nt long in the control and became 24 nt in the *Nibbler* dsRNA-treated cells, the mean miRNA length would be expected to increase by 0.056. Thus, a 0.15 increase in mean length suggests that miR-34 is not the only miRNA trimmed by Nibbler in S2 cells.

In fact, of the 36 abundantly expressed S2 cell miRNAs, 28 increased in mean length. Of these, 13 increased by more than 0.1 nt, and 9 by more than 0.33 nt (Figure 2.10A). We used a χ^2 test to assess the significance of the change in the distributions of isoform lengths in the *Nibbler (RNAi)* S2 cells for each miRNA. An increase of ~0.2 nt in mean length was the smallest change we could corroborate by Northern hybridization, an admittedly less sensitive method than high throughput sequencing. Using the 0.2 nt mean length increase as a conservative threshold, 11 S2 cell miRNAs correspond to Nibbler substrates (red filled circles, Figures 2.10A). Thus, ≥30% of S2 cell miRNAs are trimmed by Nibbler after their production by Dicer-1.

Nibbler substrates included both miRNAs derived from the 5′ arm of their pre-miRNA (4 miRNAs) and miRNAs derived from the 3′ arm of their pre-miRNA (7 miRNAs). miRNAs trimmed by Nibbler account for most of the previously identified 3′ heterogeneity of S2 cell miRNAs, because Nibbler-substrates exhibit significantly higher 3′ heterogeneity than non-substrate miRNAs (p<0.0001, Mann-Whitney U-test, Figure 2.10B). In contrast, 5′ heterogeneity, which is generally low because of the purification process associated with Argonaute loading, was unaffected by the depletion of *Nibbler* by RNAi (data not shown).

The 11 Nibbler substrate miRNAs were significantly longer in S2 cells treated with *Nibbler* dsRNA than non-Nibbler substrate miRNAs: the median of the mean lengths was 23.0 nt for Nibbler substrates versus 21.8 nt for all others (p-value = 0.02, Mann-Whitney U test, Figure 2.10C). In contrast to Nibbler substrate miRNAs, the length of endogenous siRNAs did not change after depletion of Nibbler by RNAi, suggesting that Ago2-bound small RNAs are not Nibbler substrates (data not shown). We also analyzed the effect of Nibbler depletion on the length of the 32 miRNA* strands for which we detected >10 ppm by high throughput sequencing. The overall miRNA* mean length changed from 22.00 to 22.02 nt (p-value = 0.04, Wilcoxon signed rank test), but only two miRNA* strands showed a significant increase in the *Nibbler* dsRNA-treated S2 cells when analyzed using the χ^2 test; neither of the two miRNA* strands increased more than 0.1 nt. Consistent with the proposal that miRNA trimming occurs after miRNA* strands depart from pre-RISC, those miRNA* strands

whose miRNAs were Nibbler substrates did not change significantly in length when compared to all other miRNA*s.

What destines miRNAs for trimming by Nibbler? Perhaps many Nibbler substrate miRNAs are initially produced by Dicer-1 as long isoforms that are trimmed to a more typical miRNA length. To test this idea, we incubated synthetic miR-305/miR-305* duplexes (Figure 2.10E) in *Drosophila* embryo lysate and monitored their trimming. In vivo in flies, miR-305 is efficiently trimmed. Moreover, miR-305 is abundantly expressed and efficiently trimmed in 0–2h embryos: among the 3,668 ppm miR-305 reads detected in the total small RNAs of 0–2h embryos, 23 nt (5%) and 24 nt (14%) miR-305 isoforms represent just 19% of all miR-305 reads, whereas the shorter, trimmed 21 nt (45%) and 22 nt (32%) isoforms represent 77% of all miR-305 reads. When a 24 nt 5′ ^{32}P-radiolabeled miR-305, paired to a 23 nt miR-305* strand, was incubated overnight in embryo lysate, 17% was trimmed to shorter isoforms: 10% accumulated as 23 nt, 5% as 22 nt, and 2% as 21 nt. In contrast, only 2% of a duplex comprising the 22 nt isoform of miR-305 paired to a 22 nt miR-305* strand was converted to a 21 nt form; no species shorter than 21 nt were detectable. We conclude that miRNA trimming is triggered, at least in part, by the length of the miRNA, with ~24 nt miRNAs being converted by Nibbler into the 21–22 nt length, which is more typical for miRNAs at steady-state.

Figure 2.10

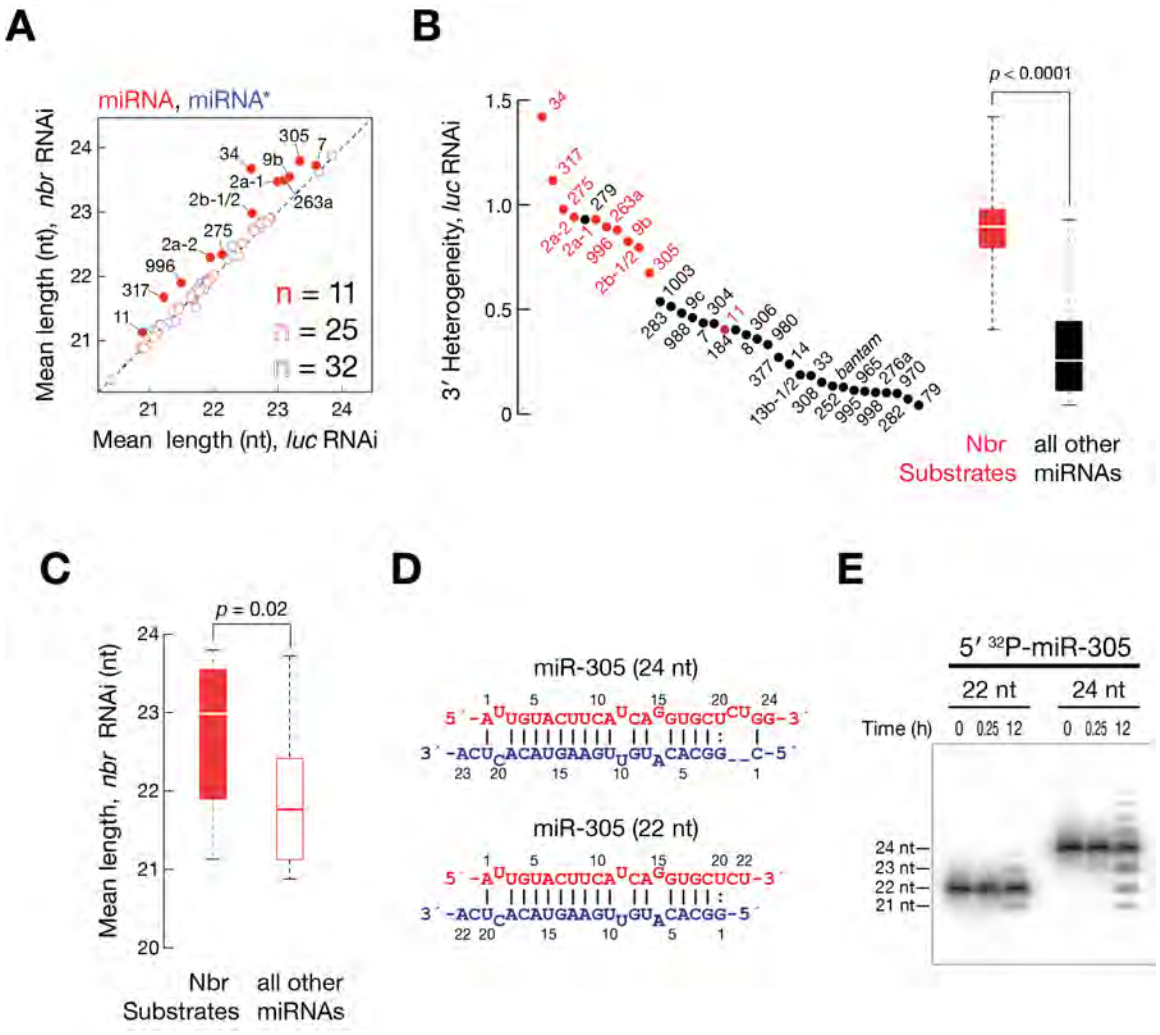

A
miRNA, miRNA*
Mean length (nt), nbr RNAi
Mean length (nt), luc RNAi
n = 11
n = 25
n = 32
B
3′ Heterogeneity, luc RNAi
p < 0.0001
Nbr Substrates
all other miRNAs
C
p = 0.02
Mean length, nbr RNAi (nt)
Nbr Substrates
all other miRNAs
D
miR-305 (24 nt)
miR-305 (22 nt)
E
5′ 32P-miR-305
22 nt
24 nt
Time (h)
0 0.25 12
24 nt
23 nt
22 nt
21 nt

Figure Legend 2.10. Nibbler Trims A Quarter of All miRNAs in S2 Cells

(A) Analysis of mean miRNA and miRNA* length in S2 cells transfected with dsRNA targeting *Nibbler* or a control dsRNA targeting firefly luciferase. miRNA, red; miRNA*, blue; filled circles indicate miRNAs with a significant increase in mean length.

(B) Nibbler trimming explains miRNA 3′ heterogeneity. 3′ heterogeneity was determined for all S2 cell miRNAs that were more abundant than 200 ppm in high throughput sequencing data. Red, the 11 Nibbler substrates identified in this study. Boxplots illustrate 3′ heterogeneity of Nibbler substrate miRNAs (red) versus all other miRNAs (black). *p*-value was determined using the Mann-Whitney U test.

(C) The mean length of Nibbler substrate miRNAs is longer than Nibbler substrate miRNAs in S2 cells treated with *Nibbler* dsRNA. *p*-value was determined using the Mann-Whitney U test.

(D, E) Synthetic miRNA/miRNA* duplexes comprising a 24 or 22 nt 5′ 32P-radiolabeled miR-305 RNA and the corresponding miRNA* strand (D) were incubated in embryo lysate, and the products analyzed by denaturing polyacrylamide gel electrophoresis (E).

Nibbler Trims miRNAs in vivo

To test the role of Nibbler in vivo, we obtained two publicly available *Drosophila* strains bearing a transposon insertion in *Nibbler*: *Nibbler*^{EY04057}, corresponding to a *P-element* insertion in the 5′ UTR of *Nibbler* and *Nibbler*^{f02257}, corresponding to a *piggyBac* insertion in the first exon of *Nibbler* (Figure 2.7A). *Nibbler*^{EY04057} was homozygous viable and showed no change in *Nibbler* mRNA abundance compared to Oregon R or *w*^{1118} control flies (Figure 2.11A). However, our preliminary data suggests that the *Nibbler* mRNAs in *Nibbler*^{EY04057} originate within the *P-element* (data not shown) and may therefore not produce wild-type levels of Nibbler protein. *Nibbler*^{f02257} was homozygous lethal, and *Nibbler*^{f02257} heterozygotes produced 38 ± 24% of the amount of *Nibbler* mRNA present in *w*^{1118} control flies (p = 0.008, Figure 2.11A). The fraction of miR-34 that was trimmed was reduced, albeit slightly, in both mutants: 57% of miR-34 was trimmed in *Sp*/*CyO* control flies, whereas 51% was trimmed in *Nibbler*^{EY04057}/*CyO* and 52% was trimmed in *Nibbler*^{f02257}/*CyO (*Figure 2.11B and 2.11C). Trimmed miR-34 accounted for only 40% of all miR-34 isoforms in *Nibbler*^{EY04057} homozygotes, and in *Nibbler*^{EY04057}/*Nibbler*^{f02257} trans-heterozygotes just 21% of miR-34 was trimmed. We conclude that trimming of miR-34 requires Nibbler in vivo. Similarly, the two *Nibbler* mutations recapitulated the effect on eleven miRNAs identified as Nibbler substrates in S2 cells (data not shown).

Nibbler^{f02257} likely corresponds to a strong allele, but this mutation is not homozygous viable. To test whether loss of Nibbler affects fly development, we

used RNAi to deplete *Nibbler* in vivo. When driven by an Actin5C-Gal4 driver, a UAS-hpRNA transgene on the second chromosome (UAS-hpRNA^{v52550}) reduced the fraction of miR-34 that was trimmed to 43% of all miR-34 isoforms, compared to 68% in flies expressing the Act5C-Gal4 driver alone or to 66% in the flies carrying only the UAS-hpRNA transgene. An insertion of the same hpRNA construct on the third chromosome (hpRNA^{v52612}), reduced the fraction of miR-34 that was trimmed to 13% of all miR-34 isoforms, compared to 60% in flies carrying only the Actin5C-Gal4 driver or 53% in flies bearing only the UAS-hpRNA transgene (Figure 2.11D and 2.11E). Notably, 29% (69 of 239) of the flies expressing UAS-hpRNA^{v52612}, the RNAi transgene with the stronger effect on miR-34 trimming, failed to eclose from their puparia. Only 5% (16 of 327) of the Actin5C-Gal4/*CyO*; *Dr*/*TM3,Sb* control flies and only 2% (5 of 298) of the +; UAS-hpRNAi^{v52612} control flies died as pupae. Although miRNAs regulate fly development and Nibbler acts on miRNAs, we currently cannot exclude the possibility that this pupal lethality reflects a requirement for Nibbler in processing substrates other than miRNAs.

Figure 2.11

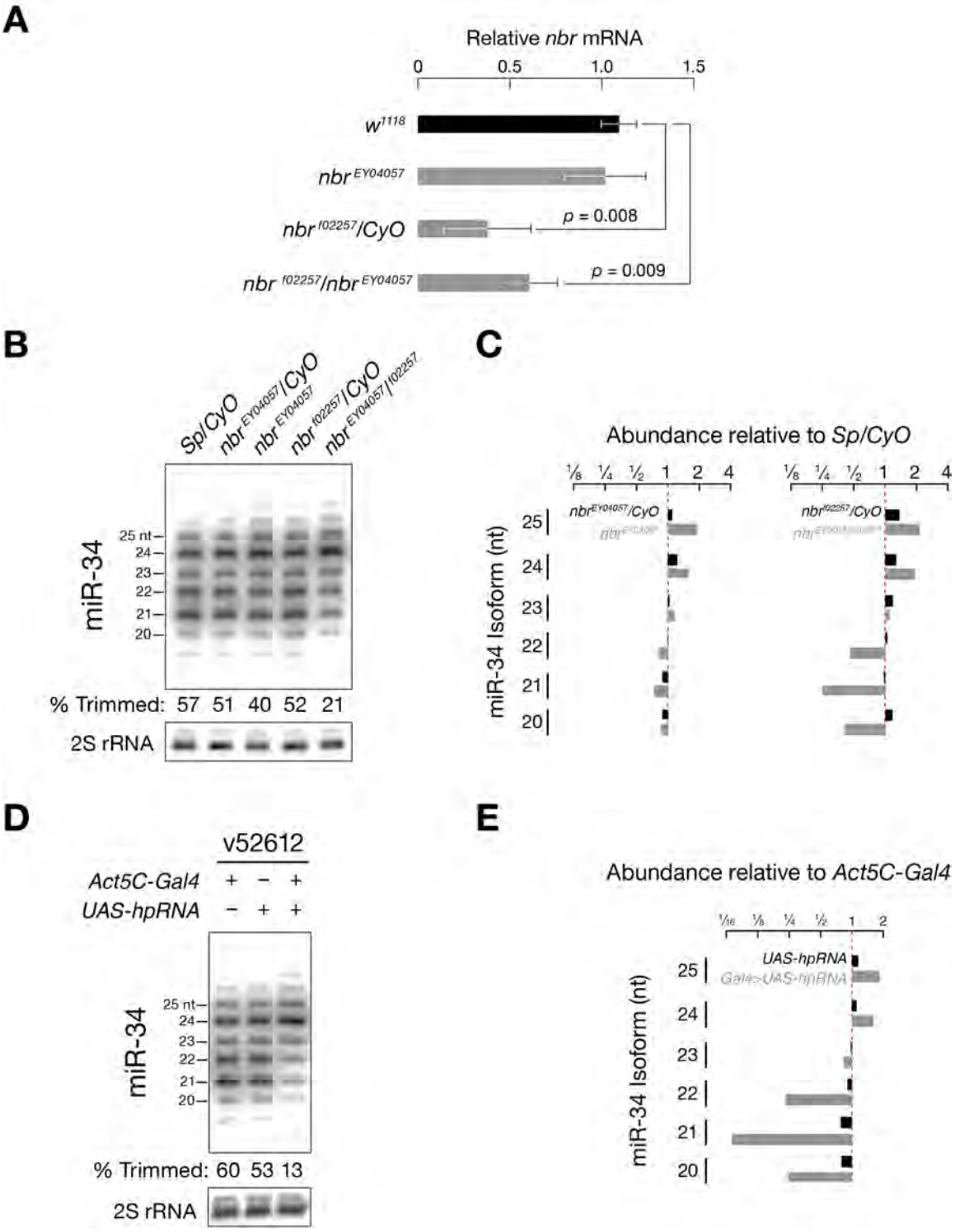

A
Relative nbr mRNA
0
0.5
1.0
1.5
w1118
nbrEY04057
nbrf02257/CyO
p = 0.008
nbrf02257/nbrEY04057
p = 0.009
B
Sp/CyO
nbrEY04057/CyO
nbrEY04057
nbrf02257/CyO
nbrEY04057/f02257
miR-34
25 nt
24
23
22
21
20
% Trimmed: 57 51 40 52 21
2S rRNA
C
Abundance relative to Sp/CyO
1/8 1/4 1/2 1 2 4
nbrEY04057/CyO
nbrf02257/CyO
miR-34 Isoform (nt)
25
24
23
22
21
20
D
v52612
Act5C-Gal4 + − +
UAS-hpRNA − + +
miR-34
25 nt
24
23
22
21
20
% Trimmed: 60 53 13
2S rRNA
E
Abundance relative to Act5C-Gal4
1/16 1/8 1/4 1/2 1 2
UAS-hpRNA
Gal4>UAS-hpRNA
miR-34 Isoform (nt)
25
24
23
22
21
20

Figure Legend 2.11. Nibbler Trims miRNAs in vivo

(A) *Nibbler* mRNA abundance in wild-type and mutant flies. *Nibbler* mRNA levels in 3–5-day-old whole male flies were measured by quantitative RT-PCR. Data were normalized to mRNA levels of ribosomal protein L32 (alternatively called rp49 or RpL32). Mean ± standard deviation for three biological replicates is shown. Student's t-test was used to determine *p*-values.

(B, D) High resolution Northern hybridization of miR-34 from 3–5 day-old male flies carrying a nibbler mutant allele (B) or in which nibbler was depleted by RNAi (D).

(C, E) miR-34 isoform abundance was measured relative to the indicated controls.

Nibbler Trimming Prevents miRNAs from Tailing

Depletion of Nibbler in S2 cells and in flies resulted in the appearance of higher molecular weight species, reminiscent of tailed small RNAs rather than bona-fide Dicer-products (Figure 2.6). Such non-templated addition of nucleotides to the 3′ ends of mature miRNAs has been implicated in miRNA turnover in plants and animals and may indicate that Nibbler-substrate miRNAs are marked for decay when not properly trimmed (Li et al., 2005; Ameres et al., 2010). To examine this idea, we analyzed the tailing of the most abundant 36 miRNAs in S2 cells treated with RNAi against *luc* or *Nbr*. Most of the miRNAs, but not miRNA*s, displayed an increase of tailing (*p*-value = 0.0004, Wilcoxon signed rank test; Figure 2.12A). Among the 11 Nibbler substrates, 10 had their tailed fraction increased (Figure 2.12B). Only miR-79, which had limited tailed species in wild-type S2 cells, failed to show increased in tailing upon *Nbr* knock-down.

Although tailing has been suggested to mark miRNA for degradation by a target-dependent pathway, the increase of tailing in *Nbr* depleted S2 cells failed to correlate with a decrease of miRNA abundance (Figure 2.12C and 2.12D). Among the 36 miRNAs we examined, 8 increased their abundance more than 2-fold while 10 had their levels more than halved.

In summary, our data suggest that Nibbler trimming protects miRNA from non-templated addition of nucleotide to their 3′ end by terminal nucleotidyl transferase.

Figure 2.12

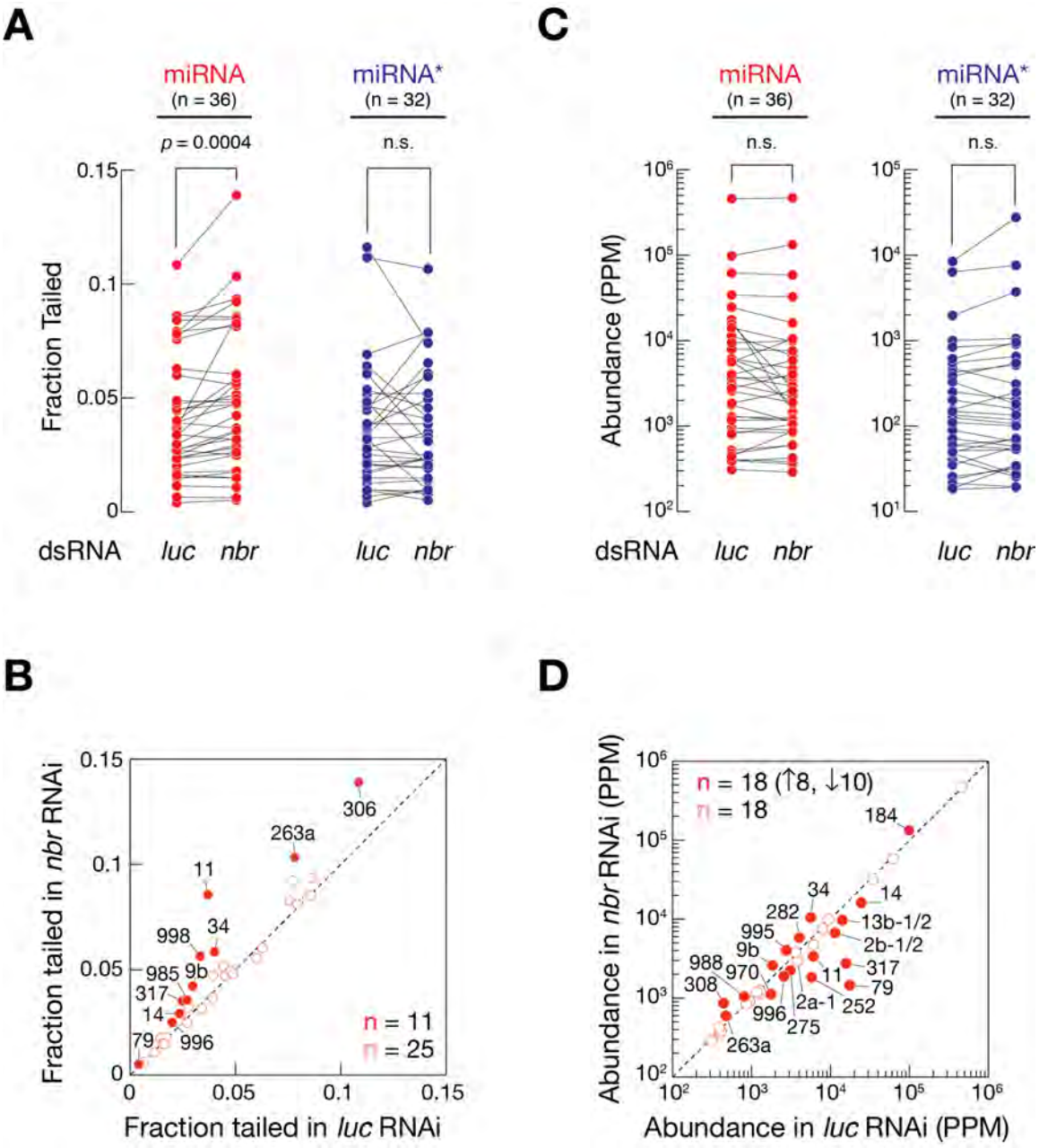

A
miRNA
(n = 36)
miRNA*
(n = 32)
p = 0.0004
n.s.
Fraction Tailed
dsRNA
luc
nbr
C
Abundance (PPM)
B
Fraction tailed in nbr RNAi
Fraction tailed in luc RNAi
n = 11
n = 25
306
263a
11
998
34
985
9b
317
14
79
996
D
Abundance in nbr RNAi (PPM)
Abundance in luc RNAi (PPM)
n = 18 (↑8, ↓10)
n = 18
184
34
14
282
13b-1/2
995
9b
2b-1/2
988
970
11
317
308
2a-1
252
79
996
275
263a

Figure Legend 2.12. Nibbler Trimming Prevents miRNA Tailing

(A, C) Paired dotplot comparing the fraction of tailed reads (A) and abundance (C) of miRNAs and miRNA*s in S2 cells treated with dsRNA against *Nibbler* or luciferase. n.s., not significant.

(B, D) Scatterplot comparing fraction of tailed reads (B) and abundance (D) of miRNAs in S2 cells treated with dsRNA against *Nibbler* or luciferase. Filled circle indicate miRNA with more than 25% of change.

Discussion

miRNA 3′ heterogeneity has been attributed to inaccurate processing by Dicer or Drosha. Our data suggest that much of the 3′ diversity of miRNAs reflects their trimming by a novel processing step catalyzed by the 3′-to-5′ exoribonuclease Nibbler. Figure 2.13 presents a revised model for the production of mature miRNAs from pre-miRNAs in flies. First, Dicer-1 converts pre-miRNAs to miRNA/miRNA* duplexes. These are then sorted between Ago1 and Ago2 to generate Ago1- and Ago-2 pre-RISC complexes, with Ago1 selecting ≥22 nt miRNAs that begin with an unpaired U or A and containing an unpaired region centered on position 9. The Ago1 sorting process helps restrict the diversity of 5′ ends of miRNAs. Next, the miRNA* strand dissociates from pre-RISC to produce RISC. We imagine that the 3′ ends of “long” miRNAs bound to Ago1 are available for trimming by Nibbler because they spend less time bound to the Ago1 PAZ domain than do 22 nt miRNAs. Once Nibbler has shortened a long miRNA to 22 nt, its 3′ end can bind the PAZ domain, protecting it from further trimming or tailing. For miR-34, we observed that trimming enhanced miRNA activity.

This model does not invoke specific recruitment of Nibbler to Ago1-RISC and is consistent with our preliminary experiments, in which we were unable to detect epitope-tagged, over-expressed Nibbler bound to immunoprecipitated Ago1 (data not shown). However, such a simple model cannot explain why some trimmed miRNAs do not accumulate isoforms longer than 22 nt even after Nibbler was depleted by RNAi (e.g., miR-11; Figure 2.10B), suggesting that

miRNA length alone does not define a Nibbler substrate. Perhaps additional proteins help recruit Nibbler to Ago1-RISC for some miRNAs.

Is miRNA-trimming conserved in other organisms? Small RNAs in the human cervical carcinoma cell line HeLa exhibit an overall miRNA 3′ heterogeneity similar to that observed for fly miRNAs. Several human miRNAs with high 3′ heterogeneity show a length distribution in HeLa cells reminiscent of Nibbler-substrates in flies (data not shown). Perhaps a human homolog of fly Nibbler processes these miRNAs. The *C. elegans* homolog of Nibbler, Mut-7, is required for the accumulation of the 22G RNAs that direct worm Piwi proteins to represses transposon expression. We do not yet know if Nibbler functions in the analogous piRNA pathway in flies or if Mut-7 has a yet undiscovered role in miRNA maturation in worms.

Figure 2.13

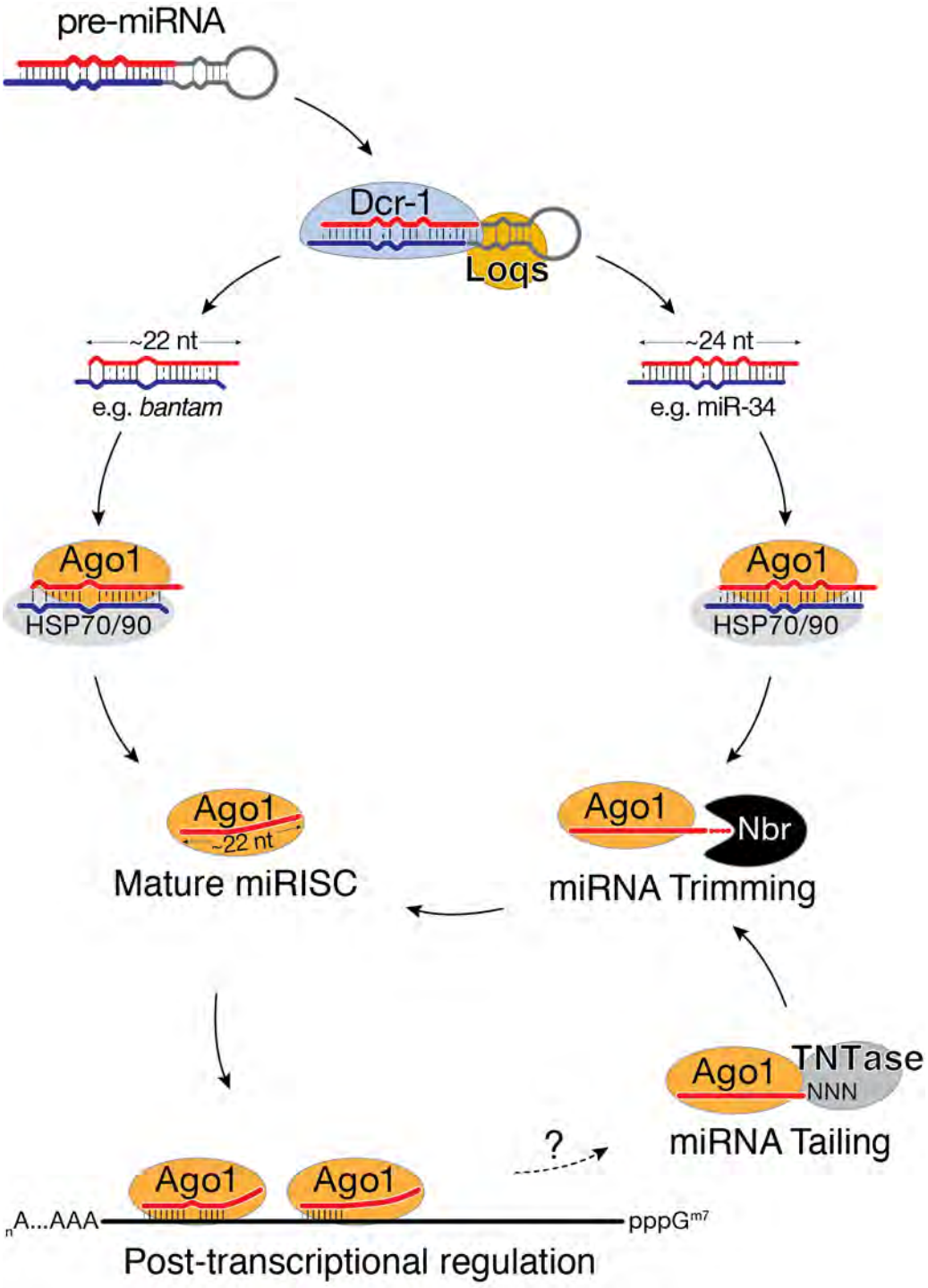

pre-miRNA
Dcr-1
Loqs
~22 nt
e.g. bantam
~24 nt
e.g. miR-34
Ago1
HSP70/90
Ago1
HSP70/90
Ago1
~22 nt
Mature miRISC
Ago1
Nbr
miRNA Trimming
Ago1
TNTase
NNN
miRNA Tailing
?
Ago1
Ago1
nA...AAA
pppGm7
Post-transcriptional regulation

Figure Legend 2.13. Revised Model of MicroRNA Biogenesis in *Drosophila*

See text for more details.

Experimental Procedures

General Methods

Preparation of embryo and S2 cell lysate (Haley et al., 2003), recombinant Dicer-1 and Loquacious-PB (Cenik et al., 2011), clonal S2 cell lines (Ameres et al., 2010), and small RNA libraries for high throughput sequencing (Ghildiyal et al., 2008) have been described previously. Northern hybridization was as described (Ameres et al., 2010), except that *N*-(3-Dimethylaminopropyl)-*N′*-ethylcarbodiimide hydrochloride (Sigma-Aldrich, St. Louis, MO, USA) was used to crosslink 5′ phosphorylated small RNAs to Hybond-NX (Amersham, GE Healthcare, Piscataway, NJ; Pall and Hamilton, 2008). Published small RNA libraries used in this study were total S2 cell RNA and *ago1* RNAi (Czech et al., 2009), total fly head RNA (Ghildiyal et al., 2008), and anti-Ago1 immunoprecipitated small RNAs (Ghildiyal et al., 2010). Sequence data generated in this study are available from the NIH Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo) using accession number GSE31689.

Fly strain $y^{1}\ w^{67c23}$; P{w^{+mC}, $y^{+mDint2}$, EPgy2}CG9247[EY04057] was from the *Drosophila* Stock Center (Bloomington, IN, USA); PBac{WH}CG9247[f02257] was from the Exelixis Collection at Harvard Medical School (Boston, MA, USA); and flies expressing hpRNAs were from the *Drosophila* RNAi Center (Vienna, Austria).

Pre-miRNA Processing and Trimming Assays

Pre-miR-34 was transcribed with T7 RNA polymerase using a double-stranded DNA oligonucleotide template, dephosphorylated with Calf Intestinal Phosphatase (New England Biolabs, Ipswich, MA, USA), and 5′ ^{32}P-radiolabeled with T4 Polynucleotide Kinase (New England Biolabs). Pre-miR-34 (2 nM) was incubated with recombinant Dicer-1/Loquacious PB (5 nM) or S2 cell or 0–2 h embryo lysate for 15 min at 25˚C in a typical RNAi reaction (Haley et al., 2003). Ago1 immuno-depletion was as described (Tomari et al., 2007).

For miRNA trimming, 5′ ^{32}P-radiolabeled RNAs (2 nM) were incubated with 0–2 h embryo lysate as described (Haley et al., 2003), except that RNase inhibitor was omitted. Products were resolved by electrophoresis through a 15% denaturing polyacrylamide sequencing gel. Gels were dried, exposed to storage phosphor screens (Fuji, Tokyo, Japan) and quantified using ImageGauge 4.22 (Science Lab 2003, Fuji).

To analyze miRNA trimming for synthetic 5′ ^{32}P-radiolabeled 24 nt miR-34, all isoforms shorter than 24 nt were considered to be trimmed. When pre-miR-34 was used as a substrate, we considered only isoforms shorter than 23 nt to be trimmed, because Dicer-1 produces 23, 24, and 25 nt miR-34 isoforms from pre-miR-34, so only isoforms shorter than 23 nt could be unambiguously considered to be trimmed. Similarly, we only considered isoforms shorter than 23 nt to be trimmed for Northern hybridization experiments. The fraction of miR-34 trimmed was defined as the sum of trimmed isoforms divided by the sum of all isoforms.

RNAi in S2 cells

Regions targeted by double-stranded RNA were from (Dietzl et al., 2007). DNA templates for in vitro transcription were amplified from genomic DNA or cDNA from Oregon R flies by PCR using primers incorporating the T7 promoter sequence. After isopropanol precipitation, PCR products were used as templates for transcription by T7 RNA polymerase. DsRNA products were purified using MEGA clear RNA purification kit (Ambion, Austin, TX, USA). S2 cells were transfected on day 1 and day 4 with 20 µg dsRNA using Dharmafect4 (Dharmacon, Lafayette, CO, USA), and then total RNA was extracted on day 7 using the mirVana kit (Ambion).

Quantitative RT-PCR

Total RNA purified from S2 cells or flies was treated with Turbo DNase (Ambion), extracted with phenol:chloroform (1:1), and precipitated with 3 volumes ethanol and 1/10th volume sodium acetate (Ambion). Purified RNA was reverse transcribed with SuperScript III (Invitrogen, Carlsbad, CA, USA), and quantitative PCR was performed using SsoFast EvaGreen Supermix (Bio-Rad, Hercules, CA, USA).

Reporter assay

S2 cells stably expressing wild-type or mutant Nibbler were seeded in 24-well plates at 1.0×10^6 cells/ml and transfected immediately after seeding using DharmaFECT Duo (Dharmacon) and 500 ng per well psiCHECK-2 bearing three sites partially complementary to miR-34 in the 3′ UTR of *Rr* luciferase, together

with 20 nM 2′-*O*-methyl-modified oligonucleotide complementary to miR-34 or *let-7*. *Rr* and *Photinus pyralis* luciferase activities were measured 72 h later. Six biological replicates were used to compare the repression conferred by miR-34 for the two cell lines; error was propagated by standard methods. *p*-values were determined using Student's t-test.

Bioinformatics Analyses and Statistics

Insert extraction, mapping and filtering was as described (Ameres et al., 2010), except that after removing the 3′ adaptor and 5′ barcode, only inserts longer than 18 nt were analyzed. 5′ and 3′ heterogeneity was determined as described (Seitz et al., 2008). Briefly, for each miRNA the heterogeneity of the termini of its isoforms was calculated as the mean of the absolute values of the distance between the 5′ or 3′ extremity of an individual read and the most abundant 5′ or 3′ end for that miRNA. For 5′ heterogeneity, all isoforms of a miRNA were examined. For 3′ heterogeneity, only the most abundant 5′ isoforms (i.e., that with the annotated seed sequence) were evaluated.

Acknowledgements

We thank Ryuya Fukunaga for providing recombinant Dicer-1/Loqs PB, Gwen Farley for technical assistance, Alicia Boucher for help with fly husbandry and members of the Zamore lab for support, discussions, and comments on the manuscript. This work was supported by NIH grants GM62868 and GM65236 to P.D.Z. and an EMBO long-term fellowship (ALTF 522-2008) and an Austrian Science Fund (FWF) Erwin Schrödinger-Auslandsstipendium (J2832-B09) to S.L.A. P.D.Z. is a member of the scientific advisory board of Regulus Therapeutics.

Chapter III The Biogenesis of PIWI-interacting RNAs

Disclaimer

This chapter was a product of a collaborative effort among the authors: Bo W Han (BWH), Wei Wang (WW), Chengjian Li (CL), Zhiping Weng (ZW) and Phillip D. Zamore (PDZ). WW first observed piRNA phasing. BWH and CL performed the fly genetics experiments and constructed the next generation sequencing libraries. BWH and WW performed the computational analyses. ZW and PDZ supervised the project.

Summary

In animal gonads, PIWI-interacting RNAs (piRNAs) protect genome integrity by suppressing transposable elements. The current view is that primary piRNAs generated by the endonuclease Zucchini produce secondary piRNAs via the "Ping-Pong" pathway—reciprocal cycles of Aubergine- and Argonaute3-mediated cleavage of transposon mRNAs and piRNA precursor transcripts. Here, we show that secondary piRNAs also initiate the production of primary piRNAs, by feeding Aubergine- and Argonaute3-cleaved RNAs to Zucchini. The first ~26 nt of these cleaved RNAs become secondary piRNAs, while the next ~26 nt become the first in a series of phased primary piRNAs that bind Piwi and Aub, allowing piRNAs to spread beyond the initial site of RNA cleavage. While the Ping-Pong pathway only amplifies the abundance of inherited and de novo piRNAs, the production of phased primary piRNAs from adjacent sequences further introduces novel sequence diversity into the piRNA pool.

Introduction

In animals, PIWI proteins guided by single-stranded, 23–36 nucleotide (nt) small RNAs, PIWI-interacting RNAs (piRNAs), suppress germline transposon expression. In *Drosophila*, piRNAs bind the PIWI proteins, Piwi, Aubergine (Aub) and Argonaute3 (Ago3; Luteijn and Ketting, 2013). Fly primary piRNAs derive from long transcripts from piRNA clusters—discrete genomic loci comprising transposon fragments (Malone et al., 2009). The endonuclease Zucchini (Zuc) is thought to cut cluster transcripts into fragments whose 5′ ends correspond to the 5′ ends of piRNAs, but whose length exceeds that of piRNAs; these piRNA precursors are loaded into Piwi and Aub and then trimmed from their 3′ ends, yielding mature primary piRNAs (Luteijn and Ketting, 2013; Voigt et al., 2012; Nishimasu et al., 2012; Ipsaro et al., 2012). In the fly oocyte, maternally inherited and primary piRNAs made *de novo* initiate production of secondary piRNAs, which subsequently self-amplify via reciprocal cycles of Aub- and Ago3-catalyzed cleavage of transposon mRNAs and cluster transcripts, a process known as the Ping-Pong pathway (Figure 3.1; Brennecke et al., 2007; Gunawardane et al., 2007). The Ping-Pong pathway increases piRNA abundance, but cannot create novel piRNA sequences. Yet piRNA populations are highly diverse, with most individual species of low abundance.

Figure 3.1

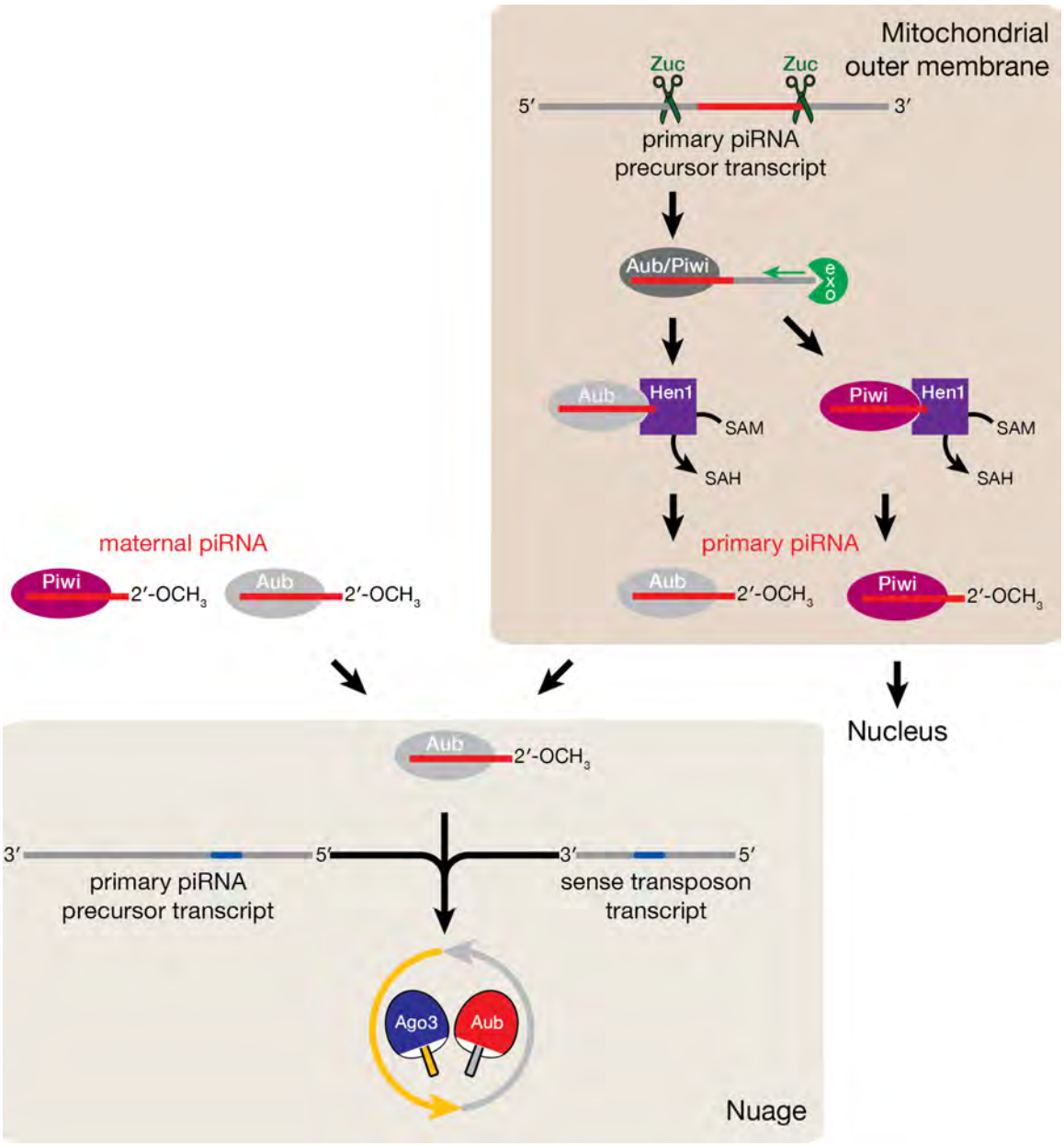

Mitochondrial outer membrane
Zuc
Zuc
5′
3′
primary piRNA precursor transcript
Aub/Piwi
exo
Aub
Hen1
SAM
SAH
Piwi
Hen1
SAM
SAH
maternal piRNA
Piwi
2′-OCH3
Aub
2′-OCH3
primary piRNA
Aub
2′-OCH3
Piwi
2′-OCH3
Nucleus
Aub
2′-OCH3
3′
5′
primary piRNA precursor transcript
3′
5′
sense transposon transcript
Ago3
Aub
Nuage

Figure Legend 3.1. Current Model of piRNA Biogenesis in *Drosophila*

See text for more details.

Results

Phasing of Primary piRNAs

We used genetic mutants to separate primary, maternal, and secondary piRNAs. To assess the mutants' effects on the germ line, we examined piRNAs from the largest piRNA cluster, *42AB* (Brennecke et al., 2007). *aub*$^{HN2/QC42}$*; ago3*$^{t2/t3}$ double-mutants lack the Ping-Pong pathway, so they contain only maternal and primary piRNAs (for *42AB*, Z_{10} = 0.6; *Z*-score ≥ 2.81 corresponds to *p*-value ≤ 0.005; Figure 3.2A). In contrast, *zuc* mutants contain maternal and secondary, but not primary piRNAs. Loss of Zuc decreased *42AB* piRNAs by a factor of 50, but the piRNAs remaining showed significant Ping-Pong amplification (*42AB* piRNAs, Z_{10} = 39; all piRNAs, Z_{10} = 42; Figure 3.2A and 3.2B), consistent with a small pool of maternal piRNAs being amplified into secondary piRNAs.

Figure 3.2

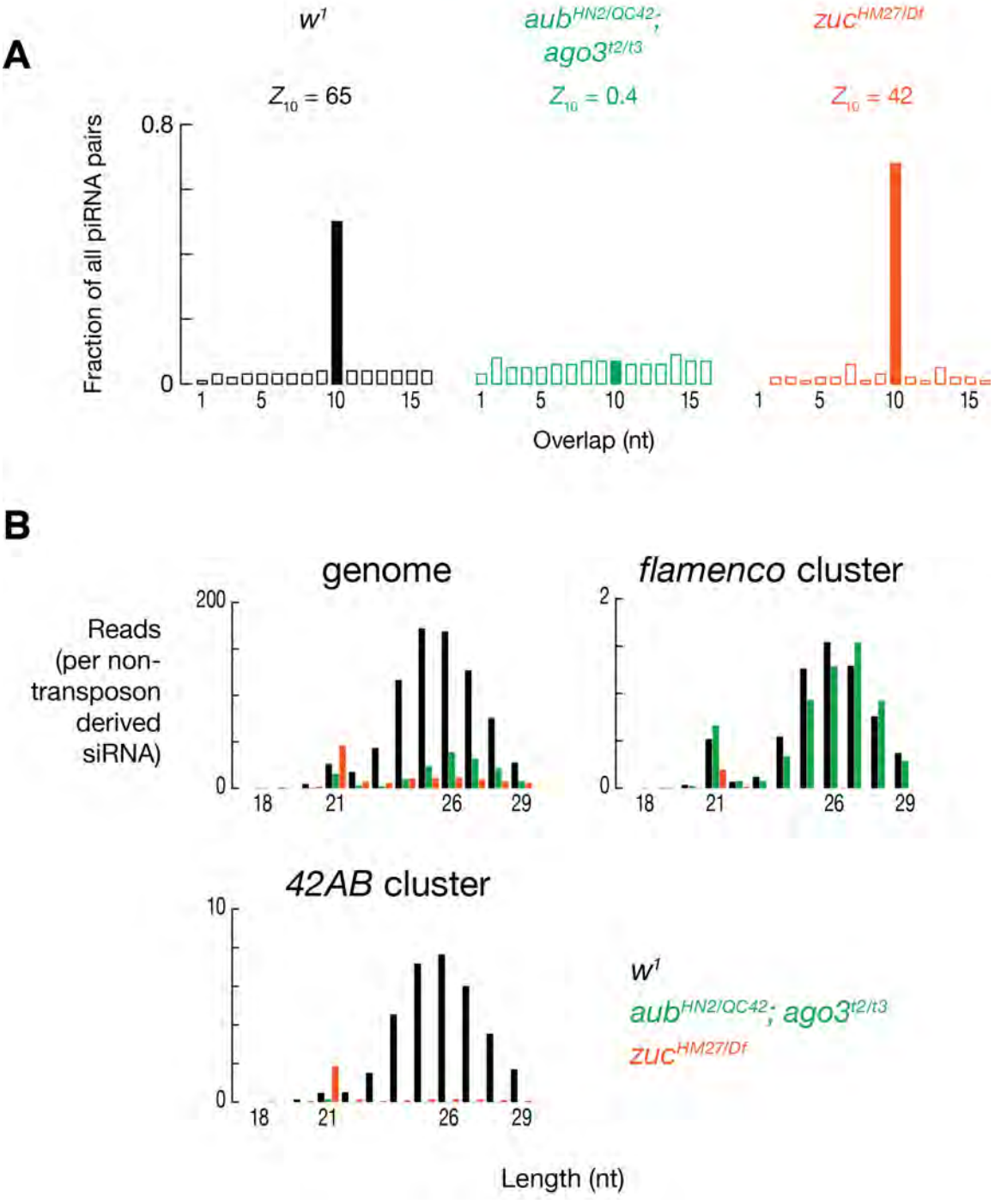

A
w1
aubHN2/QC42; ago3t2/t3
zucHM27/Df
Z10 = 65
Z10 = 0.4
Z10 = 42
Fraction of all piRNA pairs
0.8
0
1
5
10
15
Overlap (nt)
B
Reads (per non-transposon derived siRNA)
genome
flamenco cluster
42AB cluster
200
2
10
0
18
21
26
29
w1
aubHN2/QC42; ago3t2/t3
zucHM27/Df
Length (nt)

Figure Legend 3.2. Separate Primary and Secondary piRNAs in Mutants

(A) Ping-Pong analysis of all piRNAs from w^{1}, $aub^{HN2/QC42}$; $ago3^{t2/t3}$, and $zuc^{HM27/Df}$ ovaries.

(B) Length distribution of genome-, *flamenco*- and *42AB* cluster-derived, uniquely mapping piRNAs from w^{1}, $aub^{HN2/QC42}$; $ago3^{t2/t3}$, and $zuc^{HM27/Df}$ ovaries. Reads were normalized to non-transposon-derived siRNAs, including *cis*-natural antisense transcripts and structured loci.

The 5′ ends of piRNAs mapping to the same genomic strand and present in *aub*$^{HN2/QC42}$*; ago3*$^{t2/t3}$ but not *zuc*$^{HM27/Df}$ typically lay 25–28 nt apart, the same length as piRNAs themselves (Figure 3.3A). Thus, the maternal and primary piRNAs remaining in *aub*$^{HN2/QC42}$*; ago3*$^{t2/t3}$ double-mutant ovaries were phased, suggesting that a nuclease initiates production of piRNAs from one end of a piRNA precursor, moving 5′-to-3′ to clip off successive piRNAs. A broad 5′-to-5′ peak reflects the imprecision in the production of piRNA 5′ ends and impedes statistical analysis. Alternatively, we applied 3′-to-5′ analysis—the distance from the 3′ end of each piRNA to the 5′ end of the next downstream piRNA—to measure piRNA phasing (Figure 3.3B). The most common 3′-to-5′ distance was 1 nt: a single cleavage event appears to produce the 3′ end of one piRNA and the 5′ end of the adjacent, downstream piRNA more often than expected by chance (Z_1 for *w*^{1} = 6.5). Production of phased piRNAs required Zuc but not Ping-Pong: The 1-nt peak was more prominent in *aub*$^{HN2/QC42}$*; ago3*$^{t2/t3}$ ovaries (Z_1 = 22) than in *w*^{1}, but was undetectable in *zuc*$^{HM27/Df}$ (Z_1 = 1.5).

Figure 3.3

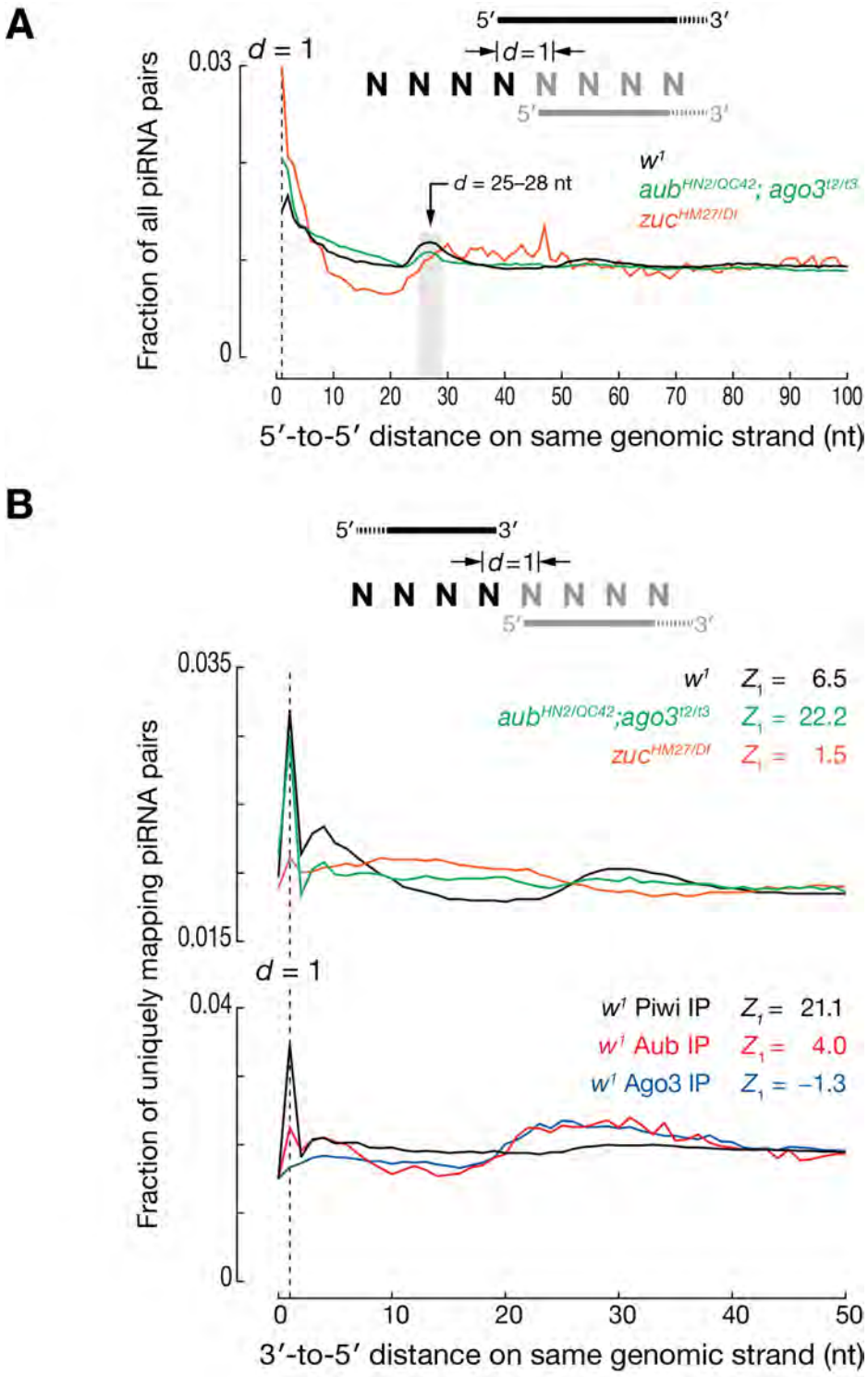

A
5′ 3′
d = 1
d=1
N N N N N N N N
5′ 3′
0.03
0
Fraction of all piRNA pairs
d = 25–28 nt
w1
aubHN2/QC42; ago3t2/t3
zucHM27/DI
0 10 20 30 40 50 60 70 80 90 100
5′-to-5′ distance on same genomic strand (nt)
B
5′ 3′
d=1
N N N N N N N N
5′ 3′
0.035
0.015
0.04
0
Fraction of uniquely mapping piRNA pairs
d = 1
w1 Z1 = 6.5
aubHN2/QC42;ago3t2/t3 Z1 = 22.2
zucHM27/DI Z1 = 1.5
w1 Piwi IP Z1 = 21.1
w1 Aub IP Z1 = 4.0
w1 Ago3 IP Z1 = −1.3
0 10 20 30 40 50
3′-to-5′ distance on same genomic strand (nt)

Figure Legend 3.3. Primary piRNAs Display Phasing

(A) Distance from 5′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs for uniquely mapping piRNAs on the same genomic strand from w^{1}, $aub^{HN2/QC42}$; $ago3^{t2/t3}$, and $zuc^{HM27/Df}$ ovaries.

(B) Distance from the 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs on the same genomic strand. The data are reported as fraction of all piRNA pairs.

Phased piRNAs are more readily detected when piRNAs are abundant, ensuring good genomic coverage. In theory, piRNAs might be phased in *zuc* mutants, but concealed by the low level of piRNA abundance in this mutant. To exclude this possibility, we randomly down-sampled the *42AB*-derived, uniquely mapping, piRNA species from w^1 to the level of *42AB*-derived piRNAs in $zuc^{HM27/Df}$. The reduced set of wild-type piRNAs gave a Z_1 score (6 ± 1) very close to that obtained when using all wild-type piRNAs (Figure 3.4A).

piRNA phasing differed among the three *Drosophila* PIWI proteins (Figure 3.3B). By 3′-to-5′ distance, Piwi-bound piRNAs displayed the most significant phasing (Z_1 = 21); Aub-bound piRNAs displayed reduced, but still significant phasing (Z_1 = 4.0); Ago3-bound piRNAs were not phased (Z_1 = −1.3). Thus, Piwi- and Aub-, but not Ago3-bound primary piRNAs are produced by a processive mechanism that requires Zuc.

piRNAs associated with Piwi and Aub, but not Ago3, typically begin with uridine (Brennecke et al., 2007). Phased piRNAs beginning with U could be produced by a processive nuclease complex measuring out ~26 nt, then cleaving at the nearest U. Alternatively, they could be made by the same nuclease measuring out ~26 nt, but cleaving at all nucleotides with similar efficiency; subsequent binding of Piwi and Aub would select for piRNAs starting with U. The first model predicts that the nucleotide immediately following the 3′ end of a piRNA—in genomic sequence but not mature piRNAs—is more likely to be U than expected by chance. The second model predicts that when one piRNA

follows another in phase, the second piRNA is more likely to begin with U because of the preference of Aub and Piwi; the genomic nucleotide following a piRNA would not have any sequence bias, because selection for a 5′ U follows piRNA precursor cleavage. To distinguish between the models, we measured the composition of the nucleotide after the 3′ ends of piRNAs ("+1U percentage"). This nucleotide was typically uridine in both w^{1} and $aub^{HN2/QC42}$; $ago3^{t2/t3}$, but not in $zuc^{HM27/Df}$ (Figure 3.4B), indicating that phased piRNAs are likely produced by direct cleavage 5′ to U, before pre-piRNAs are loaded into PIWI proteins. Because purified Zuc shows no nucleotide preference (Nishimasu et al., 2012; Ipsaro et al., 2012), we propose that other factors direct Zuc to cleave before U.

Figure 3.4

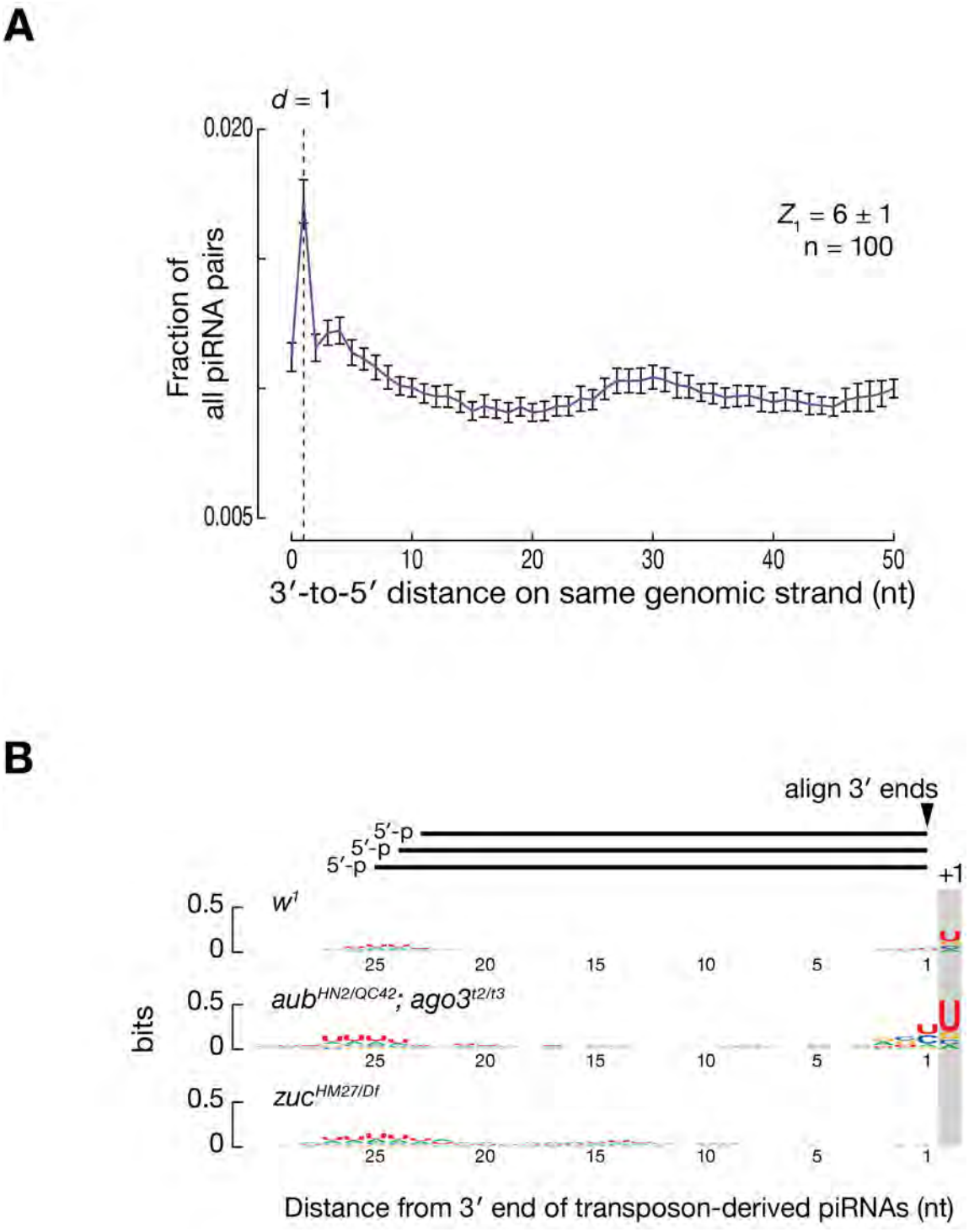

A
d = 1
0.020
0.005
Fraction of all piRNA pairs
Z1 = 6 ± 1
n = 100
0
10
20
30
40
50
3′-to-5′ distance on same genomic strand (nt)
B
align 3′ ends
5′-p
5′-p
5′-p
+1
w1
aubHN2/QC42; ago3t2/t3
zucHM27/Df
0.5
0
bits
25
20
15
10
5
1
Distance from 3′ end of transposon-derived piRNAs (nt)

Figure Legend 3.4. Primary piRNAs Display Phasing

(A) Uniquely mapping, 42AB cluster-derived piRNAs from w^{1} were randomly down-sampled 100× to the number of 42AB cluster-derived, uniquely mapping piRNA species in $zuc^{HM27/Df}$. Then, distance from 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs were calculated for each sample. Error bars report standard deviation.

(B) Nucleotide composition of piRNA species (i.e., distinct sequences irrespective of abundance) 29 nt upstream and 1 nt downstream of the 3′ ends of piRNAs.

Genetic Requirements for piRNA Phasing

Analysis of the phasing of piRNAs derived from the *42AB* cluster in 21 different piRNA pathway mutants or germline RNA interference (RNAi) strains revealed significant piRNA phasing in all mutants except those with defects in the primary piRNA pathway, including *piwi*, *zucchini*, *armitage* (*armi*), *minotaur* and *gasz* (Figure 3.5; Vagin et al., 2006; Pane et al., 2007; Malone et al., 2009; Olivieri et al., 2010; Vagin et al., 2013; Czech et al., 2013; Handler et al., 2013). Mutants defective in piRNA Ping-Pong, including *vasa*, *krimper*, *spindle-E* and *aub*, all displayed more pronounced phasing, likely because the loss of secondary piRNAs reduces the background signal. The presence or absence of piRNA phasing in a mutant accurately predicted the previously defined role of the gene in primary versus secondary piRNA production.

Figure 3.5

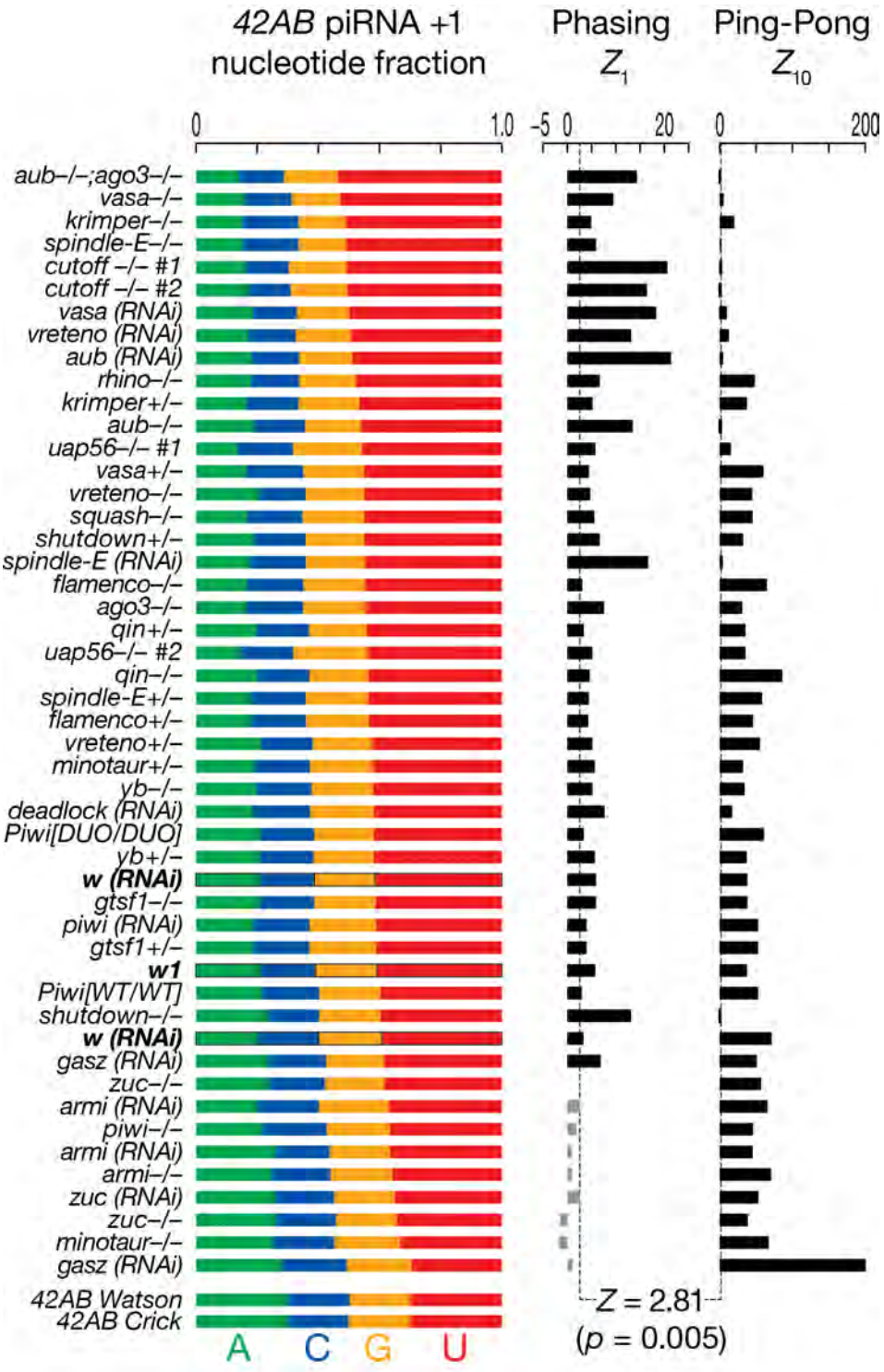

42AB piRNA +1 nucleotide fraction
Phasing Z_1
Ping-Pong Z_{10}
0
1.0
-5 0
20
0
200
aub–/–;ago3–/–
vasa–/–
krimper–/–
spindle-E–/–
cutoff –/– #1
cutoff –/– #2
vasa (RNAi)
vreteno (RNAi)
aub (RNAi)
rhino–/–
krimper+/–
aub–/–
uap56–/– #1
vasa+/–
vreteno–/–
squash–/–
shutdown+/–
spindle-E (RNAi)
flamenco–/–
ago3–/–
qin+/–
uap56–/– #2
qin–/–
spindle-E+/–
flamenco+/–
vreteno+/–
minotaur+/–
yb–/–
deadlock (RNAi)
Piwi[DUO/DUO]
yb+/–
w (RNAi)
gtsf1–/–
piwi (RNAi)
gtsf1+/–
w1
Piwi[WT/WT]
shutdown–/–
w (RNAi)
gasz (RNAi)
zuc–/–
armi (RNAi)
piwi–/–
armi (RNAi)
armi–/–
zuc (RNAi)
zuc–/–
minotaur–/–
gasz (RNAi)
42AB Watson
42AB Crick
A C G U
Z = 2.81
(p = 0.005)

Figure Legend 3.5. piRNA Phasing Requires Primary Pathway Components

Nucleotide composition of piRNA species immediately downstream of the 3′ ends of piRNAs that are uniquely mapped and derived from *42AB* cluster. *Z*-scores for Ping-Pong and phasing are shown. RNAi, germline RNA interference with double-stranded RNA or short hairpin RNA.

We also detected Zuc-dependent phasing in the piRNA cluster *flamenco* and the piRNA-producing 3′ UTR of the protein-coding *traffic jam* mRNA, two loci that produce piRNAs only in the somatic follicle cells that support oocyte development (Figure 3.6A). Cultured, somatic ovarian sheet cells (OSCs), which possess only the primary piRNA biogenesis pathway, also display piRNA phasing. Neither somatic follicle cells nor cultured OSC cells express Aub or Ago3, and both lack a secondary piRNA pathway. Thus, we conclude that phasing is an inherent feature of primary piRNA production.

Figure 3.6

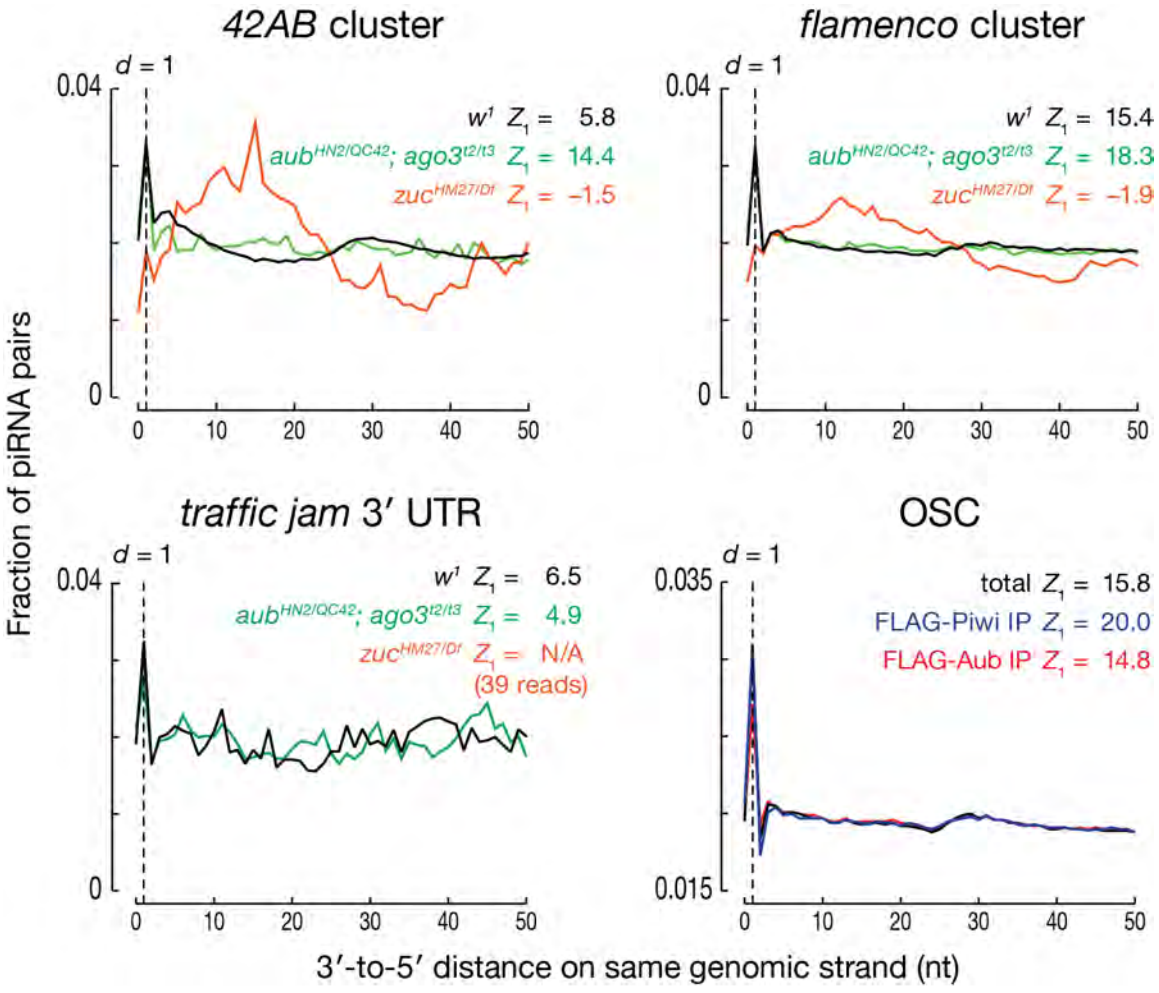

42AB cluster
flamenco cluster
traffic jam 3′ UTR
OSC
d = 1
0.04
0
0.035
0.015
w¹ Z1 = 5.8
aubHN2/QC42; ago3t2/t3 Z1 = 14.4
zucHM27/Df Z1 = –1.5
w¹ Z1 = 15.4
aubHN2/QC42; ago3t2/t3 Z1 = 18.3
zucHM27/Df Z1 = –1.9
w¹ Z1 = 6.5
aubHN2/QC42; ago3t2/t3 Z1 = 4.9
zucHM27/Df Z1 = N/A
(39 reads)
total Z1 = 15.8
FLAG-Piwi IP Z1 = 20.0
FLAG-Aub IP Z1 = 14.8
0 10 20 30 40 50
Fraction of piRNA pairs
3′-to-5′ distance on same genomic strand (nt)

Figure Legend 3.6. Somatic piRNA Display Phasing

Distance from 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs for uniquely mapping piRNAs derived from *42AB*, *flamenco*, the 3′ UTR of *traffic jam* from w^{1}, $aub^{HN2/QC42}$; $ago3^{t2/t3}$, and $zuc^{HM27/Df}$ ovaries. Few *traffic jam*-mapping piRNAs were detected for $zuc^{HM27/Df}$ and were not analyzed. Bottom-right: distance from 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs for all uniquely mapping piRNAs from cultured ovarian somatic cells (OSC), as well as those piRNAs co-purified with FLAG-HA-Piwi and FLAG-HA-Aub expressed in these cells.

Contribution of Maternal piRNAs to Phasing

To test whether the production of phased piRNAs depends on maternal piRNAs, we used a strain bearing a ~7 kilobase pair (kbp) transgene, *P{GSV6}*, inserted into *42AB*. *P{GSV6}* carries both *gfp* and w^{+mC} and produces both sense and antisense piRNAs (Figure 3.7). Both transgene piRNA abundance and Ping-Pong were greater when *P{GSV6}42A18* was inherited maternally (Figure 3.7 and 3.8; Brennecke et al., 2008; de Vanssay et al., 2012; Le Thomas et al., 2014), but primary piRNA phasing was unaltered by the parental source of the transgene (Z_1 maternal = 13; Z_1 paternal = 13). As an additional test of the idea that phased piRNAs are primary, not maternal, we sequenced piRNAs from $vasa^{D5/PH165}$ ovaries that had inherited the *P{GSV6}* transgene maternally or paternally (Figure 3.8A); Vasa is required for Ping-Pong amplification. Regardless of which parent contributed the transgene, *P{GSV6}*-derived piRNAs displayed significant phasing (paternal, Z_1 = 12; maternal, Z_1 = 9.0; wild-type, Z_1 = 13), consistent with the idea that phasing is a primary piRNA signature that requires neither maternal piRNAs nor Ping-Pong amplification.

Figure 3.7

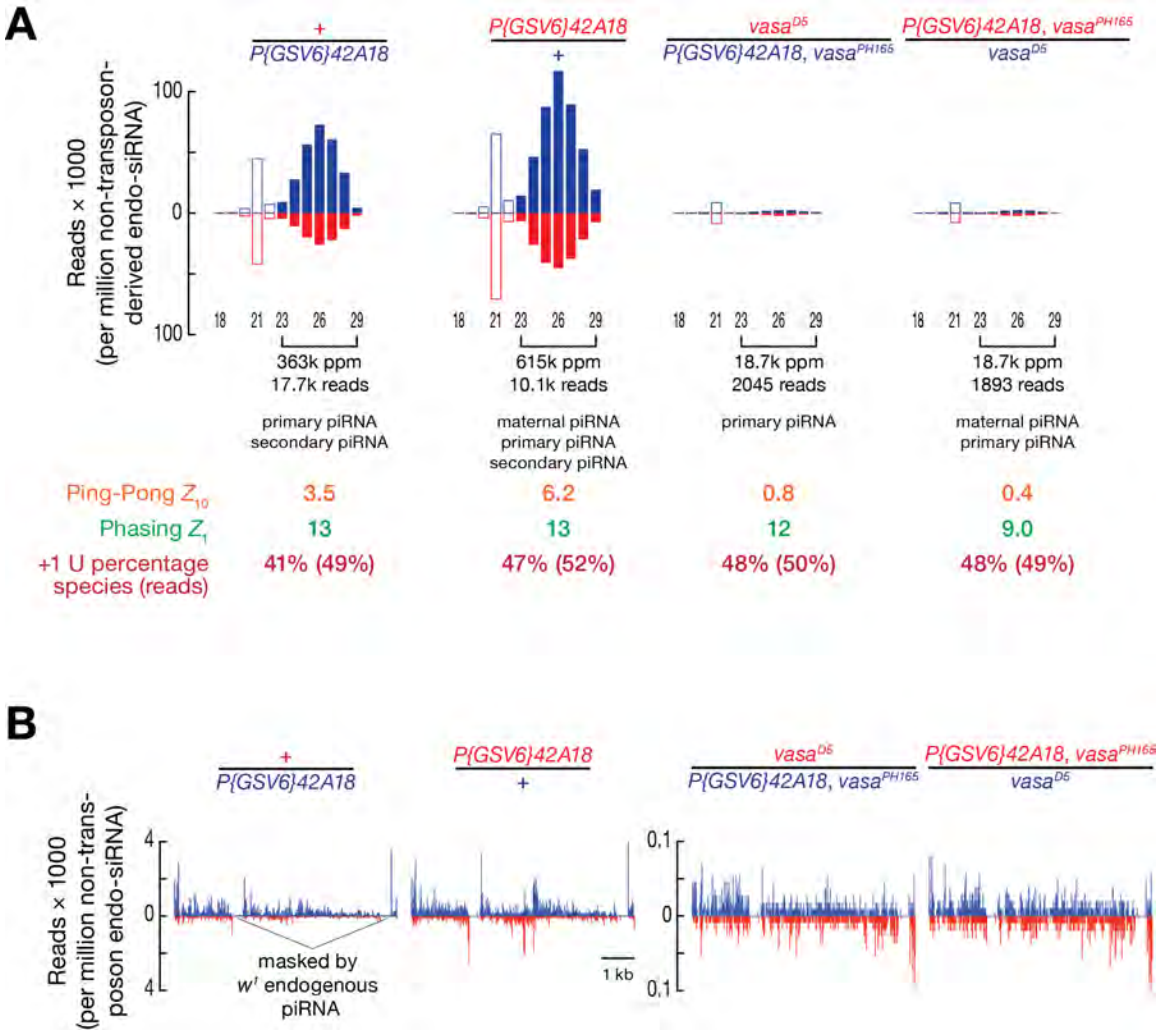

A
+ / P{GSV6}42A18
P{GSV6}42A18 / +
vasa^D5 / P{GSV6}42A18, vasa^PH165
P{GSV6}42A18, vasa^PH165 / vasa^D5
Reads × 1000 (per million non-transposon-derived endo-siRNA)
100
0
100
18 21 23 26 29
363k ppm
17.7k reads
primary piRNA
secondary piRNA
615k ppm
10.1k reads
maternal piRNA
primary piRNA
secondary piRNA
18.7k ppm
2045 reads
primary piRNA
18.7k ppm
1893 reads
maternal piRNA
primary piRNA
Ping-Pong Z10
3.5
6.2
0.8
0.4
Phasing Z1
13
13
12
9.0
+1 U percentage species (reads)
41% (49%)
47% (52%)
48% (50%)
48% (49%)
B
+ / P{GSV6}42A18
P{GSV6}42A18 / +
vasa^D5 / P{GSV6}42A18, vasa^PH165
P{GSV6}42A18, vasa^PH165 / vasa^D5
Reads × 1000 (per million non-trans-poson endo-siRNA)
4
0
4
masked by w1 endogenous piRNA
1 kb
0,1
0
0,1

Figure Legend 3.7. piRNA Production from *P{GSV6}* Inserted in *42AB*

(A) Length distribution, Ping-Pong analysis, phasing Z-score, and +1 U percentage are shown for piRNAs (23–29 nt) from *P{GSV6}42A18* in different genotypes. Red: maternally inherited allele; blue: paternally inherited allele. Reads were normalized to non-transposon-derived siRNAs from *cis*-NATs and structural loci.

(B) piRNA reads from the *P{GSV6}42A18* transgene in wild-type (w^1) or *vasa* mutant ovaries shown according to whether the transgene was inherited maternally or paternally.

Without Vasa, piRNA phasing (Z_1) and the percentage of uridine at the genomic nucleotide immediately after the 3′ ends of the piRNAs (+1U percentage) was unchanged, but the abundance of Piwi-bound piRNAs was less than one-tenth that of wild-type (Figure 3.8). Piwi is likely loaded only with primary piRNAs (Zhang et al., 2011; Sienski et al., 2012; Le Thomas et al., 2014). Why then should Vasa, a central component of the secondary piRNA pathway, affect the abundance of Piwi-bound piRNAs? One explanation is that production of Piwi-loaded, phased primary piRNAs requires precursor cleavage directed by secondary piRNAs.

Figure 3.8

A

Maternal genotype	Paternal genotype	F1 genotype	Maternal piRNA	Primary piRNA	Secondary piRNA	Phasing Z_1	+1 U percentage	Ping-Pong Z_{10}
+/+	+/*P{GSV6}42A18*	+/*P{GSV6}42A18*	No	Yes	Yes	13	41%	3.5
P{GSV6}42A18/+	+/+	*P{GSV6}42A18*/+	Yes	Yes	Yes	13	47%	6.2
vasa/+	+/*P{GSV6}42A18, vasa*	*vasa*/*P{GSV6}42A18, vasa*	No	Yes	No	12	48%	0.8
P{GSV6}42A18, vasa/+	+/*vasa*	*P{GSV6}42A18, vasa*/*vasa*	Yes	Yes	No	9.0	48%	0.4

B

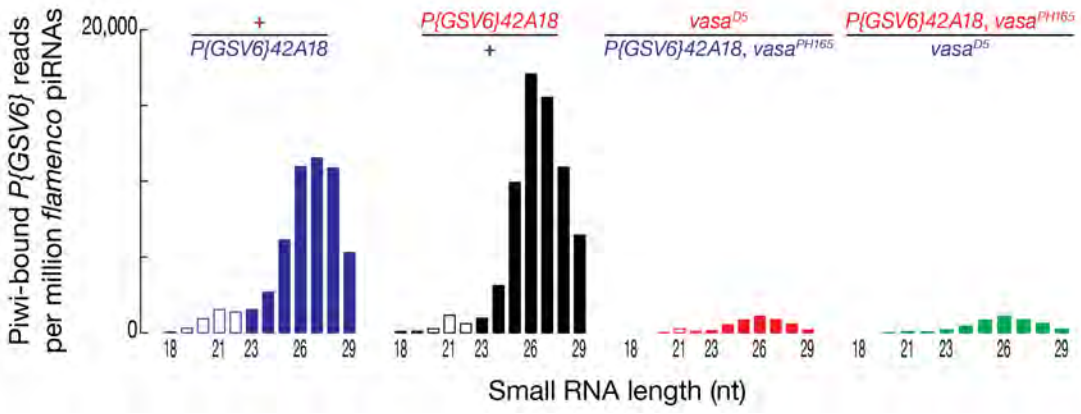


Figure Legend 3.8. Contribution of Maternal and Secondary piRNAs to Phasing

(A) Z-scores for Ping-Pong and phasing and +1 U percentage for *P{GSV6}42A18*-derived piRNAs with the transgene inherited paternally or maternally, with or without *vasa*.

(B) Length distribution of Piwi-bound, uniquely mapping piRNAs derived from *P{GSV6}42A18* in wild-type and *vasa* mutants with the transgene inherited either maternally or paternally. Reads were normalized to *flamenco*-derived, uniquely mapping piRNAs in the same library.

Phasing is Initiated from 5′ Monophosphorylated RNAs

To test this idea, we sequenced the “degradome”—RNAs >200 nt and bearing 5′ monophosphates—to detect the RNAs cleaved by secondary piRNAs bound to Aub or Ago3. In w^{1} control ovaries, we readily identified long transposon RNAs whose 5′ ends were generated by Aub or Ago3 (Figure 3.9A; Wang et al., 2014); such degradome reads were absent from $aub^{HN2/QC42}$; $ago3^{t2/t3}$ mutants (Z_{10} = 0.8). Moreover, degradome reads corresponding to Piwi-catalyzed cleavage were indistinguishable from background (Z_{10} = 0.4), consistent with Piwi silencing via transcriptional repression, rather than RNA cleavage (Sienski et al., 2012). Thus 3′ cleavage products of Aub- or Ago3-catalyzed slicing are subsequently used to produce phased primary piRNAs.

Figure 3.9

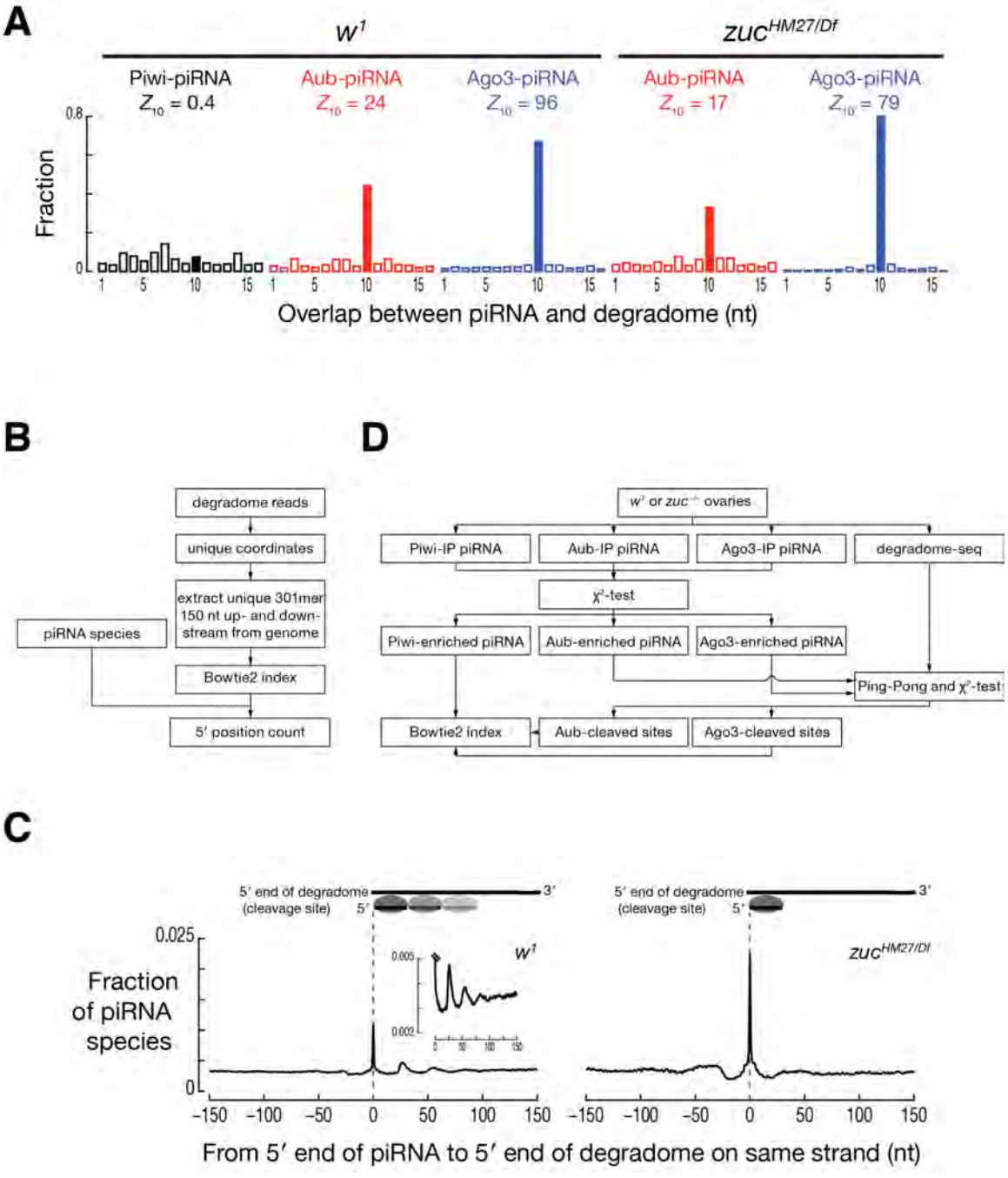

A
w1
zucHM27/Df
Piwi-piRNA
Z10 = 0.4
Aub-piRNA
Z10 = 24
Ago3-piRNA
Z10 = 96
Aub-piRNA
Z10 = 17
Ago3-piRNA
Z10 = 79
Fraction
Overlap between piRNA and degradome (nt)
B
degradome reads
unique coordinates
extract unique 301mer 150 nt up- and down-stream from genome
piRNA species
Bowtie2 index
5′ position count
D
w1 or zuc−/− ovaries
Piwi-IP piRNA
Aub-IP piRNA
Ago3-IP piRNA
degradome-seq
χ2-test
Piwi-enriched piRNA
Aub-enriched piRNA
Ago3-enriched piRNA
Ping-Pong and χ2-test
Bowtie2 index
Aub-cleaved sites
Ago3-cleaved sites
C
5′ end of degradome (cleavage site)
3′
w1
zucHM27/Df
Fraction of piRNA species
From 5′ end of piRNA to 5′ end of degradome on same strand (nt)

Figure Legend 3.9. Degradome-seq Captures Cleavage Products of Aub and Ago3

(A) Ping-Pong analysis between PIWI protein-associated piRNAs and degradome reads.

(B) Computational strategy to measure the distance from the 5′ ends of piRNAs to the 5′ ends of degradome reads.

(C) Distance from 5′ ends of transposon-derived piRNAs to the 5′ ends of degradome reads in w^{1} (left) and $zuc^{HM27/Df}$ (right).

(D) Computational strategy to identify sites cleaved by Aub or Ago3 in degradome-seq data. These sites were then used to calculate the distance to the 5′ ends of nearby PIWI-associated piRNAs.

Phased piRNAs from Aub- and Ago3-cleaved RNAs

To test whether the 3′ cleavage products of Aub- or Ago3-catalyzed slicing are subsequently used to produce phased primary piRNAs, we analyzed the positions of piRNA 5′ ends within the sequences (on the *same* genomic strand) of the cleavage products from transposon transcripts. To accomplish this, we determined the fraction of piRNA 5′ ends at each position 150 nt upstream and 150 nt downstream from the cleavage sites (Figure 3.9B). In both wild-type (w^1) and *zuc* mutant ovaries, the 5′ ends of piRNAs were far more likely to map to the cleavage site than expected by chance (w^1, $Z_0 = 27$; $zuc^{HM27/Df}$, $Z_0 = 34$; Figure 3.9C). The Ping-Pong model predicts this result: it posits that the 5′ termini of Aub- and Ago3-cleaved RNAs subsequently become the 5′ ends of secondary piRNAs. However, two additional peaks of piRNA 5′ ends were present ~26 nt and ~53 nt downstream of the cleavage sites. That is, the 5′ end of a piRNA lies immediately after the 3′ end of the secondary piRNA (i.e., ~26 nt from the cleavage site), and the 5′ end of another piRNA follows the 3′ end of that piRNA (i.e., ~53 nt from the cleavage site). The ~26 and ~53 nt peaks were readily detected in wild-type, but not in $zuc^{HM27/Df}$ ovaries. The requirement for Zuc suggests that the production of a single secondary piRNA from the 5′ end of an RNA cleaved by Aub or Ago3 is followed by the processing of the downstream sequence into phased primary piRNAs.

We separated degradome reads based on the likelihood (p-value ≤ 0.005, χ^2 test) that they were produced by Aub versus Ago3 (Figure 3.9D; Wang et al.,

2014), then analyzed the distance between the 5′ ends of Piwi-bound piRNAs and the sites of Aub- or Ago3-catalyzed cleavage. The 5′ ends of Piwi-bound piRNAs coincided with the Zuc-dependent ~26 and ~53 nt peaks for both Aub- and Ago3-cleaved RNAs (Figure 3.10). A small but significant fraction of Aub-, but not Ago3-bound piRNAs also began ~26 and ~53 nt after the Ago3-cleaved sites.

Figure 3.10

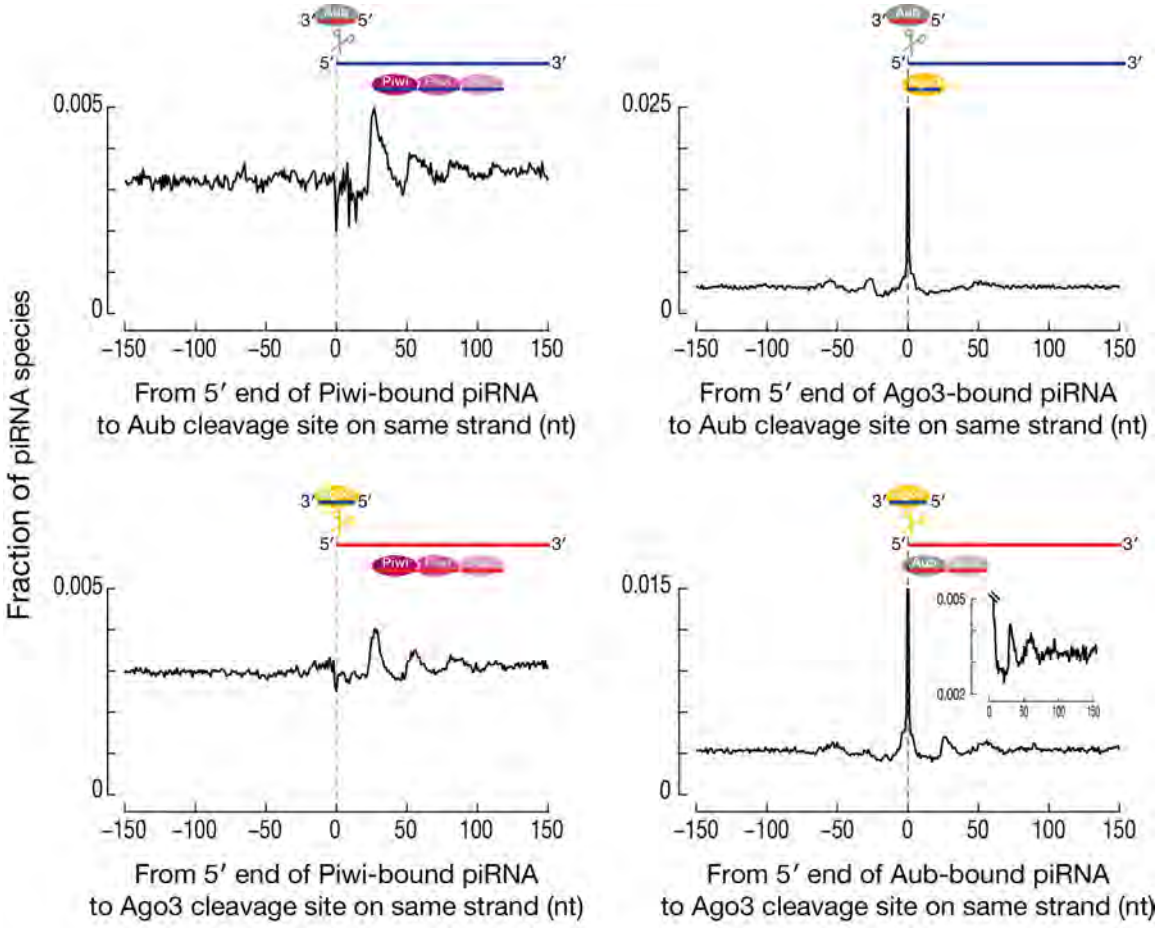

Fraction of piRNA species
0.005
0
−150 −100 −50 0 50 100 150
From 5′ end of Piwi-bound piRNA
to Aub cleavage site on same strand (nt)
0.025
0
−150 −100 −50 0 50 100 150
From 5′ end of Ago3-bound piRNA
to Aub cleavage site on same strand (nt)
0.005
0
−150 −100 −50 0 50 100 150
From 5′ end of Piwi-bound piRNA
to Ago3 cleavage site on same strand (nt)
0.015
0.005
0.002
0
−150 −100 −50 0 50 100 150
From 5′ end of Aub-bound piRNA
to Ago3 cleavage site on same strand (nt)

Figure Legend 3.10. Phasing of Piwi-piRNAs downstream of the cleavage sites of Aub and Ago3 in *w*^{1}.

The distance between the 5′ ends of Piwi- (left), Ago3- (top-right), and Aub- (bottom-right) bound piRNAs and the cleavage sites of Aub (top) and Ago3 (bottom) on the same genomic strand in *w*^{1}.

Small RNA and degradome sequencing data from *zuc* mutant ovaries unambiguously identified sites cleaved by Aub or Ago3. Using this data, the 5′ ends of Piwi-bound piRNAs in w^{1} ovaries were typically ~26 and ~53 nt downstream from where Aub or Ago3 cleaved (Figure 3.11A), a relationship not detected using degradome data from $aub^{HN2/QC42}$; $ago3^{t2/t3}$ ovaries, which lack secondary piRNAs (Figure 3.11B).

Figure 3.11

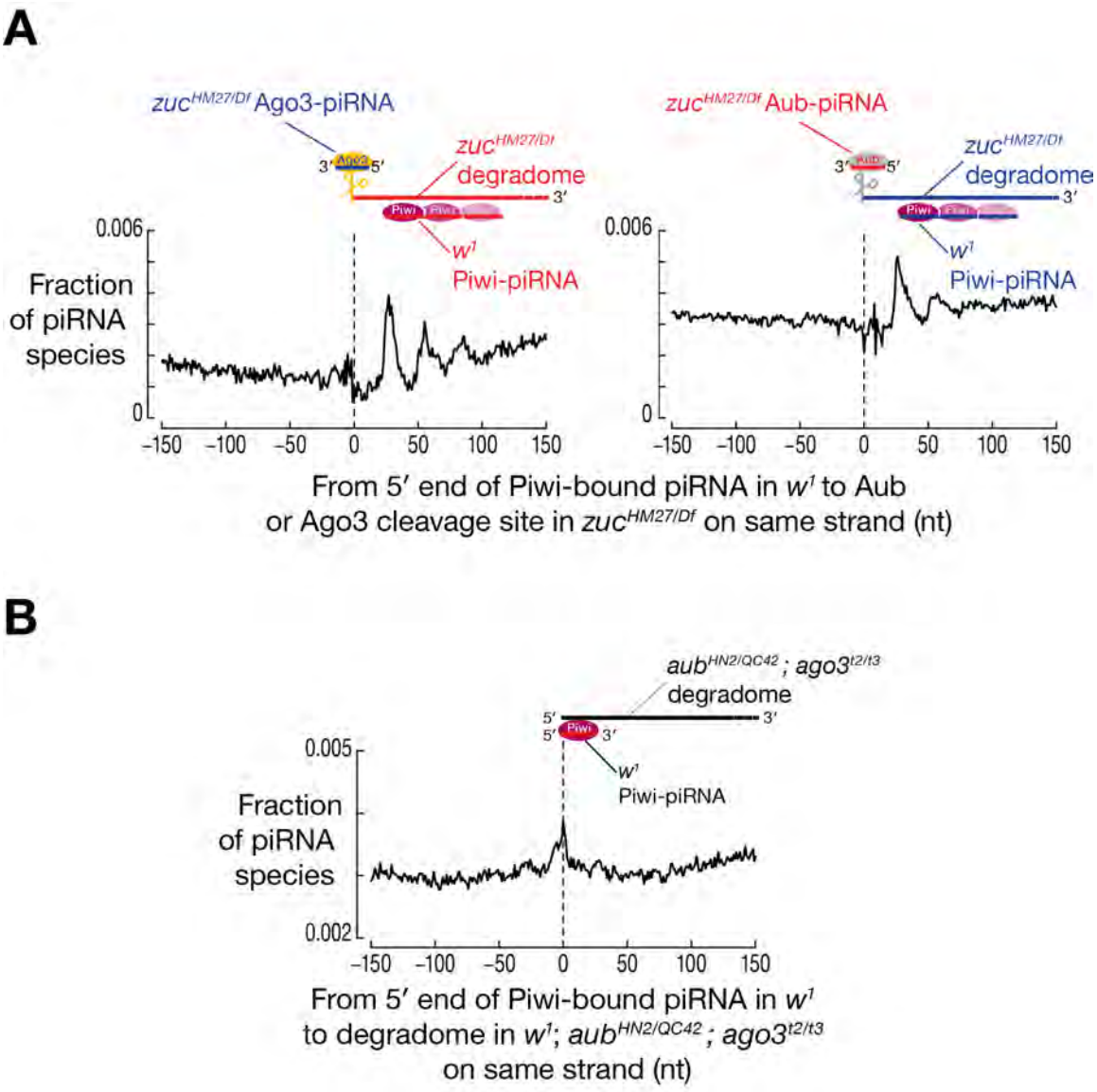

A
zuc^HM27/Df Ago3-piRNA
zuc^HM27/Df degradome
w^1 Piwi-piRNA
zuc^HM27/Df Aub-piRNA
zuc^HM27/Df degradome
w^1 Piwi-piRNA
Fraction of piRNA species
0.006
0
−150 −100 −50 0 50 100 150
From 5′ end of Piwi-bound piRNA in w^1 to Aub or Ago3 cleavage site in zuc^HM27/Df on same strand (nt)
B
aub^HN2/QC42 ; ago3^t2/t3 degradome
w^1 Piwi-piRNA
Fraction of piRNA species
0.005
0.002
−150 −100 −50 0 50 100 150
From 5′ end of Piwi-bound piRNA in w^1 to degradome in w^1; aub^HN2/QC42 ; ago3^t2/t3 on same strand (nt)

Figure Legend 3.11. Piwi-associated piRNAs Display Phasing 3′ to the Cleavage Sites of Aub and Ago3.

(A) Distance from the 5′ ends of Piwi-associated piRNAs in w^{1} to the cleavage sites of Ago3 (left) and Aub (right) identified in $zuc^{HM27/Df}$.

(B) Distance from 5′ ends of Piwi-associated piRNAs in w^{1} to the 5′ ends of degradome reads in w^{1}; $aub^{HN2/QC42}$; $ago3^{t2/t3}$.

Next, we measured the distance from the 5′ ends of Aub- and Ago3-bound piRNAs to the 5′ ends of Piwi-bound piRNAs on the same genomic strand. Again, the 5′ ends of Piwi-bound piRNAs were typically 26 nt downstream from the 5′ ends of Ago3-piRNAs and 27–29 nt downstream of the 5′ ends of Aub-piRNAs (Figure 3.12A). Similarly, the 5′ ends of Aub-bound piRNAs lay ~26 nt downstream from the 5′ ends of Ago3-bound piRNAs. In contrast, the 5′ ends of Ago3-bound piRNAs were no more likely to be ~26 nt downstream from the 5′ ends of Aub-bound piRNAs than would be expected by chance. Thus, RNAs cut by Ago3 produce phased, Aub-bound piRNAs, but RNAs cut by Aub do not make phased, Ago3-bound piRNAs.

The distance between the 5′ ends of Piwi-bound piRNAs and the 5′ ends of Ago3- or Aub-bound piRNAs on the opposite genomic strand (i.e., Ping-Pong analysis) again suggests that the 3′ cleavage product generated by Ago3 or Aub is initially processed into a secondary piRNA, and thereafter is used for the production of phased primary piRNAs loaded into Piwi (Figure 3.12B). Piwi does not directly participate in Ping-Pong, and the 5′ ends of Piwi-bound piRNAs did not map 10 nt from the 5′ ends of Aub- or Ago3-bound piRNAs. Instead, Piwi-bound piRNAs lay 15–19 nt after the 5′ ends of Aub- or Ago3-bound piRNAs. Such phased piRNAs have been detected previously, but were attributed to Ping-Pong amplification (Lau et al., 2009).

Figure 3.12

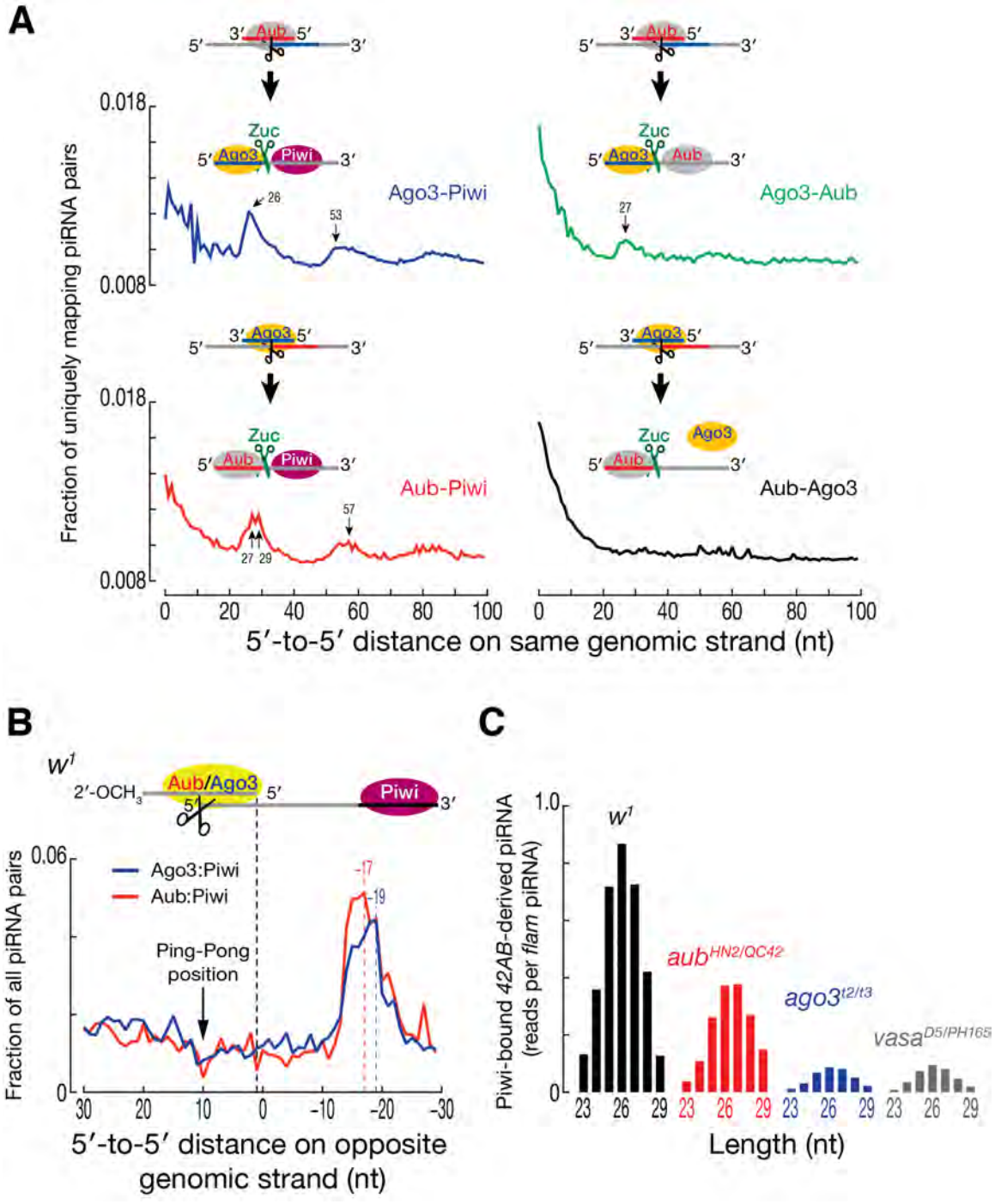

A
Fraction of uniquely mapping piRNA pairs
0.018
0.008
Ago3-Piwi
Ago3-Aub
Aub-Piwi
Aub-Ago3
Zuc
Ago3
Piwi
Aub
26
53
27
57
27 29
0 20 40 60 80 100
5′-to-5′ distance on same genomic strand (nt)
B
w1
2′-OCH3
Aub/Ago3
Piwi
Fraction of all piRNA pairs
0.06
0
Ago3:Piwi
Aub:Piwi
Ping-Pong position
-17
-19
30 20 10 0 -10 -20 -30
5′-to-5′ distance on opposite genomic strand (nt)
C
Piwi-bound 42AB-derived piRNA (reads per flam piRNA)
1.0
0
w1
aubHN2/QC42
ago3t2/t3
vasaD5/PH165
23 26 29
Length (nt)

Figure Legend 3.12. Majority of Piwi-piRNAs are Generated from the Cleavage Products of Ago3

(A) Distance from the 5′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs on the same genomic strand for piRNAs bound to each PIWI protein in w^1 ovaries.

(B) Distance from the 5′ ends of Aub- or Ago3-bound piRNAs to the 5′ ends of Piwi-bound piRNAs on the opposite genomic strand in w^1 ovaries.

Contributions of Aub and Ago3 to Phased Primary piRNAs

In the absence of Ago3 or Vasa, *42AB*-derived, Piwi-bound piRNAs decreased to ~10% of the w^1 level (Figure 3.12C). Loss of Aub had a more modest effect: *42AB*-derived, Piwi-bound piRNAs were ~47% of the w^1 level. Thus, Ago3 initiates the production of most phased, Piwi-bound primary piRNAs. These data help explain why transposon silencing requires heterotypic Aub:Ago3 Ping-Pong amplification (Zhang et al., 2011): Homotypic Aub:Aub Ping-Pong cannot generate enough Piwi-bound, antisense, primary piRNAs.

Nibbler Trims the 3′ end of piRNAs After Zuc Cleavage

Experiments in silkmoth cells and mice implicate the Tudor protein Papi in 3′ piRNA trimming (Honda et al., 2013). To examine the role of 3′ trimming in the biogenesis of phased primary piRNAs, we sequenced small RNAs from *papi* mutant fly ovaries (Figure 3.13A). The median length of piRNAs from nearly all transposon families increased 0.35 nt (p-value < 2.2×10^{-16}, Wilcoxon signed-rank test; Figure 3.13B and 3.13C) and germline and somatic piRNA phasing became more pronounced (Figure 3.13D). We propose that 3′ trimming of Piwi-bound piRNAs allows the use of uridines >26 nt after the 5′ end of a pre-piRNA as cleavage sites to make piRNAs.

Figure 3.13

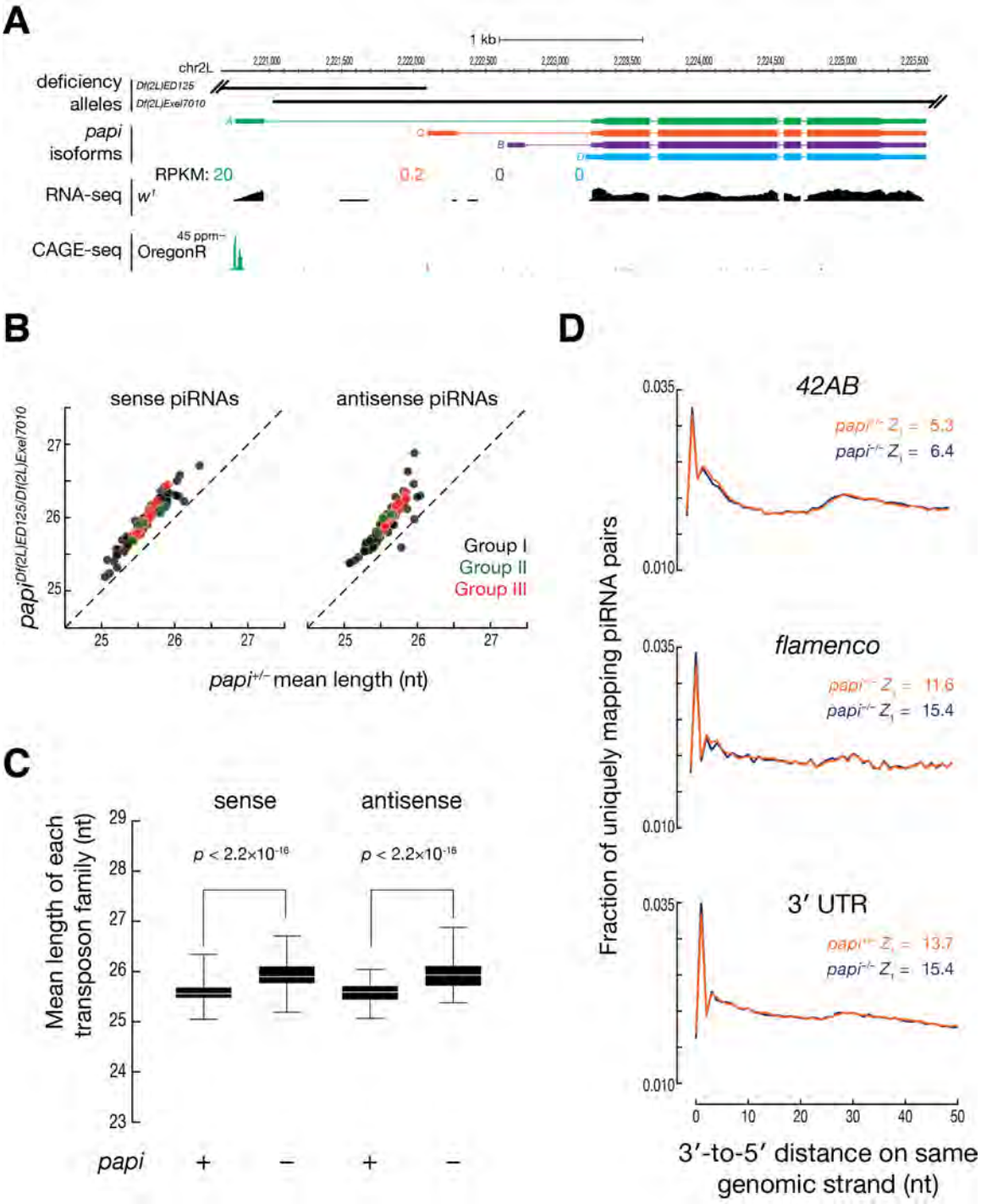

A
1 kb
chr2L
deficiency alleles
Df(2L)ED125
Df(2L)Exel7010
papi isoforms
RPKM: 20
0.2
0
0
RNA-seq
w1
CAGE-seq
45 ppm
OregonR
B
sense piRNAs
antisense piRNAs
papiDf(2L)ED125/Df(2L)Exel7010
papi+/− mean length (nt)
Group I
Group II
Group III
C
sense
antisense
p < 2.2×10−16
p < 2.2×10−16
Mean length of each transposon family (nt)
papi
D
42AB
papi+/− Z1 = 5.3
papi−/− Z1 = 6.4
flamenco
papi+/− Z1 = 11.6
papi−/− Z1 = 15.4
3′ UTR
papi+/− Z1 = 13.7
papi−/− Z1 = 15.4
Fraction of uniquely mapping piRNA pairs
3′-to-5′ distance on same genomic strand (nt)

Figure Legend 3.13. Papi and 3′ Trimming in piRNA Biogenesis

(A) Gene model for fly *papi* with RPKM values shown for each mRNA isoform calculated using RNA-seq data from w^1 ovaries. CAGE-seq data from Oregon R ovaries is also shown.

(B-C) Scatterplots (B) and boxplots (C) compare the mean lengths of piRNAs from each transposon family. *p*-values were calculated using a paired Wilcoxon test.

(D) Distance from 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs for piRNAs derived from *42AB*, *flamenco*, and 3′ UTRs.

Other than its role in miRNA 3′ end trimming, exonuclease Nibbler has been proposed to modulate the length of piRNA (Feltzin et al., 2015). We thus examined the mean length and phasing of piRNAs in *Nibbler* heterozygotes, mutant, as well as *Nibbler* mutant flies expressing transgenic, wild-type or catalytically inactive Nibbler (Figure 2.7A). Consistent with the previous finding, the median length of piRNAs increased from 25.5 nt in w^{1118} to 25.9 nt in $nibbler^{+/-}$, 26.0 nt in $nibbler^{-/-}$, and was reduced by transgenic $Nibbler^{WT}$ (25.6 nt) but not by $Nibbler^{E435,E437}$ (25.9 nt). Consistently, phasing by 3′-to-5′ analysis increased in the absence of functional Nibbler (Figure 3.14B; w^{1118} Z_1 = 3.9; $nibbler^{+/-}$ Z_1 = 5.2; $nibbler^{-/-}$ Z_1 = 8.3; $Nibbler^{WT}$ Z_1 = 4.7; $Nibbler^{E435,E437}$ Z_1 = 7.8). Our data suggests that Nibbler trimming follows the cleavage of Zuc machinery, possibly to protect piRNA from tailing enzyme (Han et al., 2015a).

Figure 3.14

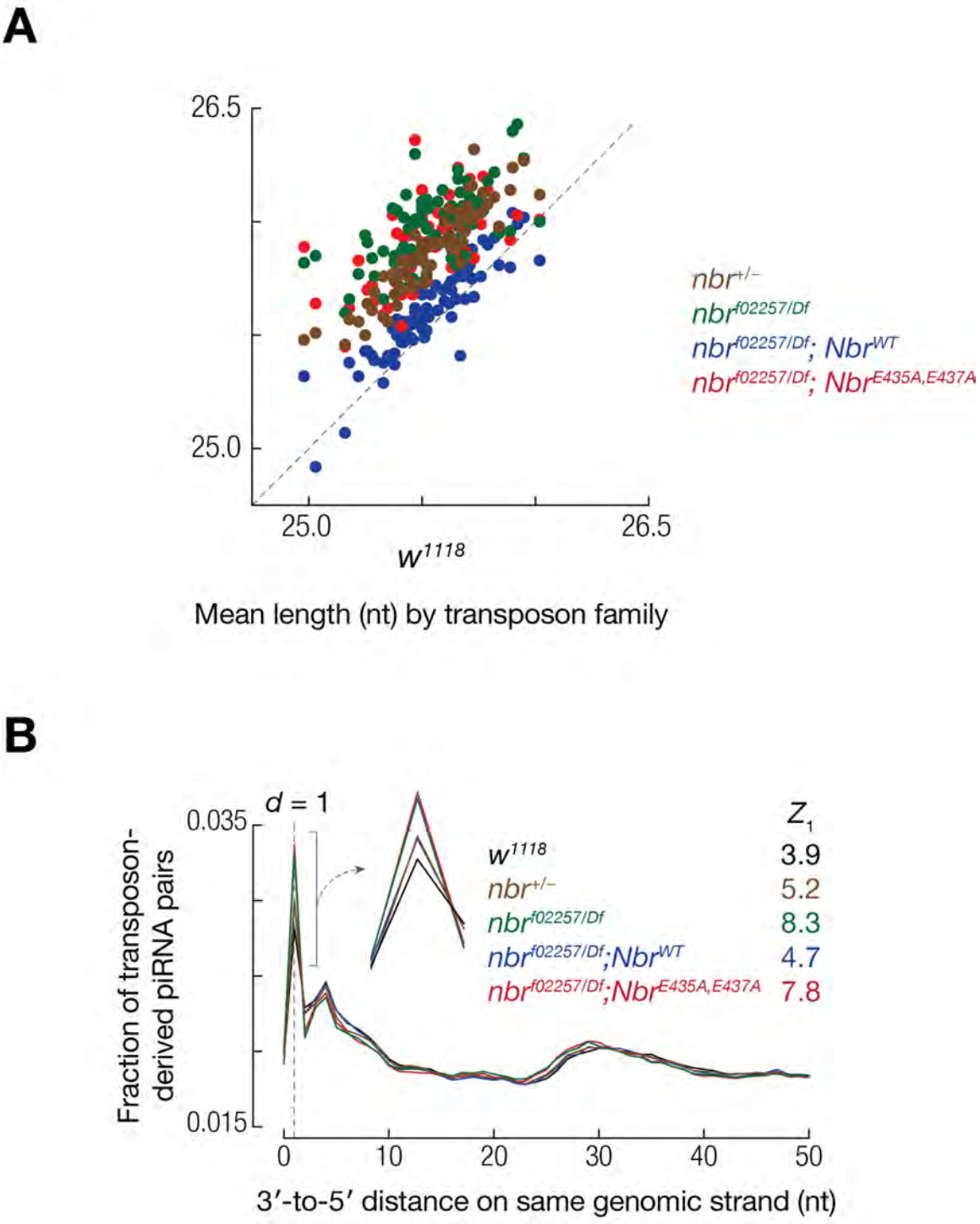

A
26.5
25.0
25.0
26.5
w1118
nbr+/−
nbrf02257/Df
nbrf02257/Df; NbrWT
nbrf02257/Df; NbrE435A,E437A
Mean length (nt) by transposon family
B
d = 1
0.035
0.015
Fraction of transposon-derived piRNA pairs
Z1
w1118 3.9
nbr+/− 5.2
nbrf02257/Df 8.3
nbrf02257/Df;NbrWT 4.7
nbrf02257/Df;NbrE435A,E437A 7.8
0
10
20
30
40
50
3′-to-5′ distance on same genomic strand (nt)

Figure Legend 3.14. Nibbler Trims piRNAs after Zuc Cleavage

(A) Scatterplots compare the mean lengths of piRNAs from each transposon family.

(B) Distance from 3′ ends of upstream piRNAs to the 5′ ends of downstream piRNAs for piRNAs derived from transposon sequences.

Phasing of mammalian piRNAs

In testes from wild-type mice, one piRNA 5′ end often lies 30–40 nt downstream from another (data not shown), possibly because mouse pre-piRNAs are longer than those in flies and require the 3′ trimming of ~3–10 nt. Analysis of Papi (*Tdrkh*$^{-/-}$) mutant testes supports this view. *Tdrkh*$^{-/-}$ testes accumulate 31–37 nt RNAs instead of 26–30 nt piRNAs, and most of these longer species share their 5′ ends with mature piRNAs from *Tdrkh*$^{+/-}$ heterozygotes (Saxe et al., 2013). At 11 days post partum (dpp), 3′-to-5′ distance analysis of piRNAs from *Tdrkh*$^{-/-}$ testes showed clear evidence for phasing (Figure 3.15A). Mouse piRNAs typically begin with uridine, and the 3′ ends of the longer RNAs in *Tdrkh*$^{-/-}$ testes were generally followed by a uridine in genomic sequence (Figure 3.15B). piRNA 5′-to-5′ distance analysis of *Tdrkh*$^{+/-}$ and *Tdrkh*$^{-/-}$ showed broad peaks at 35 to 43 nt—the same length as the pre-piRNAs detected in *Tdrkh*$^{-/-}$. We conclude mammalian primary piRNAs are phased, but are more extensively trimmed than those in flies.

Figure 3.15

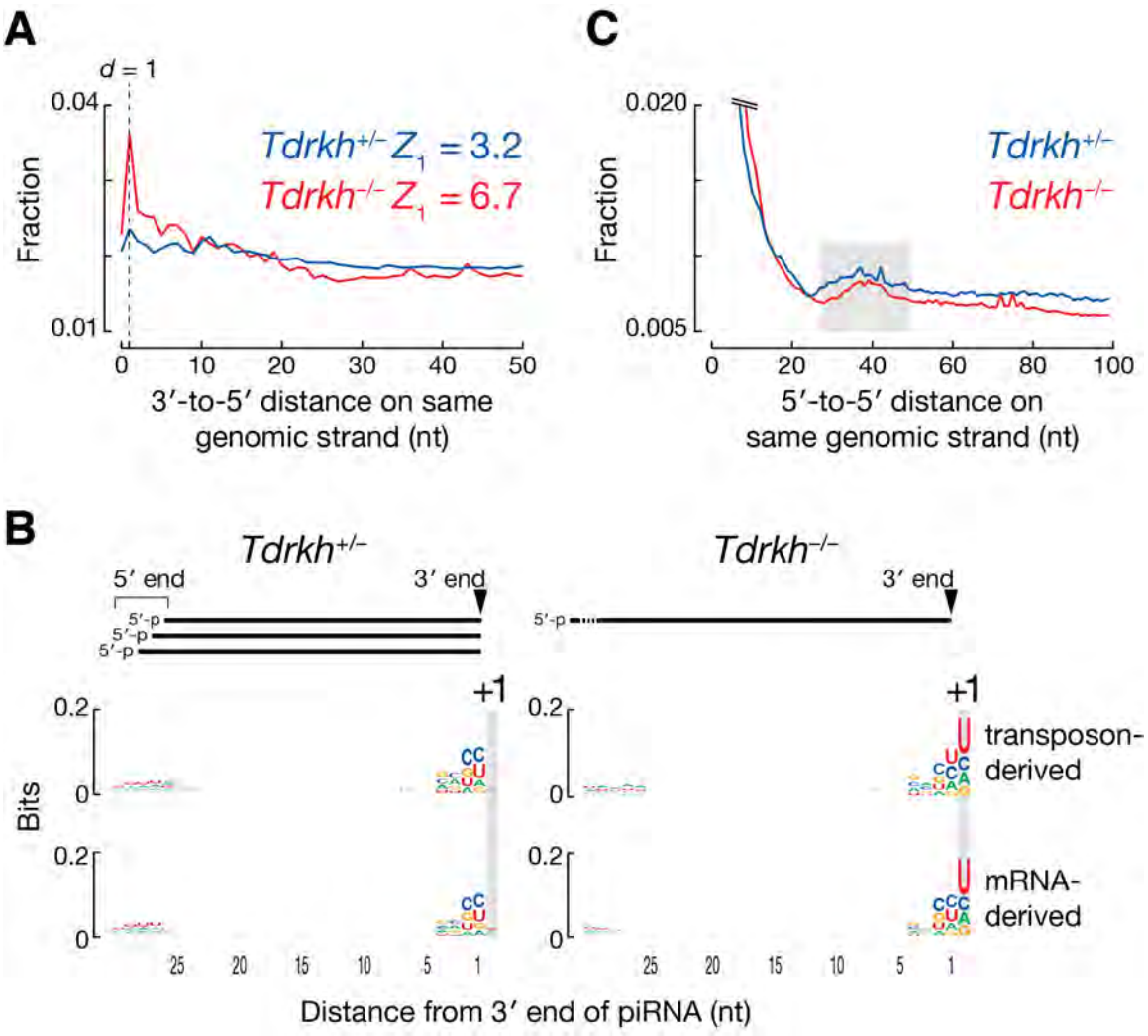

A
d = 1
0.04
0.01
Fraction
Tdrkh+/– Z1 = 3.2
Tdrkh–/– Z1 = 6.7
0
10
20
30
40
50
3′-to-5′ distance on same genomic strand (nt)
C
0.020
0.005
Fraction
Tdrkh+/–
Tdrkh–/–
0
20
40
60
80
100
5′-to-5′ distance on same genomic strand (nt)
B
Tdrkh+/–
Tdrkh–/–
5′ end
3′ end
5′-p
+1
Bits
0.2
0
transposon-derived
mRNA-derived
25
20
15
10
5
1
Distance from 3′ end of piRNA (nt)

Figure Legend 3.15. Mouse piRNAs display phasing

(A) Distance from 3′ ends of upstream piRNAs to 5′ ends of downstream piRNAs on the same genomic strand for uniquely mapping piRNAs in *Tdrkh*$^{+/-}$ and *Tdrkh*$^{-/-}$ in mouse testes at 11 dpp.

(B) The nucleotide composition, in species, of sequences 29 nt upstream and 1 nt downstream of the 3′ ends of uniquely mapping piRNAs. Pachytene piRNAs are not included because spermatogenesis arrests before the pachytene stage in *Tdrkh*$^{-/-}$.

(C) Distance from 5′ ends of upstream piRNAs to 5′ ends of downstream piRNAs on the same genomic strand for uniquely mapping piRNAs in *Tdrkh*$^{+/-}$ and *Tdrkh*$^{-/-}$ mouse testes at 11 dpp. Data are from Saxe *et al.* (GSE47151).

Discussion

Our findings suggest a substantially revised model for primary piRNA biogenesis (Figure 3.16). The model proposes that each cycle of Ping-Pong amplification can generate one secondary piRNA and multiple primary piRNAs. For example, a secondary piRNA bound to Ago3 can direct cleavage of a fully or partially complementary target RNA (Wang et al., 2014). The resulting 3′ cleavage product then binds Aub. An unknown factor, possibly Armi, recruits Zuc, which makes a second cut 26–29 nt away from the 5′ monophosphate, likely at the first uridine not occluded by Aub. The two cleavage products from this reaction follow decidedly different fates. The 5′ fragment matures into a secondary piRNA bound to Aub. We envision that some but not all of such Aub-bound RNA fragments will require 3′ trimming to achieve their characteristic length (Honda et al., 2013). The 3′ fragment becomes a substrate for the production of phased primary piRNAs by Zuc. With the aid of Armi, Zuc travels 5′-to-3′ cleaving every ~26 nt. The piRNAs released by this process load mainly into Piwi. Although as much as 90% of Piwi-associated piRNAs are generated by this mechanism in the germline, piRNAs in the soma, which lacks Aub and Ago3, must deploy a different mechanism to initiate Zuc-dependent processing.

Our data also help explain why effective transposon silencing requires heterotypic Aub:Ago3 Ping-Pong amplification. In *ago3* mutant ovaries, homotypic Aub:Aub Ping-Pong replaces heterotypic Ping-Pong. Although antisense piRNAs are produced by homotypic Ping-Pong, they fail to silence

transposon expression (Li et al., 2009; Zhang et al., 2011). We propose that homotypic Aub:Aub Ping-Pong is unable to replace heterotypic Ping-Pong, because it cannot generate enough Piwi-bound primary piRNAs.

Figure 3.16

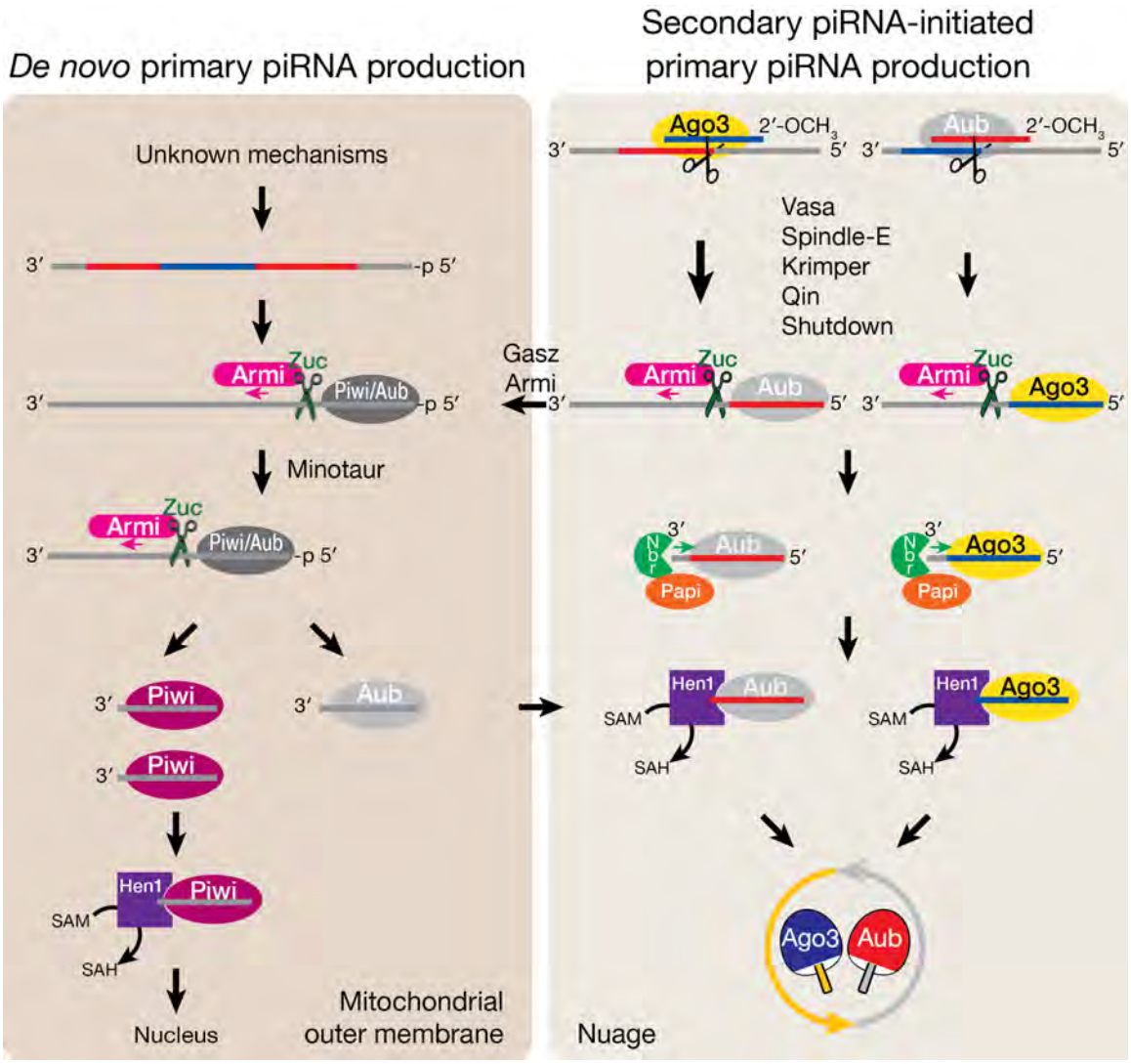

De novo primary piRNA production
Secondary piRNA-initiated primary piRNA production
Unknown mechanisms
3′
-p 5′
Armi
Zuc
Piwi/Aub
Minotaur
Piwi
Aub
Hen1
SAM
SAH
Nucleus
Mitochondrial outer membrane
Gasz
Armi
Ago3
2′-OCH3
5′
Vasa
Spindle-E
Krimper
Qin
Shutdown
Nbr
Papi
Nuage

Figure Legend 3.16. Revised Model of piRNA Biogenesis in *Drosophila*

(Left) The de novo primary piRNA pathway starts with piRNA intermediates released from piRNA cluster transcripts in an Aub- and Ago3-independent manner. Zuc slices them consecutively every ~26 nt, aided by Armi and other factors in the primary pathway (e.g., Minotaur and Gasz). Those primary piRNAs are loaded into Piwi and Aub, but not Ago3.

(Right) In nuage, cleavage by Ago3 or Aub produces piRNA intermediates with a 5′ monophosphate. The 3′ cleavage products are loaded into Aub and Ago3, followed by Zuc-dependent cleavage ~26 nt from their 5′ ends. This cleavage produces the 3′ ends of the “Ping-Pong partner” secondary piRNA and the 5′ ends of long RNAs that become substrates for Zuc, which processively cleaves the RNA to generate phased piRNAs loaded into Piwi and, to a lesser extent, Aub. We propose that the Zuc machinery chooses as its cleavage site the first uridine that is not protected by a PIWI protein. Consequently, some pre-piRNAs require Papi- and Nbr-dependent 3′ trimming before their 3′ ends are methylated by Hen1.

Finally, the Ping-Pong model does not explain the stunning diversity of piRNA sequences: each cycle of Ping-Pong increases the abundance of a pair of piRNAs, but cannot generate piRNAs with novel sequence specificity (piRNA nucleotides 2–16; Wang et al., 2014). Our data show that each RNA cleaved by Aub or Ago3 not only produces a secondary, Ping-Pong piRNA partner, but also produces primary piRNAs from the sequences immediately 3′ to the secondary piRNA. Such a spreading mechanism calls to mind features of siRNA production in *Caenorhabditis elegans* and *Arabidopsis thaliana* (Xie et al., 2005; Yoshikawa et al., 2005; Bagijn et al., 2012; Lee et al., 2012), and primed CRISPR adaptation in *Escherichia coli* (Swarts et al., 2012; Datsenko et al., 2012; Heler et al., 2014). Although the detailed mechanisms differ (e.g., slicing activity is dispensable in *C. elegans,* and an RNA-dependent RNA polymerase is required in *A. thaliana*), signal amplification and sequence diversification is clearly a recurrent theme for RNA-guided silencing in animals, plants, and bacteria.

Experimental Procedures

General methods

Stocks and crosses were grown at 25°C. All flies were in the w^{1} background, except w^{+}; *Df(2L)Prl* (zuc^{Df}) and the *papi* strains w^{1118} ; *Df(2L)ED125* and w^{1118} ; *Df(2L)Exel7010*. Ovaries were dissected in modified Robb's Buffer (55 mM CH_3COONa, 40 mM CH_3COOK, 100 mM sucrose, 10 mM glucose, 1.2 mM $MgCl_2$, 1 mM $CaCl_2$, 100 mM HEPES, pH 7.4). RNAs were purified using mirVana (Ambion, Life technologies, CA, USA).

Small RNA library construction

Total RNA (100 µg) or RNA co-immunoprecipitated with Aub, Ago3, or Piwi was purified by 15% urea polyacrylamide gel electrophoresis (PAGE), selecting for 18–30 nt long RNAs. Oxidization of RNA with $NaIO_4$ was used to deplete miRNAs and enrich for siRNAs and piRNAs (Li et al., 2009). Ligation of the 3′ adaptor (5′-rApp NNN TGG AAT TCT CGG GTG CCA AGG /ddC/-3′ or 5′-rApp TGG AAT TCT CGG GTG CCA AGG /ddC/-3′) using truncated, K227Q mutant T4 RNA Ligase 2 at 25°C for ≥16 h and subsequent size selection by 15% PAGE was as described (Li et al., 2009). To exclude 2S rRNA from sequencing libraries, 10 pmol 2S blocker oligo was added before 5′ adaptor ligation (Wickersheim and Blumenstiel, 2013); 5′ adaptor was added using T4 RNA ligase (Ambion) at 25°C for ≥ 2 h, followed by reverse-transcription using AMV reverse transcriptase (New England Biolabs, MA, USA) and PCR using Q5

polymerase (NEB). An Illumina HiSeq 2000 was used for high-throughput, single-end 50 nt or 100 nt sequencing.

Degradome-seq library construction

Freshly isolated RNA (4 µg) was subjected to two rounds of rRNA depletion (Ribo-Zero; Epicentre, WI, USA), treated with turbo DNase (Ambion), and then size-selected to isolate RNA ≥ 200 nt (DNA Clean & Concentrator™-5, ZYMO RESEARCH, CA, USA). T4 RNA ligase (Ambion) was used at 25°C for 2–4 hours for 5′ ligation. Reverse transcription with SuperScript III (Life Technologies) employed a primer containing a degenerate sequence at its 3′ end (5′-GCA CCC GAG AAT TCC ANN NNN NNN-3′). cDNA was amplified by PCR using Q5 polymerase (NEB), and 200–400 nt dsDNA was isolated using 6% native PAGE. An Illumina HiSeq 2000 was used to perform paired-end, 100 nt sequencing of the dsDNA products.

Small RNA immunoprecipitation

Anti-Piwi, Aub, and Ago3 antibodies (~10 µg) were incubated with Protein A and G Dynabeads (15 µl each; Life Technologies) in lysis buffer (30 mM HEPES-KOH, pH 7.4, 100 mM CH_3COOK, 2 mM $(CH_3COO)_2Mg$, 5 mM dithiothreitol, 0.5% [v/v] NP-40, 1 mM 4-(2-Aminoethyl)benzenesulfonyl fluoride hydrochloride, 0.3 µM Aprotinin, 40 µM Bestatin, 10 µM E-64, 10 µM Leupeptin) at 4°C for 4 h with rotation, then washed twice with lysis buffer. Next, 400–800 µl freshly prepared ovary lysate (5 µg/µl) was added and incubated at 4°C for 4 h with

rotation. After washing the beads four times with ice-cold lysis buffer, RNA was purified using Trizol (Life Technologies).

General bioinformatics analyses

Analyses were performed using piPipes v1.4 (Han et al., 2015b). Briefly, all small RNA sequencing libraries were filtered using PHRED score ≥ 5. Genome mapping using Bowtie v1.1.0 allowed no mismatches for fly and one mismatch for mouse data. Degradome mapping was performed with Bowtie2 v2.2.3 (to rRNA) and STAR v2.3.0 (to genome). Reads whose 5′ ends could not be determined precisely (soft-clipped) during alignment were removed computationally. Alignments were categorized by genomic feature using BEDTools v2.17.0. For transgene mapping, we first aligned an oxidized small RNA-seq library from w^1 (23,712,713 genome-mapping reads) to the transgene sequence, masking (turning into Ns) positions that could be mapped to piRNAs more abundant than 1 part per million. Statistical analysis in R 3.0.2 required *p*-value < 0.005. To compare piRNA abundance between two small RNA libraries, we normalized to non-transposon-derived siRNAs, rather than uniquely mapping reads of the genome, in order to avoid biasing genotypes such as *zuc*, in which piRNA abundance was decreased globally. To compare the abundance of piRNAs associated with Piwi, we normalized to *flamenco*-derived reads, which are unaffected by defects in the germline piRNA pathway.

Phasing analysis

Reads were mapped to genome, alignments that overlapped with rRNAs, tRNAs and snoRNAs were removed, and the remaining 23–29 nt RNAs (fly piRNAs) or 23–35 nt (mouse piRNAs) were analyzed. To analyze small RNAs in *Tdrkh*$^{-/-}$ *and* *Tdrkh*$^{+/-}$, all reads ≥ 23 nt were used. The score for a distance of *x* nt was calculated by ∑minimal (M_i, N_{i+x}) where M_i is the number of reads whose 3′ ends are located at position *i* and N_{i+x} is the number of reads whose 5′ ends are located at position *i*+*x*. When *x* equals 0, the 3′ and 5′ ends overlap. When *x* equals to 1, the 5′ end is immediately downstream of the 3′ end (phasing). For analyses including multi-mappers, reads were apportioned by the number of times they can be aligned to the genome. To calculate Z_1, overlaps at position 2–20 nt were used as background to calculate *Z* scores. In Ping-Pong analyses, the product, instead of the smaller value, of M and N was used.

We used three different computational strategies to evaluate the phasing of piRNAs: 5′-to-5′ end distance, 3′-to-5′ end distance, and +1U percentage. A peak in 5′-to-5′ distance analysis demonstrates that the generation of 5′ ends of piRNAs occur with a certain periodicity. Imprecision in the production of piRNA 5′ ends is expected to produce a broad 5′-to-5′ peak, impeding statistical analysis. For 3′ to 5′ analysis, a peak at a distance of 1 nt suggests that the same cleavage event generates both the 3′ end of an upstream piRNA and the 5′ end of a downstream piRNA. It also indicates the absence of 3′ end trimming. Because piRNAs are generally subjected to Papi-dependent 3′ trimming, this

analysis can fail to detect the periodicity. The +1U percentage reflects the percentage of uridines at the nucleotide immediately after the 3′ end of piRNAs. This is an indirect measurement that relies on the finding that primary piRNAs typically begin with uridine. Like the other measures, +1U analysis can be confounded by 3′ piRNA trimming. However, the +1U percentage is unaffected by sequencing depth, allowing comparison of dataset with widely varying numbers of mappable reads or species.

Assigning immunopurified small RNA reads to Piwi, Aub, or Ago3

We used a χ^2 test with a *p*-value cutoff < 0.005 to test whether a sequence was enriched in one of the three PIWI proteins. A sequence could be unambiguously assigned only when one of two conditions was met: (1) the sequence was uniquely sequenced in only one of the three libraries (two for mutants lacking one PIWI protein) or (2) the sequence passed the χ^2 test ($p < 0.005$) and was at least five-fold more abundant in one sample than the other two.

Acknowledgments

We thank Alicia Boucher, Cindy Tipping, Gwen Farley and Ellen Kittler for technical assistance; Ryuya Fukunaga and Erik Sontheimer for insightful discussions; Julius Brennecke for reagents, helpful discussions and sharing unpublished data; and members of the Weng and the Zamore laboratories for advice and critical comments on the manuscript. This work was supported in part by National Institutes of Health grants HG007000 to Z.W. and by GM62862 and GM65236 to P.D.Z. Sequencing data are available from the NCBI Sequence Read Archive using accession number SRP045930.

Chapter IV piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing.

Disclaimer

This chapter was a product of a collaborative effort among the authors: Bo W Han (BWH), Wei Wang (WW), Phillip D. Zamore (PDZ), and Zhiping Weng (ZW). BWH implemented the small RNA-seq, degradome- and CAGE-seq, ChIP-seq, and genomic DNA sequencing pipeline. BWH and WW implemented the RNA-seq pipeline. PDZ and ZW supervised the project.

Summary

piRNAs, 23–36 nucleotide (nt) small silencing RNAs, repress transposon expression in the metazoan germline, thereby protect the genome. Although high-throughput sequencing has made it possible to examine the genome and transcriptome at unprecedented resolution, extracting useful information from gigabytes of sequencing data still requires substantial computational skills. Additionally, researchers may analyze and interpret the same data differently, generating results that are difficult to reconcile. To address these issues, we developed a coordinated set of pipelines, “piPipes,” to analyze piRNA and transposon-derived RNAs from a variety of high-throughput sequencing libraries, including small RNA, RNA, degradome or 7-methyl guanosine-cap analysis of gene expression (CAGE), chromatin immunoprecipitation (ChIP), and genomic DNA sequencing. piPipes can also produce figures and tables suitable for publication. By facilitating data analysis, piPipes provides an opportunity to standardize computational methods in the piRNA field.

Introduction

piRNAs, a class of 23–36 nt long small silencing RNAs, suppress transposon expression in the metazoan germline and, in some animals, the adjacent gonadal somatic cells (Luteijn and Ketting, 2013). By preventing transposition, the piRNA pathway ensures that genetic information passes faithfully to the next generation. Disruption of the piRNA pathway typically leads to transposon mobilization, double-stranded DNA breaks, and sterility.

High-throughput sequencing technologies have been widely deployed in the study of piRNAs. Small RNA-seq reveals the identity and abundance of piRNAs (Brennecke et al., 2007); RNA-seq detects and quantifies mRNA and transposon transcripts (Reuter et al., 2011); degradome-seq (also termed RACE-seq) detects the cleavage products of PIWI-proteins guided by piRNAs (Reuter et al., 2011); ChIP-seq detects chromatin modifications directed by piRNAs or transcription factor-binding events that regulate piRNA precursor or target transcription (Sienski et al., 2012; Li et al., 2013); and genomic DNA sequencing detects new transposition events caused by transposons that escape piRNA repression (Khurana et al., 2011; Sienski et al., 2012). Correctly extracting biological knowledge from such voluminous data requires significant computational expertise and effort. Moreover, different laboratories use diverse methods to analyze and interpret data (e.g., the way of treating reads that map to multiple locations in a reference genome; Huang et al., 2013; Marinov et al., 2015; Lin et al., 2015). To provide a standardized set of tools to analyze these

diverse data types, we developed piPipes, a collection of five integrated pipelines for small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genome-seq analyses.

Methods

piPipes comprises five pipelines designed to analyze small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq or genome-seq data. The small RNA-seq pipeline reports the abundance, length distribution, nucleotide composition and 5′-to-5′ distance (Ping-Pong signature) of piRNAs assigned to genomic annotations, including individual transposon families and piRNA clusters, the initial sources of piRNA precursor transcripts. The RNA-seq pipeline reports the normalized abundance of transcripts from both genes and transposons in RPKM (Reads Per Kilobase of transcript per Million mapped reads). The degradome-seq pipeline offers methods to identify piRNA-directed cleavage products. This pipeline can also be used to analyze any long RNA sequencing method designed to define RNA 5′ ends, e.g., CAGE-seq. The ChIP-seq pipeline employs the widely used peak-calling algorithm MACS2 (Zhang et al., 2008), focusing on piRNA clusters and transposons. The genome-seq pipeline detects novel transposition events.

Besides the analysis pipelines, piPipes provides an installation pipeline to acquire genomic sequences and annotations for different genome assemblies. To achieve a generic interface across multiple genomes, piPipes uses the reference packages from the Illumina iGenome project. Additionally, piPipes comes with organized annotation files, including piRNA cluster annotations for *Drosophila melanogaster* and *Mus musculus*, two well-studied model organisms in the piRNA field (Brennecke et al., 2007; Li et al., 2013). Detailed instructions

for constructing the annotation files for other genomes can be found on the GitHub wiki.

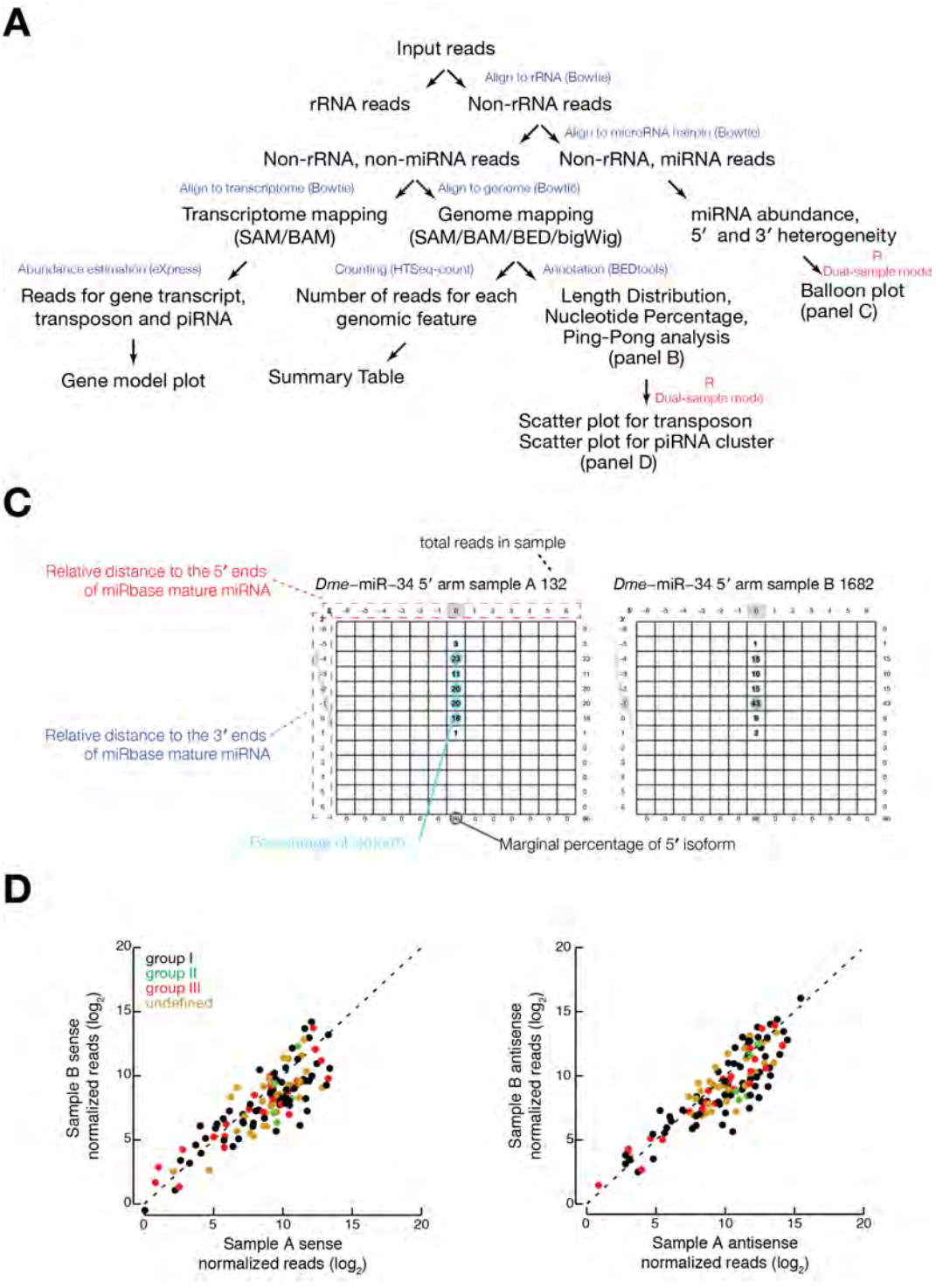
All five analysis pipelines map reads to the reference genomes, assigning mapped reads to annotated genomic features and quantifying the signal strength for each feature (computed from the number of mapped reads). piPipes uses Bowtie (Langmead et al., 2009) to map small RNA reads, Bowtie2 (Langmead and Salzberg, 2012), BWA (Li and Durbin, 2009), and mrFast (Alkan et al., 2009) to map genomic DNA reads, and STAR (Dobin et al., 2013) to map long RNA reads to a reference genome. piPipes uses two different methods to assign reads to different genomic features, including exon, intron, transposon, and piRNA clusters. The first maps all reads to the reference genome and then uses BEDTools (Quinlan and Hall, 2010) and HTSeq-count (Anders et al., 2015) to assign features to the reads based on their genomic coordinates. The second method directly aligns reads to the sequences of the features (e.g., the entire transcriptome comprising the sequences of coding and non-coding RNAs, transposon consensus sequences, and piRNA cluster sequences). eXpress (Roberts and Pachter, 2013) then quantifies reads using an expectation-maximization (EM) algorithm to assign reads that match multiple features. For each pipeline, piPipes produces summary tables of the statistics for each annotated feature, bigWig files for visualization in the UCSC (Kent et al., 2002) or IGV genome browser (Robinson et al., 2011), and publication-quality figures for presenting analysis results.

small RNA pipeline

piPipes requires that adaptors and barcodes be removed before running the pipeline. Because Bowtie (Langmead et al., 2009) does not incorporate sequence quality scores, piPipes gathers reads with the same sequence and aligns that sequence (species) once. Best practice requires filtering reads according to their PHRED score and discarding low quality reads. piPipes removes small RNA reads aligning to rRNAs (Figure 4.1A). After rRNA removal, piPipes aligns the rest of the reads to microRNA hairpins (Griffiths-Jones et al., 2006; Griffiths-Jones et al., 2008; Griffiths-Jones, 2010; Kozomara and Griffiths-Jones, 2011), and then calculates the 5′ and 3′ heterogeneity of the miRNA-mapping reads (Seitz et al., 2008). Next, piPipes uses Bowtie to align the non-rRNA, non-miRNA reads to the genome. The SAM/BAM output is converted to a modified BED format (“BED2” in piPipes) to reduce file size and computational load. BED2 replaces column four of standard BED format with the number of times a species appears in the library, replaces column five with the number of loci to which this sequence can be assigned, and replaces column seven with the sequence itself. With this design, uniquely mapping species can be retrieved by restricting column five to 1. Species mapping to multiple locations are counted by dividing column four by column five. Species calculation can be simply done by counting the unique appearance of sequences in column seven. piPipes applies BEDTools (Quinlan and Hall, 2010) to the BED2 file to assign reads to different genomic features (e.g., piRNA cluster, transposon family, genes, exon or intron).

For each genomic feature, piPipes classifies small RNA reads as siRNA or piRNA, according to length restrictions defined by the user. piPipes uses ggplot2 (Hadley, 2009) to plot length distribution (unique reads), nucleotide percentage (unique species) and Ping-Pong signature (with the abundance of reads divided by the number of loci to which the sequences can be assigned; Figure 4.1B). piPipes also aligns non-rRNA, non-miRNA reads to a reference index comprising gene transcripts, transposon consensus sequences and piRNA clusters, and then uses eXpress (Roberts and Pachter, 2013) to quantify number of reads assigned to different genomic feature. In the dual library mode, piPipes provides six different normalization methods to compare miRNA and piRNA between two samples. piPipes uses balloon plot (Warnes et al., 2008) to compare the relative abundance of different miRNA isoforms (Figure 4.1C) and scatter plots of reads from different transposon families or piRNA clusters to compare piRNA abundance (Figure 4.1D).

Figure 4.1


A
Input reads
Align to rRNA (Bowtie)
rRNA reads
Non-rRNA reads
Align to microRNA hairpin (Bowtie)
Non-rRNA, non-miRNA reads
Non-rRNA, miRNA reads
Align to transcriptome (Bowtie)
Align to genome (Bowtie)
Transcriptome mapping (SAM/BAM)
Genome mapping (SAM/BAM/BED/bigWig)
miRNA abundance, 5′ and 3′ heterogeneity
Abundance estimation (eXpress)
Counting (HTSeq-count)
Annotation (BEDtools)
R
Dual-sample mode
Reads for gene transcript, transposon and piRNA
Number of reads for each genomic feature
Length Distribution, Nucleotide Percentage, Ping-Pong analysis (panel B)
Balloon plot (panel C)
Gene model plot
Summary Table
R
Dual-sample mode
Scatter plot for transposon
Scatter plot for piRNA cluster
(panel D)
C
total reads in sample
Relative distance to the 5′ ends of miRbase mature miRNA
Dme–miR–34 5′ arm sample A 132
Dme–miR–34 5′ arm sample B 1682
Relative distance to the 3′ ends of miRbase mature miRNA
Marginal percentage of 5′ isoform
D
group I
group II
group III
undefined
Sample B sense normalized reads (log2)
Sample A sense normalized reads (log2)
Sample B antisense normalized reads (log2)
Sample A antisense normalized reads (log2)

B

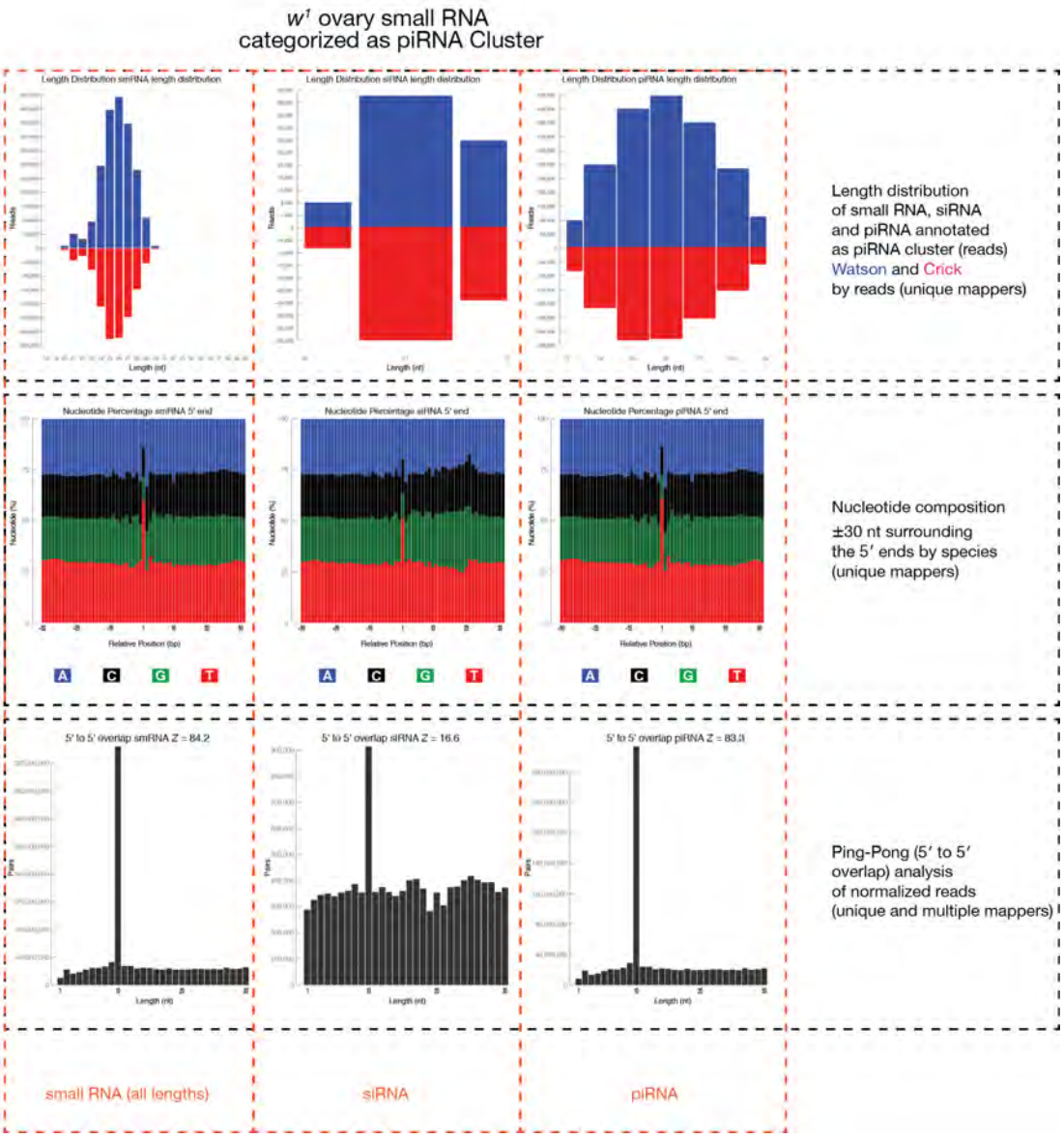

w1 ovary small RNA
categorized as piRNA Cluster
Length Distribution smRNA length distribution
Length Distribution siRNA length distribution
Length Distribution piRNA length distribution
Length distribution
of small RNA, siRNA
and piRNA annotated
as piRNA cluster (reads)
Watson and Crick
by reads (unique mappers)
Nucleotide Percentage smRNA 5' end
Nucleotide Percentage siRNA 5' end
Nucleotide Percentage piRNA 5' end
Relative Position (bp)
A C G T
Nucleotide composition
±30 nt surrounding
the 5′ ends by species
(unique mappers)
5' to 5' overlap smRNA Z = 84.2
5' to 5' overlap siRNA Z = 16.6
5' to 5' overlap piRNA Z = 83.8
Length (nt)
Ping-Pong (5′ to 5′
overlap) analysis
of normalized reads
(unique and multiple mappers)
small RNA (all lengths)
siRNA
piRNA

Figure Legend 4.1. Flowchart and Example Figures for the Small RNA Pipeline

(A) Work flow for the pipeline in single- (blue) and dual-library mode (red).

(B) An example of small RNA analysis for reads assigned to piRNA clusters in *Drosophila melanogaster*. Length distribution (first row), nucleotide composition thirty nucleotide upstream and downstream of the 5′ ends of the small RNA (second row), and local Ping-Pong signature (bottom row) for all small RNAs (left column), siRNAs (middle column), and piRNAs (right column). The length defining siRNAs versus piRNAs was set to the values defined by the users in the installing pipeline (20–22 nt for siRNA and 23–29 nt for piRNA here).

(C) Balloon plot for the pair-wise comparison of the 5′ and 3′ heterogeneity of microRNA ends. The X- (5′ ends) and Y-axes (3′ ends) report the distance to the ends of the miRBase annotated mature miRNA. The number in the “balloon” indicates the percentage of the isoform among all isoforms of the mature miRNA.

(D) A scatter plot comparing sense and antisense piRNAs abundance, classified by transposon family between two normalized data sets.

RNA-seq pipeline

Non-rRNA reads are aligned to the genome by STAR (Dobin et al., 2013). piPipes uses Cufflinks (Trapnell et al., 2010) to quantify gene expression. piPipes also uses HTSeq-count to count the uniquely mapping reads for each genomic annotation (Figure 4.2A). To quantify reads from transposons or piRNA clusters, piPipes directly aligns the non-rRNA reads to a transcriptome index that includes gene, transposon consensus and piRNA cluster sequences. Then the output is fed to eXpress (Roberts and Pachter, 2013). Because transposon-derived reads could increase in some mutants, skewing the sequencing depth calculation, piPipes uses the depth calculated by Cufflinks based on reference-compatible fragments (only reads that can be assigned to protein-coding genes and well-characterized ncRNAs but not transposons, tRNA, et al.). In dual-library mode, the output of eXpress is used to draw a scatter plot that includes coding (NM*) and non-coding (NR*) transcripts, as wells as transcripts from transposons and piRNA clusters (Figure 4.2B). For the analysis of differentially expressed genes, piPipes employs Cuffdiff (Trapnell et al., 2013) and cummeRbund (http://compbio.mit.edu/cummeRbund/).

Figure 4.2

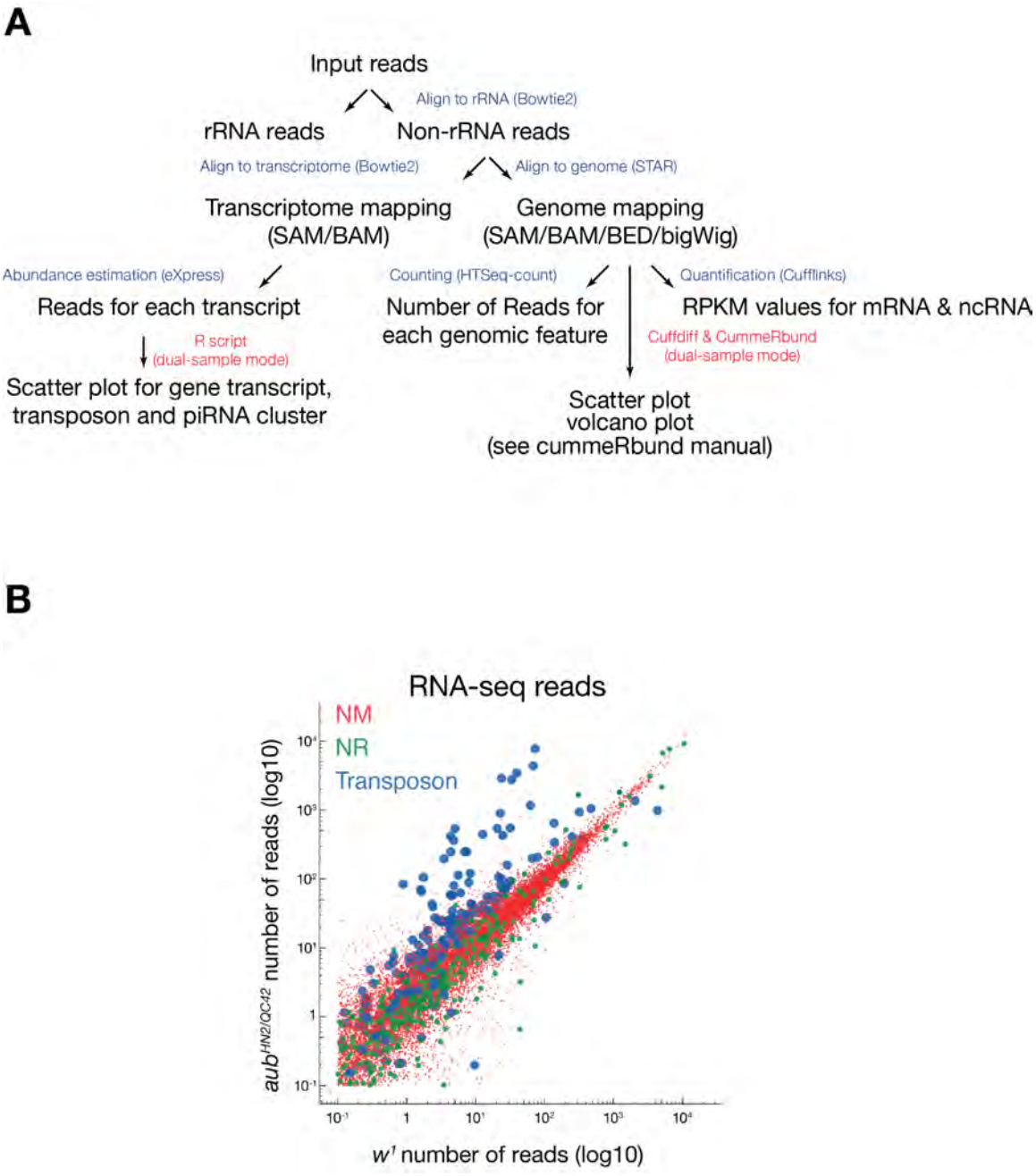

A
Input reads
Align to rRNA (Bowtie2)
rRNA reads
Non-rRNA reads
Align to transcriptome (Bowtie2)
Align to genome (STAR)
Transcriptome mapping
(SAM/BAM)
Genome mapping
(SAM/BAM/BED/bigWig)
Abundance estimation (eXpress)
Counting (HTSeq-count)
Quantification (Cufflinks)
Reads for each transcript
Number of Reads for
each genomic feature
RPKM values for mRNA & ncRNA
R script
(dual-sample mode)
Cuffdiff & CummeRbund
(dual-sample mode)
Scatter plot for gene transcript,
transposon and piRNA cluster
Scatter plot
volcano plot
(see cummeRbund manual)
B
RNA-seq reads
NM
NR
Transposon
aub^HN2/QC42 number of reads (log10)
w1 number of reads (log10)

Figure Legend 4.2. Flowchart and Example Figures For the RNA-seq Pipeline

(A) Work flow for the pipeline in single- (blue) and dual-library mode (red).

(B) Scatter plot comparing *w*^{1} to *aub*$^{HN2/QC42}$ *Drosophila* ovary RNA-seq reads assigned to mRNA (NM; red), non-coding RNA (NR; green) and transposons (blue).

Degradome- and CAGE-seq pipeline

The mapping of degradome- or CAGE-seq reads to the genome and transcriptome is similar to RNA-seq, except that the pipeline discard alignments whose 5′ ends cannot be accurately determined (e.g., soft-clipped). Like the small RNA-seq and RNA-seq pipelines, the degradome pipeline uses BEDTools to assign the alignments to different genomic features, then plots nucleotide frequency around the 5′ ends of the reads (Figure 4.3). Direct transcriptome mapping and quantification is also done as in RNA-seq pipeline.

Figure 4.3

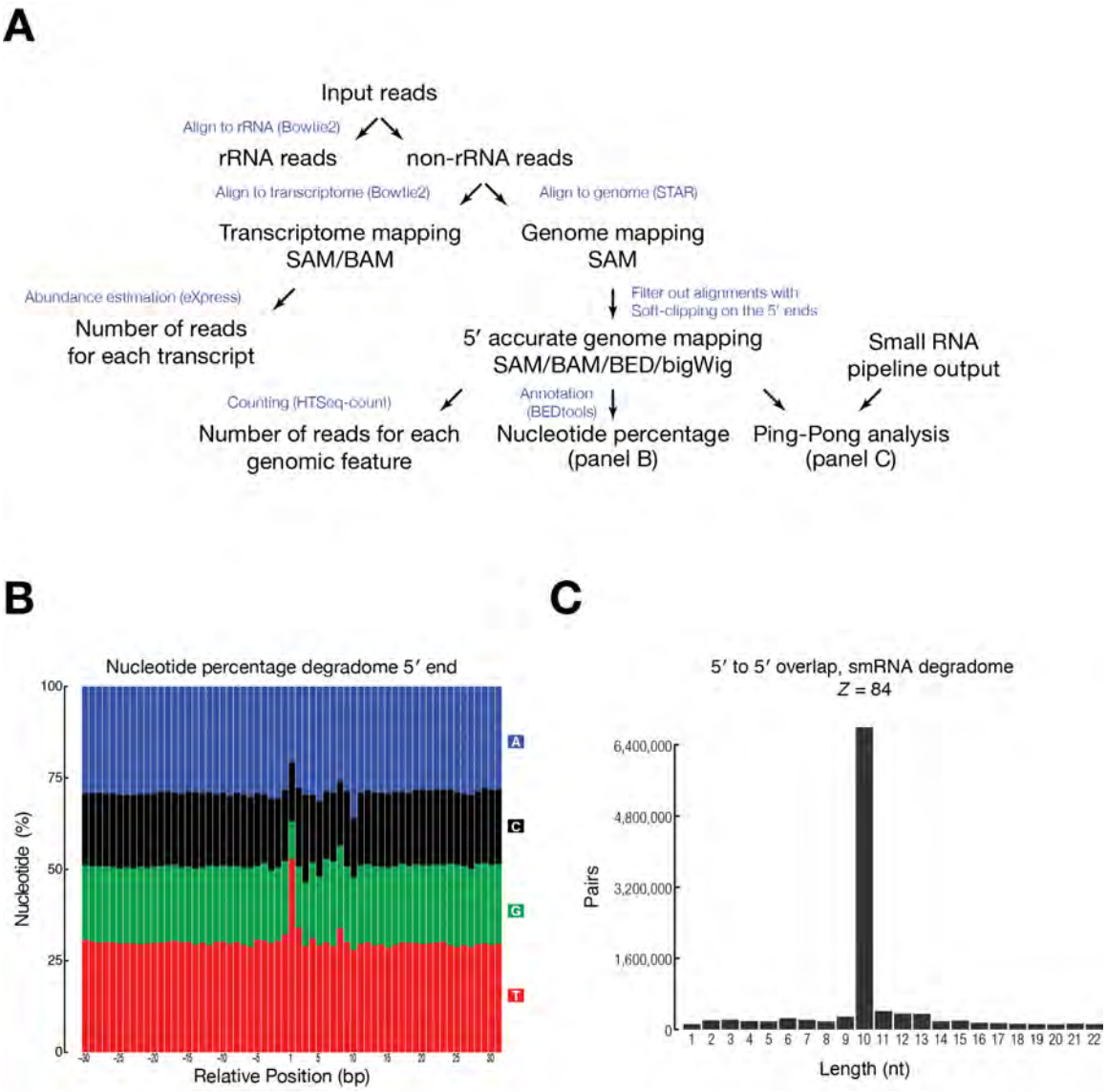

A
Input reads
Align to rRNA (Bowtie2)
rRNA reads
non-rRNA reads
Align to transcriptome (Bowtie2)
Align to genome (STAR)
Transcriptome mapping
SAM/BAM
Genome mapping
SAM
Abundance estimation (eXpress)
Number of reads
for each transcript
Filter out alignments with
Soft-clipping on the 5′ ends
5′ accurate genome mapping
SAM/BAM/BED/bigWig
Small RNA
pipeline output
Counting (HTSeq-count)
Annotation
(BEDtools)
Number of reads for each
genomic feature
Nucleotide percentage
(panel B)
Ping-Pong analysis
(panel C)
B
Nucleotide percentage degradome 5′ end
Nucleotide (%)
Relative Position (bp)
A
C
G
T
C
5′ to 5′ overlap, smRNA degradome
Z = 84
Pairs
Length (nt)

Figure Legend 4.3. Flowchart and Example Figures for the Degradome- and CAGE-seq Pipeline

(A) Work flow for the pipeline.

(B) Nucleotide percentage 30 nt upstream and downstream of the 5′ end of *Drosophila* w^1 degradome reads as-signed to piRNA clusters.

(C) Bar plot representing 5′ to 5′ overlapping (Ping-Pong signature) analysis between *Drosophila* w^1 ovary small RNA-seq and degradome-seq reads that are assigned to piRNA clusters.

ChIP-seq pipeline

piPipes aligns ChIP input and IP reads to the genome using Bowtie2 (Liu and Schmidt, 2012), then calls peaks using MACS2 (Zhang et al., 2008), which supports both narrow (e.g., transcriptional factors) and broad peaks (e.g., H3K9me3). The alignments are then converted to normalized signals, which are used by bwtool (Pohl and Beato, 2014) to perform TSS, TES, and metagene analyses for each genomic feature. In dual-library mode, piPipes uses MACS2 to call differential binding events using non-normalized alignments. TSS, TES and metagene analyses are provided for those loci that are identified to be differentially enriched (Figure 4.4).

Figure 4.4

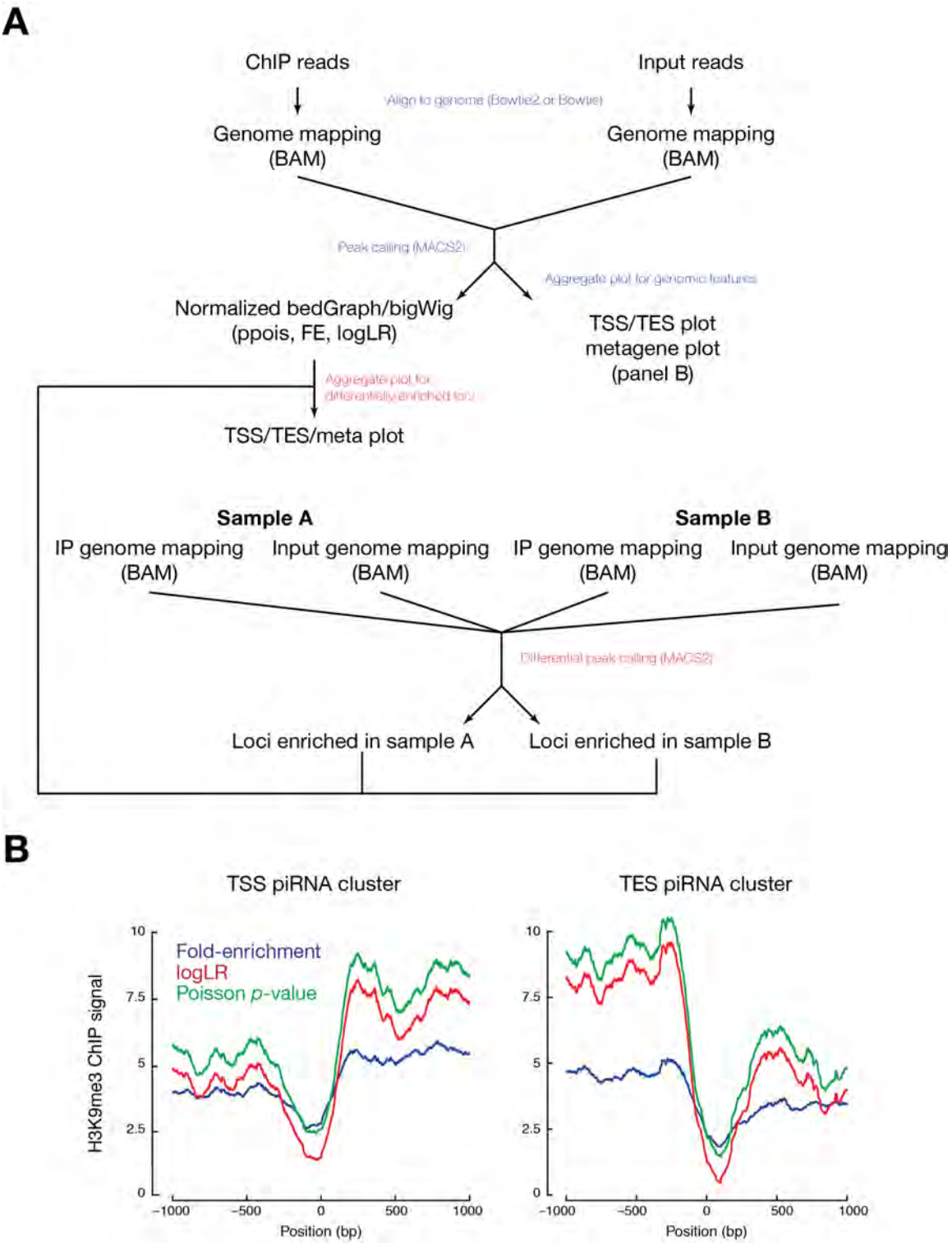

A
ChIP reads
Input reads
Align to genome (Bowtie2 or Bowtie)
Genome mapping (BAM)
Genome mapping (BAM)
Peak calling (MACS2)
Aggregate plot for genomic features
Normalized bedGraph/bigWig (ppois, FE, logLR)
TSS/TES plot metagene plot (panel B)
Aggregate plot for differentially enriched loci
TSS/TES/meta plot
Sample A
Sample B
IP genome mapping (BAM)
Input genome mapping (BAM)
IP genome mapping (BAM)
Input genome mapping (BAM)
Differential peak calling (MACS2)
Loci enriched in sample A
Loci enriched in sample B
B
TSS piRNA cluster
TES piRNA cluster
Fold-enrichment
logLR
Poisson p-value
H3K9me3 ChIP signal
Position (bp)
Position (bp)

Figure Legend 4.4. Flowchart and Example Figures of ChIP-seq Pipeline

(A) Work flow for the pipeline in single- (blue) and dual-sample mode (red).

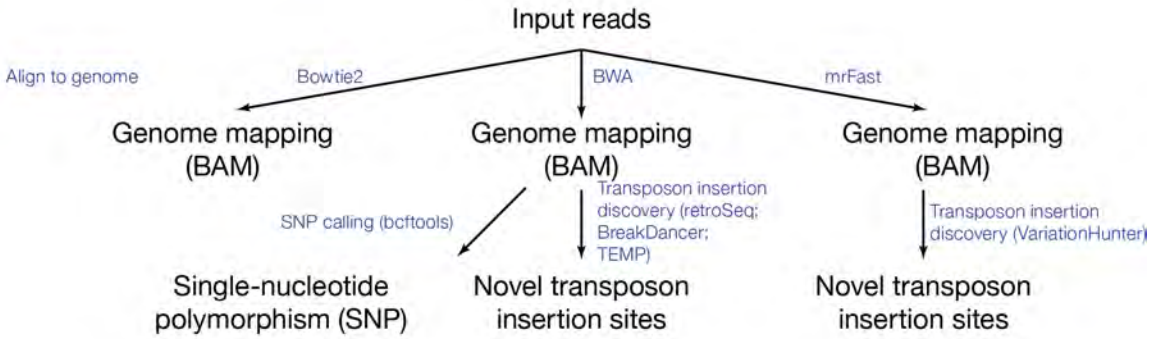
(B) Example plots for TSS (left), TES (right) analyses of H3K9me3 ChIP-seq enrichment for piRNA cluster calculated using three different statistical methods; see the MACS2 manual for detailed information. Data used here is from Drosophila ovary with RNAi against *piwi* (SRX215630).

Genome Sequencing Pipeline

piPipes aligns genomic sequencing reads to the genome with three different aligners, Bowtie2 (Liu and Schmidt, 2012), BWA (Li and Durbin, 2009) and mrFast (Alkan et al., 2009) to best fit the preference of different software used downstream. To perform Structural Variation (SV) Analysis and identify transposon insertion or deletion events, the genome-seq pipeline applies different algorithms, including BreakDancer (Chen et al., 2009b), RetroSeq (Keane et al., 2013; Hormozdiari et al., 2010), VariationHunter, and TEMP (Zhuang et al., 2014), to discover transposon insertion, deletion, and other structural variation (SV) events (Figure 4.5). piPipes uses a Circos plot (Zhang et al., 2013) to represent the variant loci discovered by each algorithm across different chromosomes.

Figure 4.5

A


Input reads
Align to genome
Bowtie2
BWA
mrFast
Genome mapping (BAM)
Genome mapping (BAM)
Genome mapping (BAM)
SNP calling (bcftools)
Transposon insertion discovery (retroSeq; BreakDancer; TEMP)
Transposon insertion discovery (VariationHunter)
Single-nucleotide polymorphism (SNP)
Novel transposon insertion sites
Novel transposon insertion sites

B

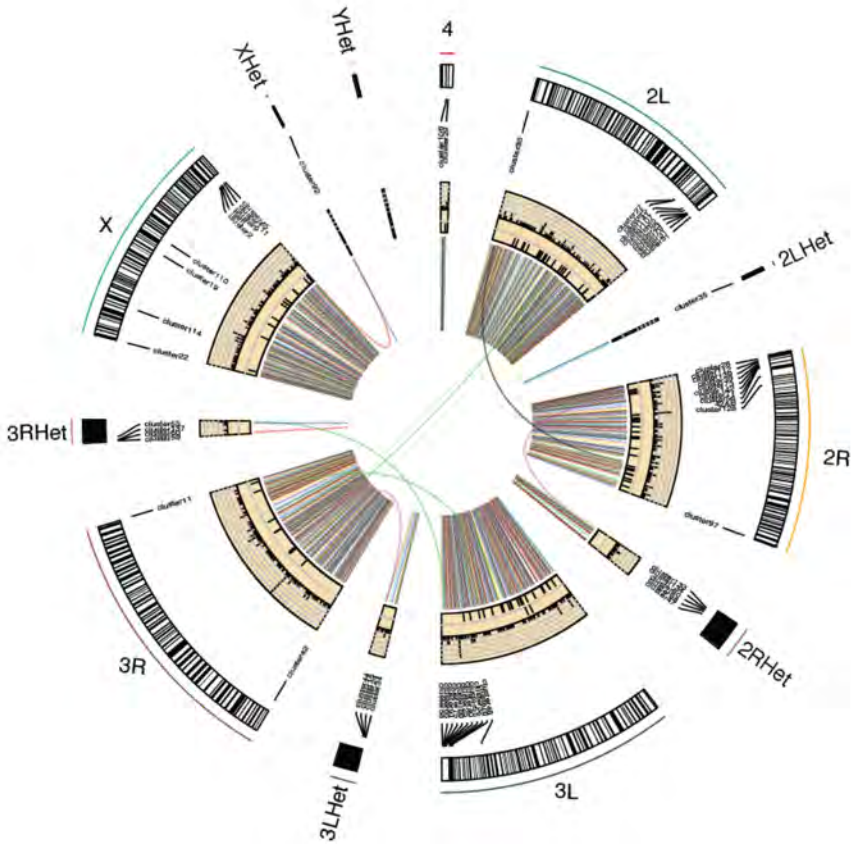

YHet
XHet
4
2L
X
2LHet
3RHet
2R
2RHet
3R
3L
3LHet

Figure Legend 4.5. Flowchart and Example Figures of Genome-seq pipeline

(A) Work flow for the Genomic-seq pipeline.

(B) Circos plot representing the locations of, from the periphery to the center, cytological position, piRNA clusters, SV discovered by TEMP (tiles), retroSeq (tiles) and Variation-Hunter (links) using genomic sequencing of 2–4 day-old ovaries from female offspring from the cross w^{1} × Harwich (SRX093065).

Dual-sample Comparison

The small RNA-seq, RNA-seq and ChIP-seq pipelines can each be run in two modes, allowing analysis of a single sample or a pair of samples. The dual-sample mode uses the output from the single-sample mode and performs pair-wise comparison as illustrated by balloon plots and scatter plots (Figure 4.1B and D). The comparison can be performed on miRNA, piRNA, or mRNA. Figure 4.2B illustrates a scatter plot showing for the mRNA abundance in the RNA-seq data set produced by the RNA-seq pipeline in the dual-sample mode. The dual-sample mode of the RNA-seq pipeline also uses Cuffdiff (Trapnell et al., 2013) to perform differential analysis on genic transcripts. In the dual-sample mode, the ChIP-seq pipeline uses MACS2 to identify differentially enriched loci (Figure 4.4).

Uniquely and Ambiguously Mapping Reads

The repetitive nature of transposons makes it desirable to analyze ambiguous mappers under some circumstances. The small RNA pipeline separately counts reads mapping to a single genomic location and reads mapping to more than one location, and then divides the abundance of each read by the number of loci to which it can be assigned. For RNA-seq and degradome/CAGE-seq, piPipes uses eXpress (Roberts and Pachter, 2013) to assign unambiguous reads with an online expectation-maximization (EM) algorithm. In ChIP-seq, piPipes calls Bowtie2 to randomly report only one of the best alignments for each ambiguous read. Incorporating multiple mappers in the analysis avoids neglecting repetitive regions, which are the chief sources or targets of piRNAs in many animals;

counting only one alignment for each read prevents artefactually enriching for repetitive regions.

Acknowledgments

We thank the members of the Zamore and Weng laboratories for helpful discussions, and Jia Xu, Jui-Hung Hung and Soo Lee for their initial work. This work was funded by National Institutes of Health [R37] to PDZ and [U41HG007000] to ZW.

Chapter V Tailor: A computational framework for detecting non-templated tailing of small silencing RNAs

Disclaimer

This chapter was a product of a collaborative effort among the authors: Min-Te Chou (MTC), Bo W Han (BWH), Chiung-Po Hsiao (CPH), Phillip D. Zamore (PDZ), Zhiping Weng (ZW), and Jui–Hung Hung (JHH). MTC and BWH implemented Tailor. BWH implemented the Bash/Shell pipeline. CPH performed quality test. PDZ, ZW and JHH supervised the project.

Summary

Small silencing RNAs, including microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs) and Piwi-interacting RNAs (piRNAs), have been shown to play important roles in fine-tuning gene expression, defending virus and controlling transposons. Loss of small silencing RNAs or components in their pathways often leads to severe developmental defects, including lethality and sterility. Recently, non-templated addition of nucleotides to the 3′ end, namely tailing, was found to associate with the processing and stability of small silencing RNAs (Ji and Chen, 2012; Li et al., 2005; Ren et al., 2012; Ameres et al., 2010; Ameres and Zamore, 2013). Next Generation Sequencing has made it possible to detect such modifications at nucleotide resolution in an unprecedented throughput. Unfortunately, detecting such events from millions of short reads confounded by sequencing errors and RNA editing is still a tricky problem. Here, we developed a computational framework, Tailor, driven by an efficient and accurate aligner specifically designed for capturing the tailing events directly from the alignments without extensive post-processing. The performance of Tailor was fully tested and compared favorably with other general-purpose aligners using both simulated and real datasets for tailing analysis. Moreover, to show the broad utility of Tailor, we used Tailor to reanalyze published datasets to reveal novel findings worth further experimental validation. The source code and the executable binaries are freely available at https://github.com/jhhung/Tailor.

Introduction

Over the past decade, small silencing RNAs, including miRNAs, endogenous small silencing RNAs (endo-siRNAs) and piRNAs have been shown to play indispensable roles in regulating gene expression, protecting against viral infection and preventing mobilization of transposable elements (Ameres and Zamore, 2013; Luteijn and Ketting, 2013; Stefani and Slack, 2008; Siomi et al., 2011; Luteijn and Ketting, 2013; Stefani and Slack, 2008; Siomi et al., 2011). Small silencing RNAs exert their silencing function by associating with Argonaute proteins to form RNA-induced silencing complex (RISC), which uses the small RNA guide to find its regulatory targets and reduce gene expression (Meister, 2013). Although the studies on the biogenesis of small silencing RNAs have made enormous progress in the past decade, the factors controlling their stability and degradation remain elusive.

Recent studies have suggested that non-templated addition to the 3′ end of small silencing RNAs, namely tailing, could play essential roles in this regard. Non-templated 3′ mono- and oligo-uridylation of the pre-microRNAs (pre-miRNAs) regulates miRNA processing by either preventing or promoting Dicer cleavage in flies (Heo et al., 2008; Heo et al., 2009; Heo et al., 2012). The 3′ mono-uridylation on small interfering RNAs in *Caenorhabditis elegans* is associated with negative regulation (van Wolfswinkel et al., 2009). *Ameres et al.* have demonstrated that highly complementary targets trigger the tailing of miRNAs and eventually lead to their degradation in flies and mammals (Ameres

et al., 2010; Xie et al., 2012); a similar mechanism has been found on some endo-siRNAs as well (Ameres et al., 2011). Identification of tailing events not only suggests the co-evolution of small silencing RNAs and their targets, but also sheds light on the mechanism of their maturation and degradation. Despite the fact that Next Generation Sequencing (NGS) has greatly facilitated the understanding of RNA tailing, computational detection of non-templated nucleotides from millions of sequencing reads is challenging. The *Ketting* group used MegaBLAST to align piRNA sequences to the genome and relied on post-processing the reported mismatches to gain insights into tailing (van Wolfswinkel et al., 2009). However, as a heuristic algorithm, BLAST is not guaranteed to find all the tailing events (Zhang et al., 2000; Altschul et al., 2013) and it is significantly slower than the NGS aligners, like MAQ (Li et al., 2008a), BWA (Li and Durbin, 2009), Bowtie (Langmead et al., 2009) and SOAP (Li et al., 2008b; Li et al., 2009). The *Chen* group used an accurate method that iterates between Bowtie alignment and 3′ clipping of unmatched reads (Zhao et al., 2012) to find all the perfect alignments of trimmed reads. A similar approach has been used for removing erroneous bases at 3′ end to increase the sensitivity of detecting miRNAs (Marco and Griffiths-Jones, 2012). Let alone that this method inevitably multiples the running time by the maximal length of tails, extra computational works are still needed to retrieve the identity of each trimmed tail. The study by *Ameres et al.* used a specialized suffix tree data structure to efficiently find all the tails without sacrificing the accuracy (Ameres et al., 2010). However, due to the

high memory footprint of suffix tree, which is about 16 to 20× of the genome size, the read mapping has to be performed for each chromosome separately (Ameres et al., 2010; Ameres et al., 2011; Xie et al., 2012). Extra processing is still required to finalize the alignments from all chromosomes.

Moreover, the task becomes even trickier when technical and biological confounding factors are taken into account for better capturing the true tailing events. For example, it is known that reads from Illumina HiSeq and Genome analyzer platforms have preferential A–C conversions (Dohm et al., 2008; Qu et al., 2009) and a high error rate at the 3′ end of reads, which frequently leads to uncalled bases, i.e. B-tails (Minoche et al., 2011; Le et al., 2013). In addition to these technical artifacts, RNA editing is another common post-transcriptional modification in small silencing RNA biology that could perplex the tools with erroneous alignment. There are two major types of RNA editing in mammals, adenosine to inosine (A-to-I) and cytidine to uridine (C-to-U) editing. The major enzymes that catalyze adenosine to inosine are the adenosine deaminases acting on RNA (ADARs), whose main substrates are RNAs with double-stranded structures (Blow et al., 2004; Kim et al., 2004; Morse et al., 2002). Since many small silencing RNAs are originated from structural RNAs, they are all likely targets of A-to-I editing (Blow et al., 2006; Luciano et al., 2004; Warnefors et al., 2014). Recent studies have shown that A-to-I editing can occur on the seed region of the miRNAs with fairly high occurrence rate (up to 80% in some cases) and have a direct impact on the selection of their regulatory targets (Kume et al.,

2014; Vesely et al., 2014). Those unmatched bases degenerate the sensitivity and accuracy of short read alignment and have a negative effect on the detection of tailing.

Most of the current methods simply ignore those confounding factors and rely on adapting existing, less specialized tools with extensive post-processing and as a consequence the performance, usefulness and application of tailing analysis is seriously compromised. A fast, accurate and straightforward approach to study tailing is still in need. To ease the cost of performing tailing analysis with dramatically increasing sequencing throughput, we here introduce Tailor–a framework that preprocesses and maps sequences to a reference, distinguishes tails from mismatches or bad alignments with a novel algorithm and reports both perfect and tailed alignment simultaneously without loss of information. Tailor is capable of analyzing the non-templated tailing for miRNA and other types of small RNAs and produce publication-quality summary figures. In addition, to better demonstrate the utility of Tailor, we reanalyzed published datasets with Tailor and unearthed several interesting observations. Although the findings still require thorough experimental validation, it is clear that Tailor would help expand the scope of the study of small silencing RNAs.

Methods

The principle of detecting non-templated bases at the 3′ end of reads is basically to find the longest common prefix (LCP) between the read and each of the suffixes of the reference and then report the remainder on the read as a tail. Given a read R (M base pairs [bp] long) and all the suffixes (S_i) of a reference sequence G (N bp long), one can find the LCP between R and S_i by finding the longest consecutive matches from the first base to the last. Since there are totally N suffixes of G, a trivial solution needs at worse $M \times N$ times of comparison to find the LCP of R and G; however the performance is unacceptably slow when G is as large as a human genome. Using index structures, such as the suffix tree or suffix array (SA), finding LCPs between the NGS reads and the reference can be solved much more efficiently (Ameres et al., 2010; Dobin et al., 2013).

Recently, the Full-text index in Minute space (FM-index) derived from the Burrows-Wheeler transform (BWT; Burrows and Wheeler, 1994; Ferragina and Manzini, 2000; Burkhardt and Kärkkäinen, 2003) is widely used in many NGS applications (Li and Durbin, 2009; Langmead et al., 2009; Li et al., 2009). The FM-index is both time and space efficient and can be built from a suffix array and requires only 3 to 4 bits per base to store the index. However, since the FM-index is originally designed for matching all bases of a read to a substring of the reference, it cannot be used directly for finding tails. One straightforward solution is to align reads without those non-templated bases by repeatedly removed one last base in each round of the alignment process until at least one perfect hit is

found, but the approach scarifies the speed greatly and requires extensive post-processing. To benefit from the space and time efficiency of the FM-index, we further modified its matching procedure and adapted the error tolerant strategy proposed by *Langmead et al.* to devise an FM-index based tail detection algorithm (Langmead et al., 2009), Tailor, which is specialized in capturing the non-templated bases at the 3′ end of reads with confounding factors, such as sequencing errors and RNA editing.

Construction of the Burrows-Wheeler Transformed Genome

Tailor first computes the suffix array of the concatenated Watson and Crick strands of the genome, which can be built by sorting all suffixes of the sequence of the concatenation of the plus and minus strand of the genome in lexicographical order. Biological sequences are usually filled with long repeats, which make the construction of SA degenerates to quadratic time and obstructs the practical use. To handle repetitions in linearithmic time, Tailor adapts the difference cover sample (DCS) data structure as proposed to accelerate the sorting (Burkhardt and Kärkkäinen, 2003). The DCS is the data structure that ensures an anchor pair (whose order is known) can be found for any pair of suffixes (order unknown) within a small offset. With the help of DCS, one can determine the relative order of two suffixes in linearithmic time even with the present of long repeats, which achieves the construction of SA of a genome in reasonable time.

For the human genome, about 15G of main memory is required for the construction of its SA. A feasible solution is to compute the corresponding block of the BWT of the complete text separately. Tailor generates the splitter as follows: First, the size of each block is pre-specified as 30 millions (M) which is taking into account the efficiency of sorting and its memory usage in the later stage. The number of blocks (k) is then calculated by k = N / 30M, where N is the length of the genome times 2 (both strands are counted). Second, Tailor randomly generates about 4k suffix indexes and keeps only the unique indexes, whose size is denoted as u, and then sorts the corresponding suffixes with the help of the DCS. Finally, from the sorted suffixes, Tailor picks the splitters with an interval of u/(k-1). Tailor then classifies all suffixes to each block by comparing the lexicographical order of the sequences of the suffixes and the splitters.

Finally, Tailor sorts the suffixes in each block using the bucket sort. The bucket sort is a non-comparison sort, and the average case time complexity is O (n+Ω), where n is sequence length, and Ω is the size of the alphabet, which contains 4 letters (A, T, C, and G). However, the worst case space complexity of the bucket sort is O (n×Ω), which is too large to store in the memory in some case, so Tailor uses multi-key quick sort and DCS to pre-sort the suffix indexes, until the size of the unsorted suffix indexes is less than 4 millions. Finally, when all blocks are sorted, the indexes are collected accordingly and the SA of the genome is constructed.

Constructing the FM-index

Since biological sequences have a relatively small alphabet, the transformed sequence (i.e., the BWT of the genome) can be further compressed to save space. For example, DNA sequences are consists of four nucleotides, A, T, C, and G. One would only need two bits to store a nucleotide, which can greatly reduced the memory usage by 75%. Another edge of compressing the BWT string is that the inverse algorithm can recover 4 nucleotides in reading one single byte and reduce the time calculating position in searching with the help of some additional lookup tables. Other two important tables in the FM-index are the C table and the Occ table. Tailor maintains the contents of them along with the construction of the SA, but since the memory usage of Occ table is also huge, so Tailor only records a part of the Occ table in an interval of 64, and others are calculated on the fly as previously suggested (Langmead et al., 2009).

Searching for Prefix Matching

As described previously, the query is first reverse complemented and Tailor starts to find the converted query in the FM-index by the backward searching algorithm. If the query contains a 3′ tail, the backward searching of the converted query will stop before reaching the end. When the backward searching stops, Tailor records a pair of indexes of the BWT, which indicating all the suffixes that share the same prefix, whose content is the same as the suffix of the converted query (i.e., the prefix of the original query). Starting from each of the suffixes marked by the pair of indexes, Tailor then uses the inverse BWT algorithm to

backtrack to the very first base of the genome (Burrows and Wheeler, 1994). The number of bases traversed before reaching the end implicating the chromosomal location of each suffix. To accelerate this process, Tailor keeps a portion of the SA along with an auxiliary data structure as a fast lookup table, which assures the chromosomal location represented by the index of the BWT can be retrieved within a bounded number of backtracking.

The system flow of the Tailor algorithm is outlined in Figure 5.1. Since searching within the FM-index initiates from the 3′ end of the query string (i.e., the read), where the non-templated nucleotides append, Tailor first makes the reverse-complement of the query sequence so that searching starts from the original 5′ end to avoid excessive exhaustive search at the early stage. To do so, the reference should be reversed complemented as well, and the coordinate of each alignment should be calculated accordingly. To allow searching against both strands simultaneously and improves the speed, Tailor concatenates the plus and minus strands of the reference and constructs one index instead of two. Tailor also stores a part of the suffix array similar to other FM-index based aligners to achieve fast calculation of the text shift for getting the coordinate of each occurrence. Any alignment whose prefix matching portion exceeds the boundary of the mapped chromosome is filtered. The searching continues until either it matches all the characters of the query to the reference (i.e., the perfect matching) or no more bases can be matched (i.e., the prefix matching). In the latter case, Tailor backtracks to the previous matched position and exhaustively

enumerates all the possible prefix matches. The unmatched part remained in the query is reported as a tail (Figure 5.1B).

Figure 5.1

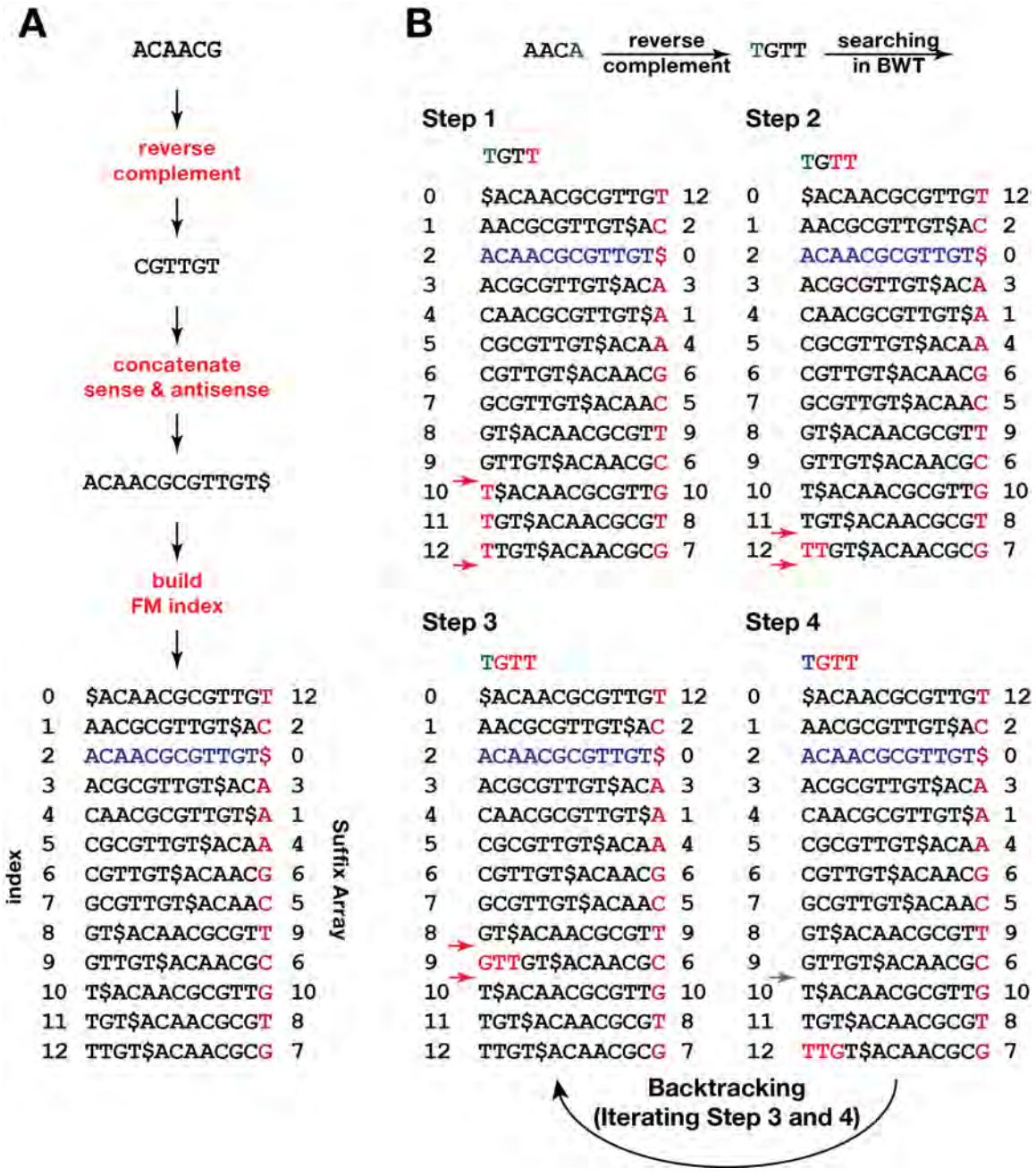

A
ACAACG
reverse complement
CGTTGT
concatenate sense & antisense
ACAACGCGTTGT$
build FM index
index
Suffix Array
0 $ACAACGCGTTGT 12
1 AACGCGTTGT$AC 2
2 ACAACGCGTTGT$ 0
3 ACGCGTTGT$ACA 3
4 CAACGCGTTGT$A 1
5 CGCGTTGT$ACAA 4
6 CGTTGT$ACAACG 6
7 GCGTTGT$ACAAC 5
8 GT$ACAACGCGTT 9
9 GTTGT$ACAACGC 6
10 T$ACAACGCGTTG 10
11 TGT$ACAACGCGT 8
12 TTGT$ACAACGCG 7
B
AACA
reverse complement
TGTT
searching in BWT
Step 1
TGTT
Step 2
TGTT
Step 3
TGTT
Step 4
TGTT
Backtracking
(Iterating Step 3 and 4)

Figure Legend 5.1. BWT-based Tailing Detection Algorithm

(A) Genomic index construction procedure.

(B) Read searching procedure.

Implementation

We implemented the core of the Tailor aligner using C++ with built-in support for multithreading. Since Tailor concatenates both strands of the chromosomes into one long reference, whose length could exceed the maximum number represented by 32 bits, we have to use 64 bits to store the indexes in all the relevant data structures, which require about 2× memory footprint than that of other FM-index based aligners. Tailor has a similar command line interface like other NGS aligners, and reports alignment in the SAM format. A tail is described as "soft-clipping" in CIGAR and the sequences are reported under "TL:Z:" in the optional fields . Mismatches, if allowed (–v), will be reported in the "MD" tag. Tailor is freely available on GitHub (http://jhhung.github.io/Tailor/) under GNU General Public License 2. The tailing pipelines were implemented in shell scripting language and R.

Results

Performance without confounding factors

To begin with, we ignored confounding factors in the following tests to compare with conventional approaches first. To assess the aligning speed directly, we indiscriminately generated 10 millions of perfectly genome-matching reads from the *Drosophila melanogaster* genome (simulated tail-free dataset) and randomly appended 1–4 genome-unmatched nucleotides to the 3′ ends (simulated tailed dataset). We compared Tailor with two most popular BWT aligners Bowtie and BWA by applying them on simulated small RNA datasets (Figure 5.2A). For the simulated tail-free dataset, Tailor outperformed Bowtie and BWA in five thread settings (using 2, 4, 8, 12, and 24 threads; Figure 5.2A, top). But for the simulated tailed dataset, Bowtie ran slightly faster than Tailor possibly due to the fact the it reported no alignment and did not perform any disk writing (Figure 5.2A, bottom). We also performed the speed test with real small RNA sequencing data from *hen1*$^{+/-}$ and *hen1*$^{-/-}$ fruitfly and zebrafish (Figure 5.2B). *hen1* encodes for a methyl-transferase that adds a methyl group to the 3′ end of siRNA and piRNA at the 2′-*O* position and prevents tailing. For both *hen1*$^{+/-}$ and *hen1*$^{-/-}$ libraries, Tailor outperformed Bowtie and BWA and reproduced the published result that siRNAs, but not miRNAs, were subjected to tailing in the absence of *hen1* (Figure 5.2B). Please note that Bowtie and BWA in the speed test setting here were not capable of detecting non-templated tails. These tests were just used to compare their execution speed but not functionality.

To prove the accuracy of Tailor when confounding factors were not considered, we then used either Tailor or the *Chen* method to identify the non-templated tailing events. To achieve maximal speed of the Chen method to our best knowledge, we used the "*-3 k*" option of Bowtie to clip *k* bases off from the 3′ end of each read. This strategy avoided calling secondary programs and ensured that minimal computational work was done other than Bowtie mapping. We started the alignment by setting *k* to 0. After the initial mapping, the unaligned reads were realigned with an incremented *k* (*k* = 1). This process was repeated four times. In the last iteration, four nucleotides were trimmed off from the 3′ end (*k* = 4) and all the tailed reads should have been mapped at this point. In the simulation test, this method finished in 67 ± 1 seconds with Bowtie been called five times (*k* = 0–4). Not surprisingly, directly mapping by Tailor finished in 22 ± 1 seconds in the same computational environment. Both methods reported the same coordinates. However, in such setting, the *Chen* method was not able to identify the tails, which requires considerable computational work and time to retrieve from the raw reads. In contrast, Tailor revealed the length and the identity of the tails in the alignment output directly.

Figure 5.2

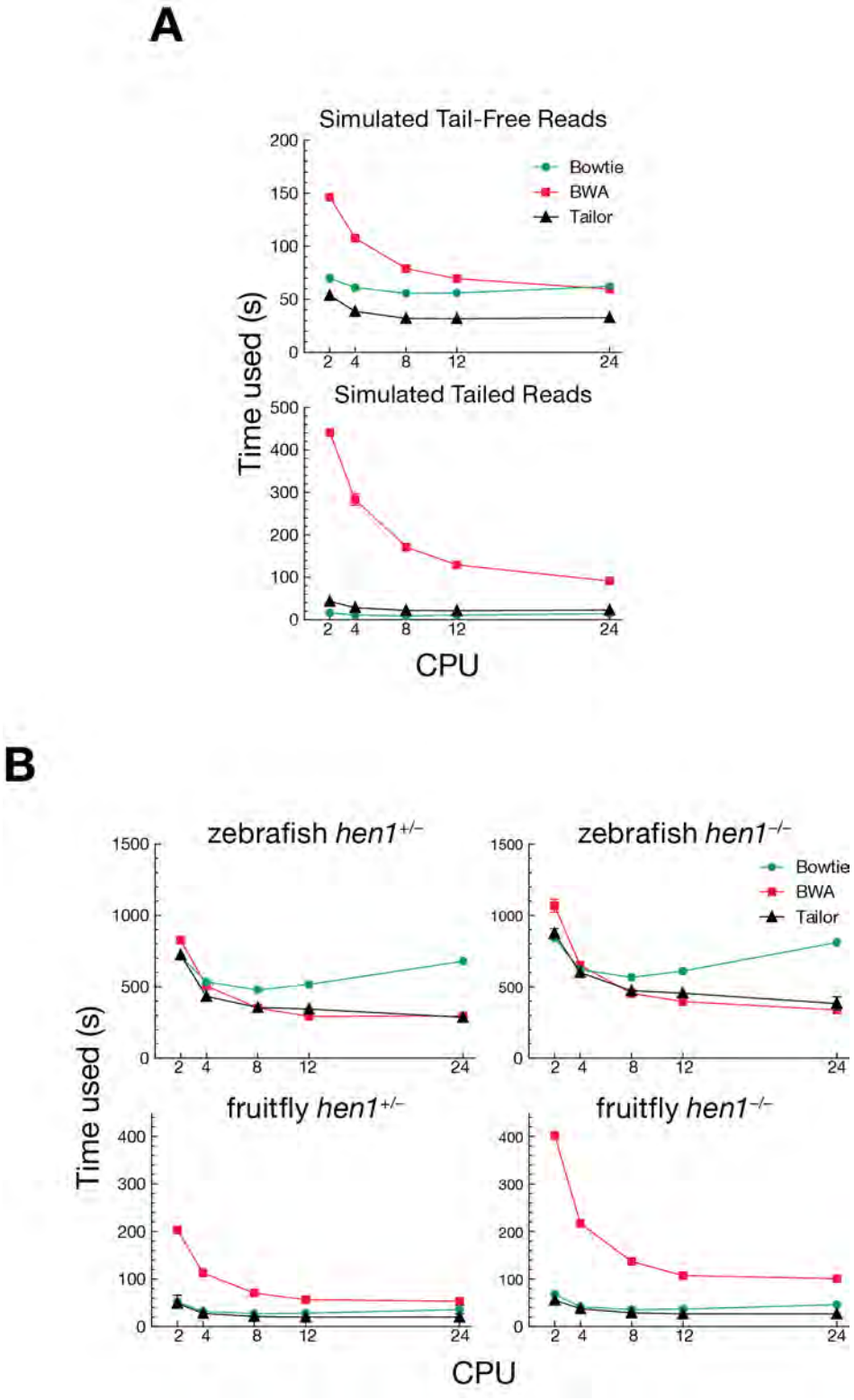

A
Simulated Tail-Free Reads
Bowtie
BWA
Tailor
Time used (s)
Simulated Tailed Reads
CPU
B
zebrafish hen1+/–
zebrafish hen1–/–
fruitfly hen1+/–
fruitfly hen1–/–
Time used (s)
CPU

Figure Legend 5.2. Speed Comparison of Tailor, Bowtie2, and BWA

(A) Speed comparison between Tailor, BWA and Bowtie using simulated 18–23 nt small RNA with (top) or without (bottom) non-templated tails. Tailor ran with the default setting, which allows no mismatch in the middle of the query. Tailed alignments were reported if perfect match could not be found. Bowtie ran with '−a –best –strata −v 0' setting to allow no mismatch while report all best alignments. BWA ran with the default setting. Five different CPU settings were used and the running time was plotted. Three replicates were performed.

(B) Speed comparison between Tailor, BWA and Bowtie using published small RNA Illumina NGS libraries from *hen1*$^{+/-}$ and *hen1*$^{-/-}$ mutants in fruitfly and zebrafish. Same settings were used as in (A).

Performance with error tolerance

It is arguable that some NGS aligners that support local alignment, such as Bowtie2 and BWA (Liu and Schmidt, 2012; Langmead and Salzberg, 2012), can recover those tails with error tolerance. We simulated two datasets (one normal, one mutated, see below) whose distribution of read length follows that of the real small RNA sequencing dataset. For the normal dataset, two million reads were randomly sampled from the reference genome. We intentionally kept reads having just one unique occurrence in the genome and then appended a 1-4 nt non-templated tail on each read. For the mutated dataset, a similar procedure was used to generate another two million reads, but one additional step was added: we introduced one substitution in the nucleotides 2-8 of each read to simulate an RNA editing event as suggested by Vesely *et al.* (Vesely et al., 2014). Again, this substitution was picked carefully to have only one occurrence in the genome with exactly one mismatch. The simulation guaranteed that there existed only one best alignment to the reference for each read in both datasets.

Then we examined the mappability of these datasets by Tailor (allow mismatch), Bowtie2, and BWA (Figure 5.3A). Tailor clearly reported more unique mapping reads than others especially in the mutated datasets. When we looked closer to those reads that were mapped to multiple positions, we found Bowtie2 and BWA were more likely to align the tails to the reference than Tailor and create many alternative alignments. Note that the seed region setting was used to aid all three tools for the alignment (*S*=20 and –v in Tailor and the

equivalences in Bowtie2 and BWA; mismatches in the seed region were allowed), and all tools should try to align the first 20 nt of each read to the genome, but Bowtie2 and BWA still generated suboptimal alignments.

We further checked whether the alignments and the tails were correctly reported (Figure 5.3B). Tailor was the only tool that gave satisfactory results reporting correct alignments and tails in the mutated dataset. There was no information in the output of BWA to recover the tails, and since most of the reads were aligned to multiple loci, it was expected that extensive post-processing would be needed for extracting the tails. The simulation clearly shows that Tailor is the only practical solution for doing tailing analysis with confounding factors.

Figure 5.3

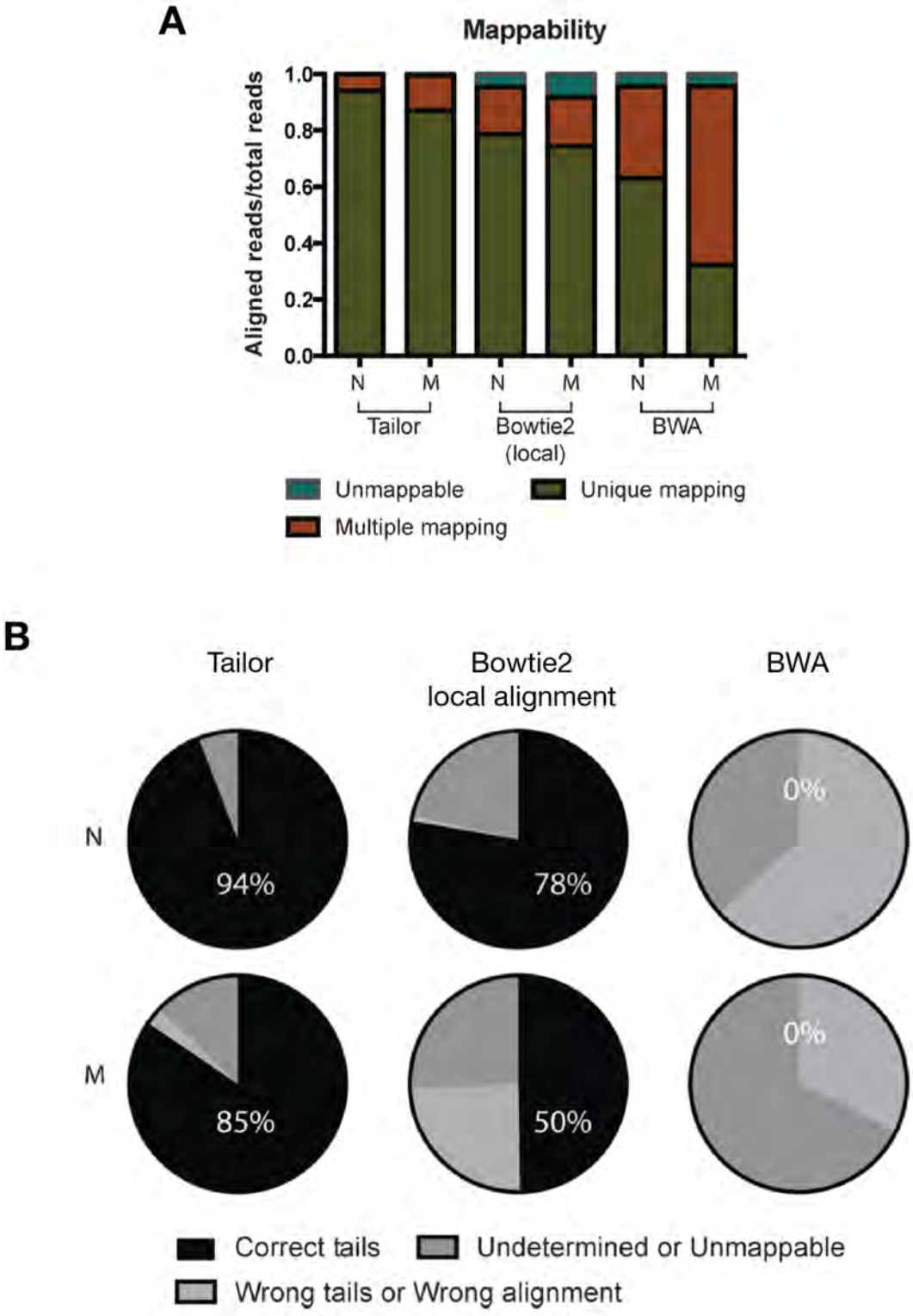

A
Mappability
Aligned reads/total reads
1.0
0.8
0.6
0.4
0.2
0.0
N
M
N
M
N
M
Tailor
Bowtie2
(local)
BWA
Unmappable
Unique mapping
Multiple mapping
B
Tailor
Bowtie2
local alignment
BWA
N
M
94%
78%
0%
85%
50%
0%
Correct tails
Undetermined or Unmappable
Wrong tails or Wrong alignment

Figure Legend 5.3. Accuracy Comparison of Tailor, Bowtie2, and BWA

(A) The mappability of the normal (N) and mutated (M) datasets aligned by Tailor, Bowtie2 (with local alignment) and BWA. Multiple mapping was deemed as misalignment since each read was guaranteed to have only one occurrence in the reference.

(B) The unique mapping reads shown in (A) were further examined to make sure they were aligned correctly and with proper tails reported (correct tails); unique mapping reads that didn't have correct alignment or tails were categorized another group (wrong tails/wrong alignment). The unmappable and multiple mapping reads were grouped together (undetermined or unmappable).

Analysis Pipeline

In order to provide a thorough and straightforward tailing analysis of deep sequencing libraries to the scientific community, we developed the interface of Tailor to take FastQ files as input and produce publication-ready figures. In brief, the input reads, with barcodes and adaptors removed, are subject to a quality-filtering step based on a PHRED score threshold provided by the user (e.g., to get rid of B-tails). The pipeline then applies Tailor to align the high-quality reads to the reference. The information on the length and identity of tails are then retrieved from the SAM formatted output and summarized to a tabular text file. Additionally, the alignments are assigned to different genomic features (miRNAs, exons, introns, et al.) using BEDTools (Quinlan and Hall, 2010). Tails from different categories are summarized. Publication quality figures depicting the length distribution are drawn using R package ggplot2 (Figure 5.4B). The pipeline also offers microRNA specific analysis. Balloon plots describing the 5′ and 3′ relative positions and the tails length are provided for a comprehensive overview (data not shown).

Figure 5.4

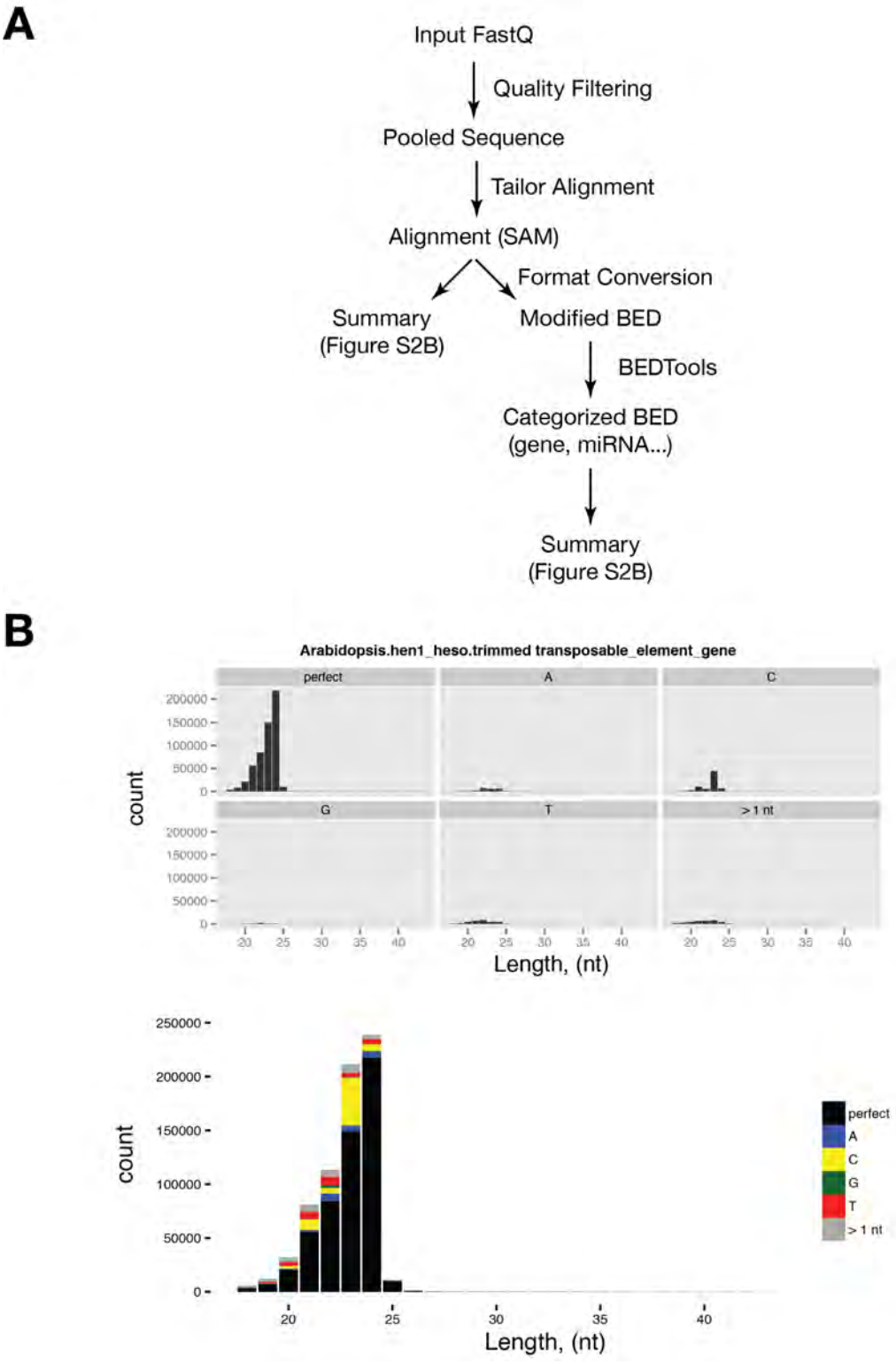

A
Input FastQ
Quality Filtering
Pooled Sequence
Tailor Alignment
Alignment (SAM)
Format Conversion
Summary
(Figure S2B)
Modified BED
BEDTools
Categorized BED
(gene, miRNA...)
Summary
(Figure S2B)
B
Arabidopsis.hen1_heso.trimmed transposable_element_gene
perfect
A
C
G
T
>1 nt
count
Length, (nt)
perfect
A
C
G
T
>1 nt

Figure Legend 5.4. Tailor Pipeline

(A) Flowchart of the Tailor pipeline.

(B) Example output of small RNAs categorized as “transposable element” in Arabidopsis *hen1*, *heso1* mutant. Perfect match, reads with one nucleotide tail (A, C, G, T), and reads with longer than one nucleotide tail are plotted separately (top) and together (bottom).

Applications—case studies

To prove the utility of Tailor, we applied Tailor to reanalyze several publicly available small RNA sequencing datasets and revealed new facts about the data that has not been reported yet. In plants, HEN1 methylates both miRNA and siRNA at their 3′ ends to protect them from non-templated uridylation catalyzed by HEN1 SUPPRESSOR1 (HESO1), a terminal nucleotidyl transferase that favors uridine as substrate (Zhao et al., 2012; Ren et al., 2012; Ren et al., 2014). We applied Tailor on small RNA sequencing libraries from WT, *hen1*$^{-/-}$ and *hen1*$^{-/-}$*; heso1*$^{-/-}$ cells of *Arabidopsis*, and the results showed that siRNAs were subjected to both non-templated uridylation and cytosylation without *HEN1* while miRNAs were mainly subjected to uridylation. Furthermore, the loss of *HESO1* only reduced the uridylation but not cytosylation of siRNAs, suggesting the existence of additional nucleotidyl transferase that prefers cytosine as substrates (Figure 5.5A).

We then applied Tailor to two NGS libraries that cloned Ago2 associated small RNA from nuclear and cytoplasmic fraction of HeLa cells respectively (Ameyar-Zazoua et al., 2012). Since RNAs were cloned using poly-A polymerase instead of 3′ adaptor ligation in the library preparation, A-tails were unable to be recovered computationally. Although most miRNAs showed very similar length distribution and tailing frequency between these two samples, one miRNA, miR-15a, exhibited a distinct pattern. In cytoplasm, miR-15a was mostly 21 nt long and had modest U tailing for its 22-mer isoform. Surprisingly, in the nuclear

fraction, miR-15a peaked at 22 nt and showed strong U tailing (Figure 5.5B). In addition, miR-15b, which shares its seed sequence with miR-15a and only has one nucleotide different from miR-15a in the first 19 nt of its mature sequence, did not exhibit obvious variation between the two samples. This suggests that, either 9–12 nt, also known as the “central site”, or the 3′ end of guide miRNA play an important role in tailing regulation.

Figure 5.5

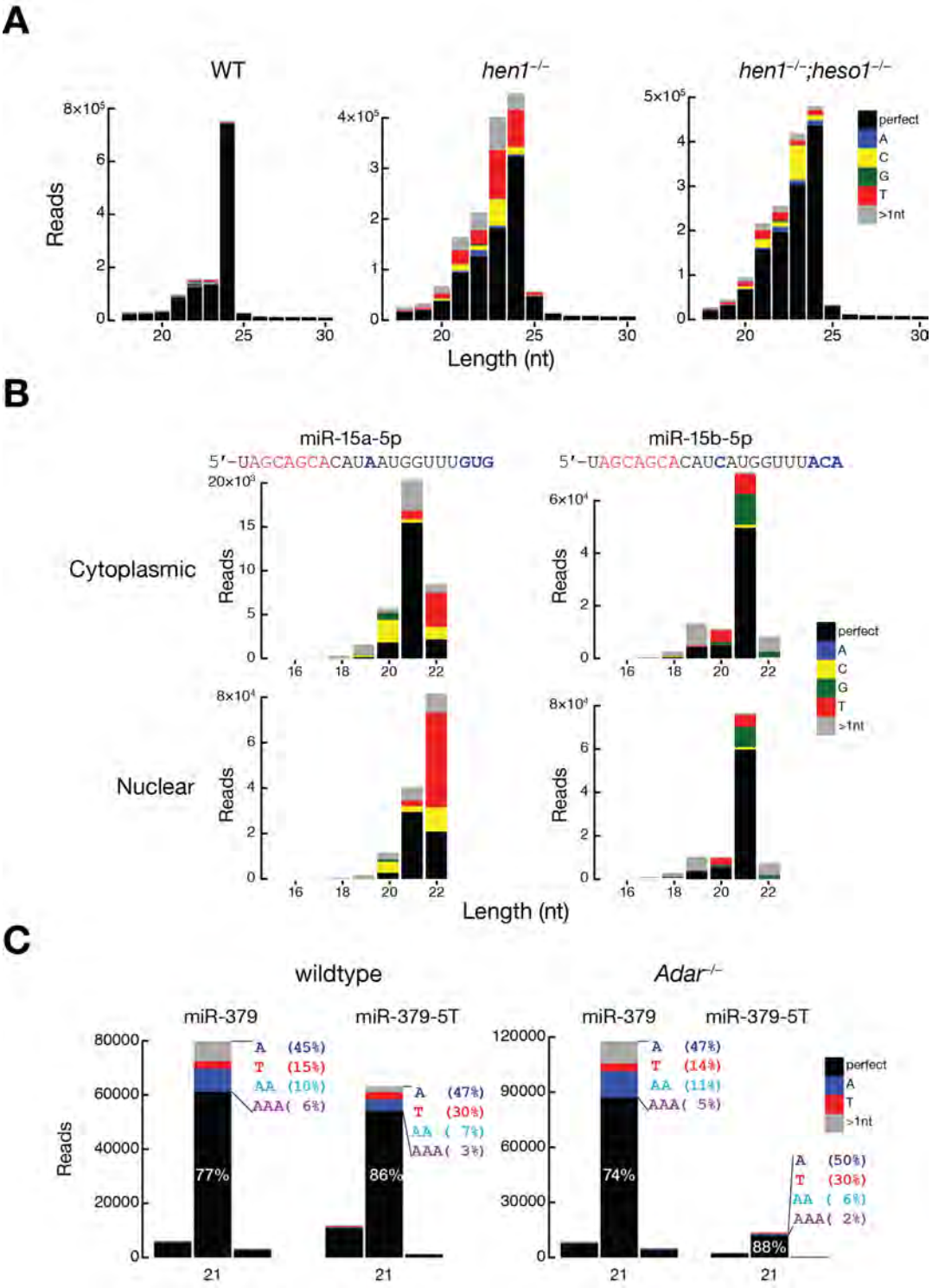

A
WT
hen1-/-
hen1-/-;heso1-/-
Reads
Length (nt)
perfect
A
C
G
T
>1nt
B
miR-15a-5p
5′-UAGCAGCACAUAAUGGUUUGUG
miR-15b-5p
5′-UAGCAGCACAUCAUGGUUUACA
Cytoplasmic
Nuclear
Reads
Length (nt)
C
wildtype
Adar-/-
miR-379
miR-379-5T
A (45%)
T (15%)
AA (10%)
AAA(6%)
77%
A (47%)
T (30%)
AA (7%)
AAA(3%)
86%
A (47%)
T (14%)
AA (11%)
AAA(5%)
74%
A (50%)
T (30%)
AA (6%)
AAA(2%)
88%
21

Figure Legend 5.5. Application of Tailor

(A) Length distribution of mRNA-derived small RNA reads with tailing information from wild type, *hen1* mutant, and *hen1*, *heso1* double mutant tissues from Arabidopsis. Raw read counts are shown without normalization. Perfect match and tailed reads are indicated in different colors.

(B) Length distribution of Ago2 associated Hsa-miR-15a (left) and Hsa-miR-15b (right) in cytoplasm (top) and nucleus (bottom) fraction of HeLa cell. Raw read count are shown without normalization. Note that since the authors of these libraries used poly-adenylation instead of 3′ ligation in their cloning strategy, it was impractical to identify A tailing.

(C) Tail composition for miR-379 and the edited form (miR-379-5G) in wildtype and *Adar*$^{-/-}$ libraries.

Finally, we applied Tailor to study the possible relationship between RNA editing and tailing in microRNAs. The miRNA libraries were constructed from the whole brain tissue cells dissected from *Adar2*$^{-/-}$ and wild-type mice (Vesely et al., 2014). Adar2 is known for its strongest effects on miRNA abundance and editing among the three isoforms of ADARs (Vesely et al., 2012). One of the highly expressed ADAR substrates, miR-379, was shown to be directly edited at the nucleotide 5 within the seed region, and about half of the mature miR-379 were edited by ADAR2 (Vesely et al., 2012). As expected, the edited form of miR-379 (i.e., miR-379-5G) was greatly reduced in *Adar*$^{-/-}$ mice. Surprisingly, we found that the normal miR-379 has much more tailing than miR-379-5G (Figure 5.5C). Mono-A and poly-A tails (the bluish portion) were depleted in miR-379-5G, which raises the probability that ADARs and the A-to-I editing could affect the affinity between the miRNAs and the unknown enzymes responsible for adenylylating the 3′ end. Since the proportion of different types of tails was unchanged upon *Adar2* knockout, the tailing machinery is less likely modulated by ADAR2 directly but by the subsequent factors after editing in the seed, such as differential targeting, RNA stability change or miRNA-Argonaute sorting.

Discussion

Tailing is a molecular phenomenon that associates with the function, processing, and stability of many small RNAs. Computational identification of the tailed sequences from the millions of NGS reads has been proven to be challenging and time-consuming. We herein present a tailing analysis framework, Tailor, which aligns reads to the reference genome, reports tailing events simultaneously, and visualizes analysis results. We assessed the accuracy of Tailor by comparing it with the *Chen* method with simulated reads and found they generated exactly the same results while Tailor only used a third of the time to align and provided more information comparing to the alternative.

When confounding factor was ignored, Tailor was not slower than other well-known fast general-purpose mappers in our tests. We demonstrated that Tailor executed in a speed that was very competitive to, if not better than, Bowtie and BWA, while providing more functionalities for detecting tailing events. When confounding factors was presented in the reads, it was arguable that advanced NGS aligners that support the local alignment mode (e.g., Bowtie2) could be competent in finding tails, but we tested them with simulated reads and showed that Tailor performed significantly better in both accuracy and efficiency.

Tailor's shell-based framework takes raw reads as input and produces comprehensive tailing analysis results and publication quality figures. We reproduced known conclusions drawn from the published tailing study by the pipeline with little extra scripting and post-processing. We also applied the

pipeline to other datasets and shed light on other possibilities of the functional roles of tailing, such as involving in RNA processing, transport, decay and storage by interacting with other RNA binding proteins (Gerstberger et al., 2014).

Our aims to design Tailor are to reduce the cost of doing tailing analysis and reinforce or even replace the conventional computational procedure in analyzing all short non-coding RNAs. We expect that Tailor could be applied to a broader scope and subsequently facilitate the understanding of biological processes related to tailing.

Acknowledgments

We thank the members of the Hung, Weng and Zamore laboratories for helpful discussion and critical testing.

Chapter VI Conclusions, discussion and future directions

Summary

miRNAs and piRNAs guide Argonaute proteins to regulate gene and transposon expression. Although enormous efforts have been made to dissect their pathways, many aspects in their biogenesis and function still remain elusive. My thesis research addresses unknown questions in the field by computationally exploring novel features of miRNA and piRNA sequences, biochemical identification of the enzymes, and validating the hypothesis with genetics and next generation sequencing strategies. The following sections in this thesis summarize my work and discuss the future challenges.

Nibbler and miRNAs

Using computational analysis and in vitro biochemical experiments, we discovered that many *Drosophila* miRNAs are released from pre-miRNA by Dicer-1 as intermediates that are longer than the lengths of mature small RNA associated with Ago proteins. Those longer isoforms are loaded into Ago1 protein as miRNA/miRNA* duplex. After the removal of miRNA*, the miRNA intermediates are subjected to 3′-to-5′ end trimming to their mature lengths. Performing a candidate RNAi screening, we identified a 3′-to-5′ exonuclease, Nibbler, that trims those intermediates to their mature lengths. Such trimming increases the diversity of miRNA sequences and explains the previously observed 3′ heterogeneity (Seitz et al., 2008).

Biochemical and structural studies suggest that the 3′ ends of small RNAs are bound by Argonaute PAZ domain and do not base-pair with their target RNAs (Tang et al., 2003; Haley and Zamore, 2004; Wee et al., 2012; Schirle et al., 2014). Consequently, the 3′ ends do not positively contribute to the efficiency of cleavage. Supporting this view, no direct evidence has linked the length of small RNA to their regulatory roles.

In fact, the study of miR-451 supports the view that miRNAs with longer 3′ ends still silence targets. Different from most miRNAs, the production of miR-451 is independent of Dicer. After being released by Drosha and exported into the cytoplasm, pre-miR-451 is loaded directly into Ago2. Using the 5′ arm as the guide strand, Ago2 cleaves the 3′ arm strand of pre-miR-451 and creates a miR-

451 intermediate with ~30 nucleotides (Cheloufi et al., 2010; Yang et al., 2010). Poly(A)-specific ribonuclease (PARN) trims the 3′ end of miR-451 intermediate to the mature length (Yoda et al., 2013). Surprisingly, the trimming activity is dispensable for the function of miR-451, indicating that RISC can tolerate miRNA guides as long as 30 nt.

The observation of tailed miRNAs in *Nbr* mutant encouraged us to link the miRNA 3′ end trimming to their stability. Tailing is observed when miRNAs encounter highly complementary targets (Ameres et al., 2010). It is hypothesized that structural rearrangements associated target binding provide the tailing enzyme access to miRNA 3′ ends, which would otherwise be buried in the PAZ domains (Yan et al., 2003; Lingel et al., 2003; Lingel et al., 2004a; Lingel et al., 2004b; Ameres and Zamore, 2013). Since longer isoforms of miRNAs are unlikely to have their 3′ ends accommodated in the PAZ domain, they are more vulnerable to Nibbler and the tailing enzyme. We thus conclude that 3′ end trimming of Nbr protects miRNA from tailing enzyme. Nonetheless, we cannot rule out the possibility that the tailed miRNA species are also substrates of Nbr.

Although tailing is often associated with target-directed miRNA degradation, the increase of tailing in the absence of Nbr fails to correlate with a decrease in their abundance (Figure 2.12D). Thus the specific function of Nbr-mediated 3′ end trimming remains elusive and demands research to understand tailing and degradation.

Nibbler and piRNAs

In Chapter III, we demonstrated that Nibbler also trims the 3′ end of piRNAs after the piRNA intermediates are generated by Zuc. In *Drosophila*, Zuc generates a majority of Piwi-associated piRNAs by processively slicing the cleavage products of Ago3. The nucleotides that are immediately downstream of the 3′ end of those piRNAs are enriched in uridines. Since those nucleotides do not exist in the mature piRNAs, we speculate that this preference is created by Zuc. We further propose that Zuc machinery chooses, as its cleavage site, the first uridine that is not protected by the PIWI proteins (a.k.a., the first uridine >26 nt from the 5′ end of piRNA intermediate). Consequently, some pre-piRNAs require 3′ trimming before their 3′ ends are methylated by Hen1, while some do not.

We can model the number of nucleotide that Nbr trims (a.k.a., the distance from the 3′ end of mature piRNA to the first uridine on its 3′ end) using a geometric distribution:

$$\sum_{i=2}^{\infty}\left[(i-1)\times\prod_{j=1}^{i-1}\left(1-p_j\right)\times p_i\right]$$

i: distance from the 3′ ends of piRNAs to the next uridine on the 3′ end (i.e., for ***i*** = 2, Nbr trims 1 nucleotide; Figure 6.1). ***p***: the percentage of uridine at the ***i*** nt position 3′ to the 3′ end of piRNA. ***p*** remains constant and equal to the percentage of uridines in piRNA clusters, assuming no preference for a uridine.

Our analysis suggests that in w^{1} and $aub^{HN2/QC42}$; $ago3^{t2/t3}$, uridines are enriched at the nucleotide immediately downstream of the 3′ end of piRNAs

(Figure 3.4B). On the other hand, *zuc*$^{HM27/Df}$ have a uridine composition equal to the percentage of uridine in the piRNA clusters. Based our simulation, it is estimated that Nbr trims ~0.7 nt on average (Figure 6.1), close to the average length increase (0.5 nt) in *Nbr*$^{-/-}$ compared to *w*^{1118} (Figure 3.14). Due to the small increase of piRNA length, it is unlikely that the 3′ end trimming plays an important function in the piRNA pathway.

Unexpectedly, in *w*^{1} and *aub*$^{HN2/QC42}$; *ago3*$^{t2/t3}$, but not *zuc*$^{HM27/Df}$, the chance to encounter the first uridine two nucleotides away from the 3′ end of piRNAs is lower than expected (Figure 6.1; ***i*** = 2). Our data suggest a possibility that the second nucleotide is also involved in the selection site of Zuc cleavage.

Despite the its role in the 3′ end trimming of miRNA and piRNA in flies, Nbr does not have homolog in silk moth or mouse, the two other model organisms widely used in piRNA studies. The 5′-to-5′ distance analysis of silk moth and mouse piRNA displayed a peak at ~35 nt, which is longer than the lengths of their mature piRNAs. These data suggest that their piRNAs undergo more 3′ trimming than do fly piRNAs. Consistently, *Tdrkh*$^{-\backslash-}$ mutant mice accumulate 30–36 nt piRNA intermediates. More importantly, *Tdrkh*$^{-\backslash-}$ male mice are sterile, indicating the 3′ trimming is indispensable in the piRNA pathway. The identification of the trimming enzyme in silk moth and mouse remain an important open question for the future study of the piRNA pathway.

Figure 6.1

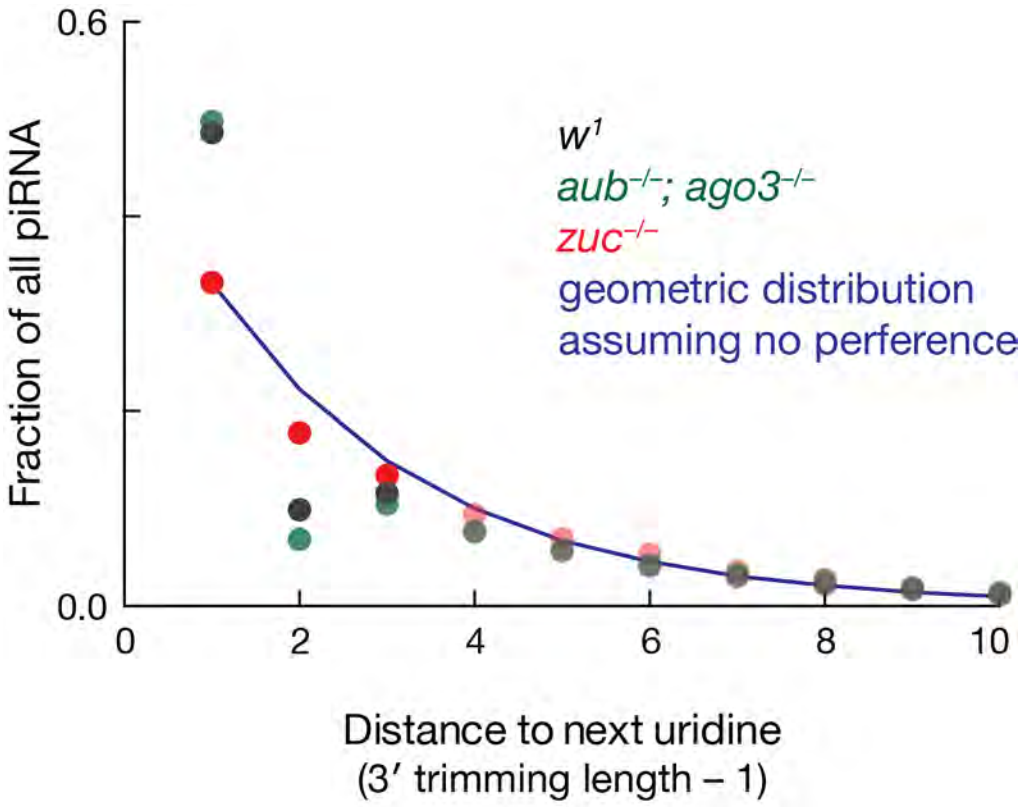

0.6
0.0
Fraction of all piRNA
0
2
4
6
8
10
Distance to next uridine
(3′ trimming length – 1)
w¹
aub⁻/⁻; ago3⁻/⁻
zuc⁻/⁻
geometric distribution
assuming no perference

Figure Legend 6.1. Modeling the Length of piRNA 3′ Trimming

Distance from the 3′ end of piRNAs to the next uridine in the genome. Only transposon-derived, unique mapping piRNAs from w^{1}, $aub^{HN2/QC42}$; $ago3^{t2/t3}$, $zuc^{HM27/Df}$ (dots) are included in this analysis. We used a geometric distribution, with a constant *p*, to model the expected distribution assuming no nucleotide preference towards uridine (line).

Ping-Pong Cycle and Transcriptional Silencing

In Chapter III, we analyzed piRNA sequences in different mutants and discovered that primary piRNAs display phasing—the 5′ ends of piRNAs exhibit a periodicity of ~26 nt in the genome. Similar to endogenous siRNAs generated by Dicer-2 (Vagin et al., 2006; Ghildiyal et al., 2008; Czech et al., 2008), phased piRNAs also rely on processive activity of the endonuclease Zucchini for their production. Further analyses on piRNAs associating with different PIWI proteins revealed that Piwi-piRNAs display the strongest phasing, Aub-piRNAs display modest phasing, while Ago3-piRNAs fail to show any signature of phasing. We subsequently observed that the abundance of Piwi-associated piRNA dropped dramatically in *ago3* and *vasa* mutants. Our data suggest that the production of primary piRNAs is downstream of the Ping-Pong cycle.

To test this hypothesis, we cloned the degradation intermediate RNAs using degradome sequencing and identified the potential cleavage products of Aub and Ago3 using the 10 bp cleavage signature. Further analysis of the 5′ end of Piwi-piRNA and those cleavage sites suggest that Piwi-piRNAs are initiated from the cleavage products of Aub and, more frequently, Ago3. Our data suggested a new model for piRNA biogenesis: the primary piRNAs are essentially produced from the cleavage product of the secondary piRNAs: after Ago3 slicing, the 3′ end cleavage products become a substrate of Zuc machinery, which preferentially slices the phosphodiester bond upstream of a uridine. This cleavage not only produces the 3′ end of the secondary piRNA, but also

generates a 5′ end of a piRNA precursor. This precursor is likely transferred to the outer membrane of mitochondria and further processed by Zuc to produce phased piRNAs that are mainly loaded into Piwi.

Without an intact Ping-Pong pathway, the level of germline Piwi-piRNAs drop while the somatic Piwi-piRNA level remains unchanged. Importantly, Piwi-piRNAs decrease to ~10–20% in *ago3* and *vas* mutant while only to ~40–60% in *aub* mutant. Those data indicate that most Piwi-associated piRNAs are generated from the cleavage products of Ago3 but not Aub. It is consistently with the early observation that Ago3 mainly associates with piRNAs in the sense orientation of the transposons while Aub- and Piwi-piRNAs are predominantly antisense.

Our data also explain the mysterious function of the Tudor protein Qin. In *qin* mutant ovaries, homotypic Aub:Aub dominates and heterotypic Aub:Ago3 Ping-Pong is greatly reduced (Zhang et al., 2011; Zhang et al., 2014a). However, the abundance of piRNAs and the strength of Ping-Pong show no significant difference. It remained elusive why Aub:Aub Ping-Pong cannot replace heterotypic Aub:Ago3 Ping-Pong. Our data suggest that Aub cleavage cannot produce antisense substrates for Zuc to generate Piwi-associated, antisense piRNAs. Thus, Qin ensures the antisense bias of Piwi-bound piRNAs by enforcing heterotypic Ago3:Aub Ping-Pong. However, how Qin ensures heterotypic Ping-Pong still remains mysterious. Qin could either inhibit homotypic Aub:Aub Ping-Pong or promote heterotypic Aub:Ago3 Ping-Pong. An epigenetic

strategy using *ago3*$^{t2/t3}$ and *ago3*$^{t2/t3}$, *qin*$^{1/Df}$ double mutant flies can answer this question. If Qin represses Aub:Aub Ping-Pong, then Ping-Pong will increase in *ago3*$^{t2/t3}$; *qin*$^{1/Df}$ compared to *ago3*$^{t2/t3}$. However, if Qin represses Aub:Ago3 Ping-Pong, then we predict that little change on the Ping-Pong level will be observed between *ago3*$^{t2/t3}$ single-mutant and *ago3*$^{t2/t3}$; *qin*$^{1/Df}$ double-mutant.

Our discovery that the Ping-Pong pathway not only amplifies piRNA reads bound to Aub and Ago3 but also produces piRNAs loaded into Piwi completely revises the current model of the piRNA pathway. It also raises many new questions. For example, two functions of the Ping-Pong cycle have been revealed: to repress transposons post-transcriptionally and to initiate the production of Piwi-bound piRNAs. Which one is its major function? To answer this question, we propose to compare the transposon silencing in *piwi*$^{2/Nt}$ single-mutant, *piwi*$^{2/Nt}$; *ago3*$^{t2/t3}$, and *piwi*$^{2/Nt}$; *aub*$^{HN2/QC42}$ double-mutants. If the major function of Ago3 is to generate Piwi-associated piRNAs and promote transcriptional silencing, we expect little transposon level increase when comparing *piwi*$^{2/Nt}$ single-mutant to *piwi*$^{2/Nt}$; *ago3*$^{t2/t3}$ double-mutant. On the other hand, if transposons further increase in *piwi*$^{2/Nt}$; *ago3*$^{t2/t3}$ compared to *piwi*$^{2/Nt}$, this portion of increase must be derived from the loss of heterotypic Aub:Ago3 Ping-Pong.

Aub silences transposon expression by cleaving transposon RNA and amplifying sense piRNA guides for Ago3, which can then initiate primary, antisense piRNA production for Piwi. *aub*; *piwi* double mutants would, of course,

be ideal for determining the relative importance of these two Aub functions, but the *aub* and *piwi* genes are too close to generate such a genotype (10 kbp and 0.04–0.06 centimorgan apart). The high efficiency of CRISPR might be useful in generating an *aub*; *piwi* double mutant fly line simultaneously. However, a *piwi* null mutant has degenerated ovaries due to its additional function in stem cell maintenance (Lin and Spradling, 1997; Cox et al., 1998; Klenov et al., 2011; Jin et al., 2013). Thus the $piwi^{Nt}$ allele needs to be included in our epistasis analysis, further increasing the difficulty of testing this hypothesis.

Despite those difficulties, the future is big for those small RNAs.

BIBLIOGRAPHY

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the *Arabidopsis* Degradome. Curr Biol *18*, 758-762.

Alkan, C., Kidd, J. M., Marques-Bonet, T., Aksay, G., Antonacci, F., Hormozdiari, F., Kitzman, J. O., Baker, C., Malig, M., Mutlu, O., Sahinalp, S. C., Gibbs, R. A., and Eichler, E. E. (2009). Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet *41*, 1061-1067.

Altschul, S., Demchak, B., Durbin, R., Gentleman, R., Krzywinski, M., Li, H., Nekrutenko, A., Robinson, J., Rasband, W., Taylor, J., and Trapnell, C. (2013). The anatomy of successful computational biology software. Nat Biotechnol *31*, 894-897.

Ameres, S. L., Horwich, M. D., Hung, J. H., Xu, J., Ghildiyal, M., Weng, Z., and Zamore, P. D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. Science *328*, 1534-1539.

Ameres, S. L., Hung, J. H., Xu, J., Weng, Z., and Zamore, P. D. (2011). Target RNA-directed tailing and trimming purifies the sorting of endo-siRNAs between the two Drosophila Argonaute proteins. RNA *17*, 54-63.

Ameres, S. L., and Zamore, P. D. (2013). Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol *14*, 475-488.

Ameyar-Zazoua, M., Rachez, C., Souidi, M., Robin, P., Fritsch, L., Young, R., Morozova, N., Fenouil, R., Descostes, N., Andrau, J. C., Mathieu, J., Hamiche,

A., Ait-Si-Ali, S., Muchardt, C., Batsché, E., and Harel-Bellan, A. (2012). Argonaute proteins couple chromatin silencing to alternative splicing. Nat Struct Mol Biol *19*, 998-1004.

Anders, S., Pyl, P. T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M. J., Kuramochi-Miyagawa, S., Nakano, T., Chien, M., Russo, J. J., Ju, J., Sheridan, R., Sander, C., Zavolan, M., and Tuschl, T. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203-207.

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., and Hannon, G. J. (2008). A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell *31*, 785-799.

Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744-747.

Aravin, A. A., van der Heijden, G. W., Castañeda, J., Vagin, V. V., Hannon, G. J., and Bortvin, A. (2009). Cytoplasmic compartmentalization of the fetal piRNA pathway in mice. PLoS Genet *5*, e1000764.

Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., and Gvozdev, V. A. (2001). Double-stranded RNA-mediated silencing of genomic tandem

repeats and transposable elements in the *D. melanogaster* germline. Curr Biol *11*, 1017-1027.

Arkov, A. L., Wang, J. Y., Ramos, A., and Lehmann, R. (2006). The role of Tudor domains in germline development and polar granule architecture. Development *133*, 4053-4062.

Baek, D., Villén, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Bagijn, M. P., Goldstein, L. D., Sapetschnig, A., Weick, E. M., Bouasker, S., Lehrbach, N. J., Simard, M. J., and Miska, E. A. (2012). Function, targets, and evolution of Caenorhabditis elegans piRNAs. Science *337*, 574-578.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215-233.

Bazzini, A. A., Lee, M. T., and Giraldez, A. J. (2012). Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. Science *336*, 233-237.

Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363-366.

Blow, M., Futreal, P. A., Wooster, R., and Stratton, M. R. (2004). A survey of RNA editing in human brain. Genome Res *14*, 2379-2387.

Blow, M. J., Grocock, R. J., van Dongen, S., Enright, A. J., Dicks, E., Futreal, P. A., Wooster, R., and Stratton, M. R. (2006). RNA editing of human microRNAs. Genome Biol *7*, R27.

Bohnsack, M. T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a Ran GTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA *10*, 185-191.

Bolcun-Filas, E., Bannister, L. A., Barash, A., Schimenti, K. J., Hartford, S. A., Eppig, J. J., Handel, M. A., Shen, L., and Schimenti, J. C. (2011). A-MYB (MYBL1) transcription factor is a master regulator of male meiosis. Development *138*, 3319-3330.

Borchert, G. M., Lanier, W., and Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

Boswell, R. E., and Mahowald, A. P. (1985). tudor, a gene required for assembly of the germ plasm in Drosophila melanogaster. Cell *43*, 97-104.

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

Brennecke, J., Malone, C. D., Aravin, A. A., Sachidanandam, R., Stark, A., and Hannon, G. J. (2008). An epigenetic role for maternally inherited piRNAs in transposon silencing. Science *322*, 1387-1392.

Burkhardt, S., and Kärkkäinen, J. (2003). Fast lightweight suffix array construction and checking. *Combinatorial Pattern Matching*, 55-69.

Burrows, M., and Wheeler, D. J. (1994). A block-sorting lossless data compression algorithm. Technical Report *123*, Digital Equipment Corporation.

Cai, X., Hagedorn, C. H., and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. RNA *10*, 1957-1966.

Carmell, M. A., Girard, A., van de Kant, H. J., Bourc'his, D., Bestor, T. H., de Rooij, D. G., and Hannon, G. J. (2007). MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell *12*, 503-514.

Carmell, M. A., Xuan, Z., Zhang, M. Q., and Hannon, G. J. (2002). The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. Genes Dev *16*, 2733-2742.

Cenik, E. S., Fukunaga, R., Lu, G., Dutcher, R., Wang, Y., Tanaka Hall, T. M., and Zamore, P. D. (2011). Phosphate and R2D2 Restrict the Substrate Specificity of Dicer-2, an ATP-Driven Ribonuclease. Mol Cell *42*, 172-184.

Cenik, E. S., and Zamore, P. D. (2011). Argonaute proteins. Curr Biol *21*, R446-R449.

Chatterjee, S., and Grosshans, H. (2009). Active turnover modulates mature microRNA activity in Caenorhabditis elegans. Nature *461*, 546-549.

Cheloufi, S., Dos Santos, C. O., Chong, M. M., and Hannon, G. J. (2010). A dicer-independent miRNA biogenesis pathway that requires Ago catalysis. Nature *465*, 584-589.

Chen, C., Jin, J., James, D. A., Adams-Cioaba, M. A., Park, J. G., Guo, Y., Tenaglia, E., Xu, C., Gish, G., Min, J., and Pawson, T. (2009a). Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. Proc Natl Acad Sci U S A *106*, 20336-20341.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L., and Mardis, E. R. (2009b). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods *6*, 677-681.

Chuma, S., Hosokawa, M., Kitamura, K., Kasai, S., Fujioka, M., Hiyoshi, M., Takamune, K., Noce, T., and Nakatsuji, N. (2006). Tdrd1/Mtr-1, a tudor-related gene, is essential for male germ-cell differentiation and nuage/germinal granule formation in mice. Proc Natl Acad Sci U S A *103*, 15894-15899.

Cox, D. N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by *piwi* are essential for stem cell self-renewal. Genes Dev *12*, 3715-3727.

Czech, B., and Hannon, G. J. (2011). Small RNA sorting: matchmaking for Argonautes. Nat Rev Genet *12*, 19-31.

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., and Brennecke, J. (2008). An endogenous small interfering RNA pathway in *Drosophila*. Nature *453*, 798-802.

Czech, B., Preall, J. B., McGinn, J., and Hannon, G. J. (2013). A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway. Mol Cell *50*, 749-761.

Czech, B., Zhou, R., Erlich, Y., Brennecke, J., Binari, R., Villalta, C., Gordon, A., Perrimon, N., and Hannon, G. J. (2009). Hierarchical rules for Argonaute loading in *Drosophila*. Mol Cell *36*, 445-456.

Darricarrère, N., Liu, N., Watanabe, T., and Lin, H. (2013). Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. Proc Natl Acad Sci U S A *110*, 1297-1302.

Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat Commun *3*, 945.

De Fazio, S., Bartonicek, N., Di Giacomo, M., Abreu-Goodger, C., Sankar, A., Funaya, C., Antony, C., Moreira, P. N., Enright, A. J., and O'Carroll, D. (2011). The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. Nature *480*, 259-263.

de Vanssay, A., Bougé, A. L., Boivin, A., Hermant, C., Teysset, L., Delmarre, V., Antoniewski, C., and Ronsseray, S. (2012). Paramutation in Drosophila linked to emergence of a piRNA-producing locus. Nature *490*, 112-115.

Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell *2*, 819-830.

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231-235.

Dietzl, G., Chen, D., Schnorrer, F., Su, K. C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., Couto, A., Marra, V., Keleman, K., and Dickson, B. J. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. Nature *448*, 151-156.

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

Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res *36*, e105.

Farh, K. K., Grimson, A., Jan, C., Lewis, B. P., Johnston, W. K., Lim, L. P., Burge, C. B., and Bartel, D. P. (2005). The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science *310*, 1817-1821.

Feltzin, V. L., Khaladkar, M., Abe, M., Parisi, M., Hendriks, G. J., Kim, J., and Bonini, N. M. (2015). The exonuclease Nibbler regulates age-associated traits and modulates piRNA length in Drosophila. Aging Cell

Ferragina, P., and Manzini, G. (2000). Opportunistic data structures with applications. *Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on*, 390-398.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. Nature *391*, 806-811.

Forstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E., and Zamore, P. D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. PLoS Biol *3*, e236.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2009). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res *19*, 92-105.

Frost, R. J., Hamra, F. K., Richardson, J. A., Qi, X., Bassel-Duby, R., and Olson, E. N. (2010). MOV10L1 is necessary for protection of spermatocytes against retrotransposons by Piwi-interacting RNAs. Proc Natl Acad Sci U S A *107*, 11847-11852.

German, M. A., Pillay, M., Jeong, D. H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol *26*, 941-946.

Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat Rev Genet *15*, 829-845.

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

Ghildiyal, M., Xu, J., Seitz, H., Weng, Z., and Zamore, P. D. (2010). Sorting of Drosophila small silencing RNAs partitions microRNA* strands into the RNA interference pathway. RNA *16*, 43-56.

Ghildiyal, M., and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. Nat Rev Genet *10*, 94-108.

Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199-202.

Gonzalez-Reyes, A., Elliott, H., and St Johnston, D. (1997). Oocyte determination and the origin of polarity in Drosophila: the role of the spindle genes. Development *124*, 4927-4937.

Goriaux, C., Desset, S., Renaud, Y., Vaury, C., and Brasset, E. (2014). Transcriptional properties and splicing of the flamenco piRNA cluster. EMBO Rep *15*, 411-418.

Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235-240.

Griffiths-Jones, S. (2010). miRBase: microRNA sequences and annotation. Curr Protoc Bioinformatics *Chapter 12*, Unit 12.9.1-Unit 12.910.

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res *34*, D140-D144.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. Nucleic Acids Res *36*, D154-D158.

Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell *27*, 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., and Bartel, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G., and Mello, C. C. (2001). Genes and Mechanisms Related to RNA Interference Regulate Expression of the Small Temporal RNAs that Control *C. elegans* Developmental Timing. Cell *106*, 23-34.

Grivna, S. T., Beyret, E., Wang, Z., and Lin, H. (2006a). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev *20*, 1709-1714.

Grivna, S. T., Pyhtila, B., and Lin, H. (2006b). MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. Proc Natl Acad Sci U S A *103*, 13415-13420.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M. C. (2007). A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5′ End Formation *in Drosophi*la. Science *315*, 1587-1590.

Guo, H., Ingolia, N. T., Weissman, J. S., and Bartel, D. P. (2010). Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature *466*, 835-840.

Haase, A. D., Fenoglio, S., Muerdter, F., Guzzardo, P. M., Czech, B., Pappin, D. J., Chen, C., Gordon, A., and Hannon, G. J. (2010). Probing the initiation and effector phases of the somatic piRNA pathway in *Drosophila*. Genes Dev *24*, 2499-2504.

Haase, A. D., Jaskiewicz, L., Zhang, H., Laine, S., Sack, R., Gatignol, A., and Filipowicz, W. (2005). TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. EMBO Rep *6*, 961-967.

Hadley, W. (2009). ggplot2: Elegant graphics for data analysis.

Haley, B., Tang, G., and Zamore, P. D. (2003). In vitro analysis of RNA interference in Drosophila melanogaster. Methods *30*, 330-336.

Haley, B., and Zamore, P. D. (2004). Kinetic analysis of the RNAi enzyme complex. Nat Struct Mol Biol *11*, 599-606.

Han, B. W., Wang, W., Li, C., Weng, Z., and Zamore, P. D. (2015a). Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science *348*, 817-821.

Han, B. W., Wang, W., Zamore, P. D., and Weng, Z. (2015b). piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics *31*, 593-595.

Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev *18*, 3016-3027.

Handler, D., Meixner, K., Pizka, M., Lauss, K., Schmied, C., Gruber, F. S., and Brennecke, J. (2013). The genetic makeup of the Drosophila piRNA pathway. Mol Cell *50*, 762-777.

Handler, D., Olivieri, D., Novatchkova, M., Gruber, F. S., Meixner, K., Mechtler, K., Stark, A., Sachidanandam, R., and Brennecke, J. (2011). A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. EMBO J *30*, 3977-3993.

Heler, R., Marraffini, L. A., and Bikard, D. (2014). Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. Mol Microbiol *93*, 1-9.

Heo, I., Ha, M., Lim, J., Yoon, M. J., Park, J. E., Kwon, S. C., Chang, H., and Kim, V. N. (2012). Mono-Uridylation of Pre-MicroRNA as a Key Step in the Biogenesis of Group II let-7 MicroRNAs. Cell *151*, 521-532.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., and Kim, V. N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. Mol Cell *32*, 276-284.

Heo, I., Joo, C., Kim, Y. K., Ha, M., Yoon, M. J., Cho, J., Yeom, K. H., Han, J., and Kim, V. N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. Cell *138*, 696-708.

Honda, S., Kirino, Y., Maragkakis, M., Alexiou, P., Ohtaki, A., Murali, R., Mourelatos, Z., and Kirino, Y. (2013). Mitochondrial protein BmPAPI modulates the length of mature piRNAs. RNA *19*, 1405-1418.

Hormozdiari, F., Hajirasouliha, I., Dao, P., Hach, F., Yorukoglu, D., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2010). Next-generation VariationHunter: combinatorial algorithms for transposon insertion discovery. Bioinformatics *26*, i350-i357.

Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P. D. (2007). The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. Curr Biol *17*, 1265-1272.

Hosokawa, M., Shoji, M., Kitamura, K., Tanaka, T., Noce, T., Chuma, S., and Nakatsuji, N. (2007). Tudor-related proteins TDRD1/MTR-1, TDRD6 and

TDRD7/TRAP: domain composition, intracellular localization, and function in male germ cells in mice. Dev Biol *301*, 38-52.

Huang, H. Y., Houwing, S., Kaaij, L. J., Meppelink, A., Redl, S., Gauci, S., Vos, H., Draper, B. W., Moens, C. B., Burgering, B. M., Ladurner, P., Krijgsveld, J., Berezikov, E., and Ketting, R. F. (2011). Tdrd1 acts as a molecular scaffold for Piwi proteins and piRNA targets in zebrafish. EMBO J *30*, 3298-3308.

Huang, X. A., Yin, H., Sweeney, S., Raha, D., Snyder, M., and Lin, H. (2013). A Major Epigenetic Programming Mechanism Guided by piRNAs. Dev Cell *24*, 502-516.

Hutvagner, G., and Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. Nat Rev Mol Cell Biol *9*, 22-32.

Hutvagner, G. (2001). A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the let-7 Small Temporal RNA. Science *293*, 834-838.

Ipsaro, J. J., Haase, A. D., Knott, S. R., Joshua-Tor, L., and Hannon, G. J. (2012). The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis. Nature *491*, 279-283.

Iwasaki, S., Kobayashi, M., Yoda, M., Sakaguchi, Y., Katsuma, S., Suzuki, T., and Tomari, Y. (2010). Hsc70/Hsp90 chaperone machinery mediates ATP-dependent RISC loading of small RNA duplexes. Mol Cell *39*, 292-299.

Ji, L., and Chen, X. (2012). Regulation of small RNA stability: methylation and beyond. Cell Res *22*, 624-636.

Jiang, F., Ye, X., Liu, X., Fincher, L., McKearin, D., and Liu, Q. (2005). Dicer-1 and R3D1-L catalyze microRNA maturation in *Drosophila*. Genes Dev *19*, 1674-1679.

Jin, Z., Flynt, A. S., and Lai, E. C. (2013). Drosophila piwi Mutants Exhibit Germline Stem Cell Tumors that Are Sustained by Elevated Dpp Signaling. Curr Biol *23*, 1442-1448.

Kawaoka, S., Izumi, N., Katsuma, S., and Tomari, Y. (2011). 3' end formation of PIWI-interacting RNAs in vitro. Mol Cell *43*, 1015-1022.

Keane, T. M., Wong, K., and Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics *29*, 389-390.

Kennedy, S., Wang, D., and Ruvkun, G. (2004). A conserved siRNA-degrading RNase negatively regulates RNA interference in *C. elegans*. Nature *427*, 645-649.

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

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., and Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. Genes Dev *15*, 2654-2659.

Ketting, R. F., Haverkamp, T. H., van Luenen, H. G., and Plasterk, R. H. (1999). Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. Cell *99*, 133-141.

Khurana, J. S., Wang, J., Xu, J., Koppetsch, B. S., Thomson, T. C., Nowosielska, A., Li, C., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2011). Adaptation to P element transposon invasion in Drosophila melanogaster. Cell *147*, 1551-1563.

Kim, D. D., Kim, T. T., Walsh, T., Kobayashi, Y., Matise, T. C., Buyske, S., and Gabriel, A. (2004). Widespread RNA editing of embedded alu elements in the human transcriptome. Genome Res *14*, 1719-1725.

Kirino, Y., Kim, N., de Planell-Saguer, M., Khandros, E., Chiorean, S., Klein, P. S., Rigoutsos, I., Jongens, T. A., and Mourelatos, Z. (2009). Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. Nat Cell Biol *11*, 652-658.

Klenov, M. S., Lavrov, S. A., Korbut, A. P., Stolyarenko, A. D., Yakushev, E. Y., Reuter, M., Pillai, R. S., and Gvozdev, V. A. (2014). Impact of nuclear Piwi elimination on chromatin state in Drosophila melanogaster ovaries. Nucleic Acids Res *42*, 6208-6218.

Klenov, M. S., Sokolova, O. A., Yakushev, E. Y., Stolyarenko, A. D., Mikhaleva, E. A., Lavrov, S. A., and Gvozdev, V. A. (2011). Separation of stem cell maintenance and transposon silencing functions of Piwi protein. Proc Natl Acad Sci U S A *108*, 18760-18765.

Kozomara, A., and Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res *39*, D152-D157.

Krutzfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M., and Stoffel, M. (2005). Silencing of microRNAs in vivo with ‘antagomirs’. Nature *438*, 685-689.

Kume, H., Hino, K., Galipon, J., and Ui-Tei, K. (2014). A-to-I editing in the miRNA seed region regulates target mRNA selection and silencing efficiency. Nucleic Acids Res *42*, 10050-10060.

Kuramochi-Miyagawa, S., Kimura, T., Yomogida, K., Kuroiwa, A., Tadokoro, Y., Fujita, Y., Sato, M., Matsuda, Y., and Nakano, T. (2001). Two mouse piwi-related genes: miwi and mili. Mech Dev *108*, 121-133.

Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T. W., Hata, K., Li, E., Matsuda, Y., Kimura, T., Okabe, M., Sakaki, Y., Sasaki, H., and Nakano, T. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev *22*, 908-917.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell *121*, 713-724.

Lai, E. C. (2002). MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet *30*, 363-364.

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. Curr Biol *14*, 2162-2167.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.

Lau, N. C., Robine, N., Martin, R., Chung, W. J., Niki, Y., Berezikov, E., and Lai, E. C. (2009). Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. Genome Res *19*, 1776-1785.

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

Le Thomas, A., Stuwe, E., Li, S., Du, J., Marinov, G., Rozhkov, N., Chen, Y. C., Luo, Y., Sachidanandam, R., Toth, K. F., Patel, D., and Aravin, A. A. (2014). Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. Genes Dev *28*, 1667-1680.

Le, H. S., Schulz, M. H., McCauley, B. M., Hinman, V. F., and Bar-Joseph, Z. (2013). Probabilistic error correction for RNA sequencing. Nucleic Acids Res *41*, e109.

Lee, H. C., Gu, W., Shirayama, M., Youngman, E., Conte, D., and Mello, C. C. (2012). *C. elegans* piRNAs Mediate the Genome-wide Surveillance of Germline Transcripts. Cell *150*, 78-87.

Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993). The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell *75*, 843-854.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415-419.

Lee, Y., Hur, I., Park, S. Y., Kim, Y. K., Suh, M. R., and Kim, V. N. (2006). The role of PACT in the RNA silencing pathway. EMBO J *25*, 522-532.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J *23*, 4051-4060.

Li, C., Vagin, V. V., Lee, S., Xu, J., Ma, S., Xi, H., Seitz, H., Horwich, M. D., Syrzycka, M., Honda, B. M., Kittler, E. L., Zapp, M. L., Klattenhoff, C., Schulz, N., Theurkauf, W. E., Weng, Z., and Zamore, P. D. (2009). Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals Somatic piRNAs in Flies. Cell *137*, 509-521.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754-1760.

Li, H., Ruan, J., and Durbin, R. (2008a). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res *18*, 1851-1858.

Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. Curr Biol *15*, 1501-1507.

Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008b). SOAP: short oligonucleotide alignment program. Bioinformatics *24*, 713-714.

Li, R., Yu, C., Li, Y., Lam, T. W., Yiu, S. M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966-1967.

Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., Xu, J., Moore, M. J., Schimenti, J. C., Weng, Z., and Zamore, P. D. (2013). An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol Cell *50*, 67-81.

Lim, A. K., and Kai, T. (2007). Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster*. Proc Natl Acad Sci U S A *104*, 6714-6719.

Lin, H., Chen, M., Kundaje, A., Valouev, A., Yin, H., Liu, N., Neuenkirchen, N., Zhong, M., and Snyder, M. (2015). Reassessment of Piwi Binding to the Genome and Piwi Impact on RNA Polymerase II Distribution. Dev Cell *32*, 772-774.

Lin, H., and Spradling, A. C. (1997). A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the *Drosophila* ovary. Development *124*, 2463-2476.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain. Nature *426*, 465-469.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2004a). NMR assignment of the Drosophila Argonaute2 PAZ domain. J Biomol NMR *29*, 421-422.

Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2004b). Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. Nat Struct Mol Biol *11*, 576-577.

Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J. J., Hammond, S. M., Joshua-Tor, L., and Hannon, G. J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

Liu, Y., and Schmidt, B. (2012). Long read alignment based on maximal exact match seeds. Bioinformatics *28*, i318-i324.

Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of *Scarecrow-Like* mRNA Targets Directed by a Class of *Arabidopsis* miRNA. Science *297*, 2053-2056.

Luciano, D. J., Mirsky, H., Vendetti, N. J., and Maas, S. (2004). RNA editing of a miRNA precursor. RNA *10*, 1174-1177.

Luteijn, M. J., and Ketting, R. F. (2013). PIWI-interacting RNAs: from generation to transgenerational epigenetics. Nat Rev Genet *14*, 523-534.

Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., Sachidanandam, R., and Hannon, G. J. (2009). Specialized piRNA Pathways Act in Germline and Somatic Tissues of the *Drosophila* Ovary. Cell *137*, 522-535.

Marco, A., and Griffiths-Jones, S. (2012). Detection of microRNAs in color space. Bioinformatics *28*, 318-323.

Marinov, G. K., Wang, J., Handler, D., Wold, B. J., Weng, Z., Hannon, G. J., Aravin, A. A., Zamore, P. D., Brennecke, J., and Toth, K. F. (2015). Pitfalls of Mapping High-Throughput Sequencing Data to Repetitive Sequences: Piwi's Genomic Targets Still Not Identified. Dev Cell *32*, 765-771.

Mathioudakis, N., Palencia, A., Kadlec, J., Round, A., Tripsianes, K., Sattler, M., Pillai, R. S., and Cusack, S. (2012). The multiple Tudor domain-containing protein TDRD1 is a molecular scaffold for mouse Piwi proteins and piRNA biogenesis factors. RNA *18*, 2056-2072.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. Nat Rev Genet *14*, 447-459.

Minoche, A. E., Dohm, J. C., and Himmelbauer, H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. Genome Biol *12*, R112.

Mohn, F., Sienski, G., Handler, D., and Brennecke, J. (2014). The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. Cell *157*, 1364-1379.

Morse, D. P., Aruscavage, P. J., and Bass, B. L. (2002). RNA hairpins in noncoding regions of human brain and Caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA. Proc Natl Acad Sci U S A *99*, 7906-7911.

Napoli, C., Lemieux, C., and Jorgensen, R. (1990). Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. Plant Cell *2*, 279-289.

Nishida, K. M., Okada, T. N., Kawamura, T., Mituyama, T., Kawamura, Y., Inagaki, S., Huang, H., Chen, D., Kodama, T., Siomi, H., and Siomi, M. C. (2009). Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in Drosophila germlines. EMBO J *28*, 3820-3831.

Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M. K., Bonnefond, L., Matsumoto, N., Nishizawa, T., Nakanaga, K., Aoki, J., Ishitani, R., Siomi, H., Siomi, M. C., and Nureki, O. (2012). Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature *491*, 284-287.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. (2007). The Mirtron Pathway Generates microRNA-Class Regulatory RNAs in *Drosophila*. Cell *130*, 89-100.

Olivieri, D., Sykora, M. M., Sachidanandam, R., Mechtler, K., and Brennecke, J. (2010). An in vivo RNAi assay identifies major genetic and cellular requirements for primary piRNA biogenesis in Drosophila. EMBO J *29*, 3301-3317.

Orban, T. I., and Izaurralde, E. (2005). Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. RNA *11*, 459-469.

Pal-Bhadra, M., Bhadra, U., and Birchler, J. A. (1997). Cosuppression in Drosophila: gene silencing of Alcohol dehydrogenase by white-Adh transgenes is Polycomb dependent. Cell *90*, 479-490.

Pall, G. S., and Hamilton, A. J. (2008). Improved northern blot method for enhanced detection of small RNA. Nat Protoc *3*, 1077-1084.

Pan, J., Goodheart, M., Chuma, S., Nakatsuji, N., Page, D. C., and Wang, P. J. (2005). RNF17, a component of the mammalian germ cell nuage, is essential for spermiogenesis. Development *132*, 4029-4039.

Pandey, R. R., Tokuzawa, Y., Yang, Z., Hayashi, E., Ichisaka, T., Kajita, S., Asano, Y., Kunieda, T., Sachidanandam, R., Chuma, S., Yamanaka, S., and Pillai, R. S. (2013). Tudor domain containing 12 (TDRD12) is essential for secondary PIWI interacting RNA biogenesis in mice. Proc Natl Acad Sci U S A *110*, 16492-16497.

Pane, A., Wehr, K., and Schüpbach, T. (2007). zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Dev Cell *12*, 851-862.

Parker, J. S., Parizotto, E. A., Wang, M., Roe, S. M., and Barford, D. (2009). Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. Mol Cell *33*, 204-214.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. Nature *408*, 86-89.

Patil, V. S., Anand, A., Chakrabarti, A., and Kai, T. (2014). The Tudor domain protein Tapas, a homolog of the vertebrate Tdrd7, functions in the piRNA pathway to regulate retrotransposons in germline of Drosophila melanogaster. BMC Biol *12*, 61.

Patil, V. S., and Kai, T. (2010). Repression of retroelements in Drosophila germline via piRNA pathway by the Tudor domain protein Tejas. Curr Biol *20*, 724-730.

Pelisson, A., Song, S. U., Prud'homme, N., Smith, P. A., Bucheton, A., and Corces, V. G. (1994). Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the *Drosophila* flamenco gene. EMBO J *13*, 4401-4411.

Pohl, A., and Beato, M. (2014). bwtool: a tool for bigWig files. Bioinformatics *30*, 1618-1619.

Prud'homme, N., Gans, M., Masson, M., Terzian, C., and Bucheton, A. (1995). Flamenco, a gene controlling the gypsy retrovirus of *Drosophila melanogaster*. Genetics *139*, 697-711.

Qu, W., Hashimoto, S., and Morishita, S. (2009). Efficient frequency-based de novo short-read clustering for error trimming in next-generation sequencing. Genome Res *19*, 1309-1315.

Quinlan, A. R., and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841-842.

Ramachandran, V., and Chen, X. (2008). Degradation of microRNAs by a family of exoribonucleases in *Arabidopsis*. Science *321*, 1490-1492.

Ren, G., Chen, X., and Yu, B. (2012). Uridylation of miRNAs by hen1 suppressor1 in *Arabidopsis*. Curr Biol *22*, 695-700.

Ren, G., Xie, M., Zhang, S., Vinovskis, C., Chen, X., and Yu, B. (2014). Methylation protects microRNAs from an AGO1-associated activity that uridylates 5’ RNA fragments generated by AGO1 cleavage. Proc Natl Acad Sci U S A *111*, 6365-6370.

Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C., Sachidanandam, R., and Pillai, R. S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature *480*, 264-267.

Robert, V., Prud’homme, N., Kim, A., Bucheton, A., and Pelisson, A. (2001). Characterization of the flamenco region of the *Drosophila melanogaster* genome. Genetics *158*, 701-713.

Roberts, A., and Pachter, L. (2013). Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods *10*, 71-73.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Romano, N., and Macino, G. (1992). Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences. Mol Microbiol *6*, 3343-3353.

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

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell *127*, 1193-1207.

Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Saito, K., Inagaki, S., Mituyama, T., Kawamura, Y., Ono, Y., Sakota, E., Kotani, H., Asai, K., Siomi, H., and Siomi, M. C. (2009). A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature *461*, 1296-1299.

Saito, K., Nishida, K. M., Mori, T., Kawamura, Y., Miyoshi, K., Nagami, T., Siomi, H., and 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.

Saito, K., Sakaguchi, Y., Suzuki, T., Suzuki, T., Siomi, H., and Siomi, M. C. (2007). Pimet, the *Drosophila* homolog of HEN1, mediates *2′*-O-methylation of Piwi- interacting RNAs at their 3′ ends. Genes Dev *21*, 1603-1608.

Saxe, J. P., Chen, M., Zhao, H., and Lin, H. (2013). Tdrkh is essential for spermatogenesis and participates in primary piRNA biogenesis in the germline. EMBO J *32*, 1869-1885.

Schirle, N. T., and MacRae, I. J. (2012). The crystal structure of human Argonaute2. Science *336*, 1037-1040.

Schirle, N. T., Sheu-Gruttadauria, J., and MacRae, I. J. (2014). Structural basis for microRNA targeting. Science *346*, 608-613.

Seitz, H., Ghildiyal, M., and Zamore, P. D. (2008). Argonaute loading improves the 5’ precision of both microRNAs and their miRNA* strands in flies. Curr Biol *18*, 147-151.

Shirayama, M., Seth, M., Lee, H. C., Gu, W., Ishidate, T., Conte, D., and Mello, C. C. (2012). piRNAs Initiate an Epigenetic Memory of Nonself RNA in the *C. elegans* Germline. Cell *150*, 65-77.

Shoji, M., Tanaka, T., Hosokawa, M., Reuter, M., Stark, A., Kato, Y., Kondoh, G., Okawa, K., Chujo, T., Suzuki, T., Hata, K., Martin, S. L., Noce, T., Kuramochi-Miyagawa, S., Nakano, T., Sasaki, H., Pillai, R. S., Nakatsuji, N., and Chuma, S. (2009). The TDRD9-MIWI2 complex is essential for piRNA-mediated retrotransposon silencing in the mouse male germline. Dev Cell *17*, 775-787.

Sienski, G., Dönertas, D., and 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., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. Nat Rev Mol Cell Biol *12*, 246-258.

Song, J. J., Smith, S. K., Hannon, G. J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. Science *305*, 1434-1437.

Stark, A., Brennecke, J., Bushati, N., Russell, R. B., and Cohen, S. M. (2005). Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. Cell *123*, 1133-1146.

Stefani, G., and Slack, F. J. (2008). Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol *9*, 219-230.

Swarts, D. C., Mosterd, C., van Passel, M. W., and Brouns, S. J. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS One *7*, e35888.

Szakmary, A., Reedy, M., Qi, H., and Lin, H. (2009). The Yb protein defines a novel organelle and regulates male germline stem cell self-renewal in *Drosophila* melanogaster. J Cell Biol *185*, 613-627.

Tanaka, T., Hosokawa, M., Vagin, V. V., Reuter, M., Hayashi, E., Mochizuki, A. L., Kitamura, K., Yamanaka, H., Kondoh, G., Okawa, K., Kuramochi-Miyagawa, S., Nakano, T., Sachidanandam, R., Hannon, G. J., Pillai, R. S., Nakatsuji, N., and Chuma, S. (2011). Tudor domain containing 7 (Tdrd7) is essential for

dynamic ribonucleoprotein (RNP) remodeling of chromatoid bodies during spermatogenesis. Proc Natl Acad Sci U S A *108*, 10579-10584.

Tang, G., Reinhart, B. J., Bartel, D. P., and Zamore, P. D. (2003). A biochemical framework for RNA silencing in plants. Genes Dev *17*, 49-63.

Tomari, Y., Du, T., and Zamore, P. D. (2007). Sorting of Drosophila small silencing RNAs. Cell *130*, 299-308.

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., and Pachter, L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol *31*, 46-53.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

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

Vagin, V. V., Yu, Y., Jankowska, A., Luo, Y., Wasik, K. A., Malone, C. D., Harrison, E., Rosebrock, A., Wakimoto, B. T., Fagegaltier, D., Muerdter, F., and Hannon, G. J. (2013). Minotaur is critical for primary piRNA biogenesis. RNA *19*, 1064-1077.

van der Heijden, G. W., and Bortvin, A. (2009). Defending the genome in tudor style. Dev Cell *17*, 745-746.

van der Krol, A. R., Mur, L. A., Beld, M., Mol, J. N., and Stuitje, A. R. (1990). Flavonoid genes in petunia: addition of a limited number of gene copies may lead to a suppression of gene expression. Plant Cell *2*, 291-299.

van Wolfswinkel, J. C., Claycomb, J. M., Batista, P. J., Mello, C. C., Berezikov, E., and Ketting, R. F. (2009). CDE-1 affects chromosome segregation through uridylation of CSR-1-bound siRNAs. Cell *139*, 135-148.

Vasileva, A., Tiedau, D., Firooznia, A., Müller-Reichert, T., and Jessberger, R. (2009). Tdrd6 Is Required for Spermiogenesis, Chromatoid Body Architecture, and Regulation of miRNA Expression. Curr Biol *19*, 630-639.

Vesely, C., Tauber, S., Sedlazeck, F. J., Tajaddod, M., von Haeseler, A., and Jantsch, M. F. (2014). ADAR2 induces reproducible changes in sequence and abundance of mature microRNAs in the mouse brain. Nucleic Acids Res *42*, 12155-12168.

Vesely, C., Tauber, S., Sedlazeck, F. J., von Haeseler, A., and Jantsch, M. F. (2012). Adenosine deaminases that act on RNA induce reproducible changes in abundance and sequence of embryonic miRNAs. Genome Res *22*, 1468-1476.

Voigt, F., Reuter, M., Kasaruho, A., Schulz, E. C., Pillai, R. S., and Barabas, O. (2012). Crystal structure of the primary piRNA biogenesis factor Zucchini reveals similarity to the bacterial PLD endonuclease Nuc. RNA *18*, 2128-2134.

Vourekas, A., Zheng, K., Fu, Q., Maragkakis, M., Alexiou, P., Ma, J., Pillai, R. S., Mourelatos, Z., and Wang, P. J. (2015). The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. Genes Dev

Wang, J., Saxe, J. P., Tanaka, T., Chuma, S., and Lin, H. (2009a). Mili Interacts with Tudor Domain-Containing Protein 1 in Regulating Spermatogenesis. Curr Biol *19*, 640-644.

Wang, W., Yoshikawa, M., Han, B. W., Izumi, N., Tomari, Y., Weng, Z., and Zamore, P. D. (2014). The Initial Uridine of Primary piRNAs Does Not Create the Tenth Adenine that Is the Hallmark of Secondary piRNAs. Mol Cell *56*, 708-716.

Wang, Y., Juranek, S., Li, H., Sheng, G., Tuschl, T., and Patel, D. J. (2008). Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. Nature *456*, 921-926.

Wang, Y., Juranek, S., Li, H., Sheng, G., Wardle, G. S., Tuschl, T., and Patel, D. J. (2009b). Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. Nature *461*, 754-761.

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., and Patel, D. J. (2008). Structure of the guide-strand-containing argonaute silencing complex. Nature *456*, 209-213.

Warnefors, M., Liechti, A., Halbert, J., Valloton, D., and Kaessmann, H. (2014). Conserved microRNA editing in mammalian evolution, development and disease. Genome Biol *15*, R83.

Warnes, G. R., Bolker, B., and Lumley, T. (2008). gplots: various R programming tools for plotting data, v2. 6.0. The Comprehensive R Archive Network

Watanabe, T., Chuma, S., Yamamoto, Y., Kuramochi-Miyagawa, S., Totoki, Y., Toyoda, A., Hoki, Y., Fujiyama, A., Shibata, T., Sado, T., Noce, T., Nakano, T.,

Nakatsuji, N., Lin, H., and Sasaki, H. (2011). MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline. Dev Cell *20*, 364-375.

Wee, L. M., Flores-Jasso, C. F., Salomon, W. E., and Zamore, P. D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. Cell *151*, 1055-1067.

Wickersheim, M. L., and Blumenstiel, J. P. (2013). Terminator oligo blocking efficiently eliminates rRNA from Drosophila small RNA sequencing libraries. Biotechniques *55*, 269-272.

Wightman, B., Ha, I., and Ruvkun, G. (1993). Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. Cell *75*, 855-62.

Xie, J., Ameres, S. L., Friedline, R., Hung, J. H., Zhang, Y., Xie, Q., Zhong, L., Su, Q., He, R., Li, M., Li, H., Mu, X., Zhang, H., Broderick, J. A., Kim, J. K., Weng, Z., Flotte, T. R., Zamore, P. D., and Gao, G. (2012). Long-term, efficient inhibition of microRNA function in mice using rAAV vectors. Nat Methods *9*, 403-409.

Xie, Z., Allen, E., Wilken, A., and Carrington, J. C. (2005). DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A *102*, 12984-12989.

Xiol, J., Spinelli, P., Laussmann, M. A., Homolka, D., Yang, Z., Cora, E., Couté, Y., Conn, S., Kadlec, J., Sachidanandam, R., Kaksonen, M., Cusack, S.,

Ephrussi, A., and Pillai, R. S. (2014). RNA Clamping by Vasa Assembles a piRNA Amplifier Complex on Transposon Transcripts. Cell *157*, 1698-1711.

Yabuta, Y., Ohta, H., Abe, T., Kurimoto, K., Chuma, S., and Saitou, M. (2011). TDRD5 is required for retrotransposon silencing, chromatoid body assembly, and spermiogenesis in mice. J Cell Biol *192*, 781-795.

Yan, K. S., Yan, S., Farooq, A., Han, A., Zeng, L., and Zhou, M. M. (2003). Structure and conserved RNA binding of the PAZ domain. Nature *426*, 468-474.

Yang, J. S., Maurin, T., Robine, N., Rasmussen, K. D., Jeffrey, K. L., Chandwani, R., Papapetrou, E. P., Sadelain, M., O'Carroll, D., and Lai, E. C. (2010). Conserved vertebrate mir-451 provides a platform for Dicer-independent, Ago2-mediated microRNA biogenesis. Proc Natl Acad Sci U S A *107*, 15163-15168.

Yekta, S. (2004). MicroRNA-Directed Cleavage of HOXB8 mRNA. Science *304*, 594-596.

Yi, R., Qin, Y., Macara, I. G., and Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev *17*, 3011-3016.

Yoda, M., Cifuentes, D., Izumi, N., Sakaguchi, Y., Suzuki, T., Giraldez, A. J., and Tomari, Y. (2013). Poly(A)-Specific Ribonuclease Mediates 3'-End Trimming of Argonaute2-Cleaved Precursor MicroRNAs. Cell Rep *5*, 715-726.

Yoshikawa, M., Peragine, A., Park, M. Y., and Poethig, R. S. (2005). A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis. Genes Dev *19*, 2164-2175.

Zamparini, A. L., Davis, M. Y., Malone, C. D., Vieira, E., Zavadil, J., Sachidanandam, R., Hannon, G. J., and Lehmann, R. (2011). Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila. Development *138*, 4039-4050.

Zhang, F., Wang, J., Xu, J., Zhang, Z., Koppetsch, B. S., Schultz, N., Vreven, T., Meignin, C., Davis, I., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2012). UAP56 Couples piRNA Clusters to the Perinuclear Transposon Silencing Machinery. Cell *151*, 871-884.

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track plots. BMC Bioinformatics *14*, 244.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., and Liu, X. S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol *9*, R137.

Zhang, Z., Koppetsch, B. S., Wang, J., Tipping, C., Weng, Z., Theurkauf, W. E., and Zamore, P. D. (2014a). Antisense piRNA amplification, but not piRNA production or nuage assembly, requires the Tudor-domain protein Qin. EMBO J *33*, 536-539.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol *7*, 203-214.

Zhang, Z., Wang, J., Schultz, N., Zhang, F., Parhad, S. S., Tu, S., Vreven, T., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2014b). The HP1 Homolog

Rhino Anchors a Nuclear Complex that Suppresses piRNA Precursor Splicing. Cell *157*, 1353-1363.

Zhang, Z., Xu, J., Koppetsch, B. S., Wang, J., Tipping, C., Ma, S., Weng, Z., Theurkauf, W. E., and Zamore, P. D. (2011). Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. Mol Cell *44*, 572-584.

Zhao, Y., Yu, Y., Zhai, J., Ramachandran, V., Dinh, T. T., Meyers, B. C., Mo, B., and Chen, X. (2012). The *Arabidopsis* nucleotidyl transferase HESO1 uridylates unmethylated small RNAs to trigger their degradation. Curr Biol *22*, 689-694.

Zheng, K., Xiol, J., Reuter, M., Eckardt, S., Leu, N. A., McLaughlin, K. J., Stark, A., Sachidanandam, R., Pillai, R. S., and Wang, P. J. (2010). Mouse MOV10L1 associates with Piwi proteins and is an essential component of the Piwi-interacting RNA (piRNA) pathway. Proc Natl Acad Sci U S A *107*, 11841-11846.

Zhuang, J., Wang, J., Theurkauf, W., and Weng, Z. (2014). TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Res *42*, 6826-6838.