Transposon activity and control in the *Drosophila melanogaster* germline

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Felix Mürdter
aus Göppingen

Tübingen
2013

Tag der mündlichen Qualifikation:	22.01.2014
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Gerd Jürgens
2. Berichterstatter:	Prof. Dr. Gregory J. Hannon

Contents

Abbreviations

A	adenosine
adh	alcohol dehydrogenase
AGO	Argonaute
APOBEC	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
bp	base pair
C	cytidine
C. elegans	Caenorhabditis elegans
cDNA	complementary DNA
CDS	coding sequence
ChIP-seq	chromatin immunoprecipitation sequencing
DNA	deoxyribonucleic acid
Dnmt	DNA methyltransferase
dsDNA	double stranded DNA
dsRNA	double stranded RNA
G	guanosine
GFP	green fluorescent protein
H3K4	histone 3 lysine 4
H3K9me3	histone 3 lysine 9 trimethylation
hpRNA	hairpin RNA
IS	insertion sequence
kb	kilobase pairs
kD	kilo Dalton
LINE	long interspersed element
LTR	long terminal repeat
miRNA	micro RNA
mRNA	messenger RNA
nt	nucleotide
ORF	open reading frame

OSS	ovarian somatic sheet
PCR	polymerase chain reaction
piRNA	piwi-interacting RNA
polyA	poly-adenosine
PTGS	post-transcriptional gene silencing
qPCR	quantitative real-time PCR
RISC	RNA induced silencing complex
RNA	ribonucleic acid
RNA polymerase II	Pol II
RNA-seq	RNA high-throughput sequencing
RNAi	RNA interference
RNase	ribonuclease
RNP	ribonucleoprotein
rRNA	ribosomal RNA
RT	reverse transcriptase
siRNA	small interfering RNA
sRNA	small RNA
T	thymidine
TAS	telomere associated sequence
TE	transposable element
TF	transcription factor
TGS	transcriptional gene silencing
tRNA	transfer RNA
U	uridine
UAS	upstream activating sequence
UTR	untranslated region

Zusammenfassung

Springende genetische Elemente sind ein grundlegender Bestandteil unseres Genoms. Ihre Präsenz und Mobilität kann einen starken Einfluss auf Genregulation und Genomevolution ausüben. Dieser Einfluss kann jedoch auch negative Auswirkungen haben, weshalb Transposons einer strikten zellulären Kontrolle unterliegen. Unkontrollierte Mobilisierung kann die Integrität des Genomes gefährden, zu Mutationen und in extremen Fällen zur Sterilität des Organismus führen. Aus diesen Gründen haben sich mehrere Mechanismen entwickelt, welche die Aktivität von Transposons überwachen. In Keimzellen von *Drosophila melanogaster* übernehmen diese Rolle kleine RNAs zusammen mit einer Gruppe von evolutionär hochkonservierten Proteinen der Argonaute Klasse: Piwi, Argonaute3 und Aubergine. Diese Enzyme können aktive Transposons anhand der gebundenen kleinen RNA erkennen und konsequenterweise abbauen oder ihre Transkription gänzlich verhindern. Die gebundenen kleinen RNAs, oder Piwi-interagierenden RNAs (piRNAs), werden aus langen, einzelsträngigen RNA Vorläufern gebildet, welche Homologie zu Transposons besitzen. Es ist bislang unklar auf welchem Weg Transposons ihre Sequenzinformation zu diesen sogenannten piRNA Clustern hinzufügen. Um diesen Vorgang besser zu verstehen, haben wir exogene Sequenzen in mehrere solche Cluster inseriert, um zu testen ob diese in kleine RNAs prozessiert werden. Unsere Ergebnisse bestätigen daß dies der Fall ist, selbst wenn der piRNA Cluster von seiner üblichen Umgebung im Genom separiert wird.

Auf welcher Weise diese kleinen RNAs aus ihren Vorläufer-Molekülen hergestellt werden und welche Gene an der Biogenese und der daraus folgenden Transposon Kontrolle beteiligt sind ist größtenteils unbekannt. Aus diesem Grund haben wir ein Genom-weites Screening unternommen, um nach Möglichkeit alle an diesen Prozessen beteiligten Gene zu identifizieren. Abschaltung von 87 Genen in einer Zellkultur hergeleitet von Drosophila Follikel Zellen führte zu dramatisch erhöhter Expression von einer Spezialklasse von Transposons. Durch weitere in vivo Experimente konnten wir nachweisen daß mehrere dieser Gene in der piRNA Biogenese beteiligt sind. Andere Gene, wie zum Beispiel *CG3893*, welches wir *asterix* tauften, scheinen einen starken Einfluss auf heterochromatische Histon Modifikationen zu haben, welche normalerweise mit stillgelegten Transposons assoziiert sind. Eine detaillierte Charakterisierung von *asterix* Mutanten zeigte, daß dieses Protein im Nukleus lokalisiert ist, keinen Einfluss auf piRNA Biogenese hat, und deshalb mit hoher Wahrscheinlichkeit zusammen mit Piwi dem transkriptionellen Transposonregulations Komplex angehört.

Summary

Mobile genetic elements are a fundamental part of our genome. Their presence can provide a rich source of regulatory elements and has a strong impact on genome evolution. Nonetheless, their activity can lead to insertional mutagenesis and thus has to be tightly controlled. Failure to do so can compromise genomic integrity and lead to sterility. Several pathways have evolved that exert this control. Germ cells, as the carriers of the inheritable genome, are protected by a highly conserved small RNA pathway: the piRNA pathway. Argonaute proteins of the Piwi-clade together with their bound small RNAs (the piRNAs) comprise this pathway and target transposable elements for degradation and transcriptional gene silencing. In animals, piRNAs are derived from discrete genomic loci called piRNA clusters. These transposon graveyards act as a molecular memory of all elements an organism or its ancestors were exposed to. In an effort to better understand piRNA mediated resistance against a new invader, we investigated the molecular events that take place upon *de novo* insertion of any sequence into such a cluster. We sequenced small RNA populations from flies and mice expressing ectopic piRNA clusters tagged with artificial sequences, as well as clusters divorced from their native genomic location. We detect artificial piRNAs against these sequences and at least in one case demonstrate their silencing capabilities.

Because little is known about the details of how piRNAs are produced from their precursors, and how mature piRNAs silence transposons, we decided to perform a genome-wide screen for factors involved in these processes. We identified 87 genes that are essential for transposon silencing. Follow up *in vivo* experiments showed that some of these factors are involved in piRNA biogenesis. Others, such as *CG3893* (*asterix*) have a strong impact on silencing histone marks over transposon loci and may therefore be part of the Piwi-mediated transcriptional silencing complex. Detailed analysis on *asterix* mutants revealed that knockdown of this gene did not change mature piRNA levels, which together with its nuclear localization points towards an involvement in the effector step of transposon silencing.

Acknowledgments

In April 2009 I arrived in Cold Spring Harbor with the set goal to only stay for as long as it would take to finish my Diploma studies, but definitely not longer. Today, more than 4 years later, I am still here, finishing my thesis work. I never regret having made the decision to stay, mostly because of the people I met along the way. I would like to thank these people, friends and colleagues alike, because my success and happiness rests on them as much as it does on myself.

First and foremost I would like to thank Greg for taking a leap of faith and letting me join his lab as an undergrad. Being part of your lab is truly an amazing experience, both at the bench and at random food stands in the alleys of Xian.

I am very grateful to Gerd, who allowed me to do my research abroad. I am very proud to be able to call the Eberhard Karls Universität my *alma mater*, which would not be possible without your open mindedness and generosity.

I would like to thank Paloma, mi nena, for always being there with me and for me. With you (screening through) a hurricane feels like a breeze.

I am very thankful to Leah, Paloma and Antoine for reading and commenting on this manuscript and for many fruitful discussions.

I am grateful to Sabrina, who managed all the logistics and supplies for every experiment mentioned in this thesis. Especially the screen would not have been possible without you. I would also like to thank Jo, for all the help and support throughout the years.

I am greatly indebted to Gordy and Oliver for all the support with bioinformatics and for not loosing your patience with me.

I would like to thank all current and former members of the Hannon lab for creating such a great work environment. Many of you have helped me and contributed to this thesis, namely but not limited to: Ben, Alexei, Astrid, Jon, Sho, Andres, Emily Lee, YY, Yicheng, Nik, Adam and Vasily. I would also like to thank my collaborators Caifu, Richard, Jesse, Molly, Stephanie, Norbert and Alon.

I am very fortunate to have met many good friends here, which will hopefully stay with me beyond my time in grad school. Thank you Colin, Johannes, Fred and Saya for being great housemates. Thank you Emily, Antoine, Simon, Audrey, Brian, Katie, Ian, Kaja, Mike, Anna, Shane, Leah, Roberto, Filipe, Kata, Ingrid and Krista for many unforgettable moments.

Zu guter Letzt gilt mein Dank meiner Familie, meinen Eltern und meinem Bruder für nimmerendende Unterstützung.

Manuscripts included in this thesis

Muerdter, F.[*], Guzzardo, P.M.[*], Gillis, J., Luo, Y., Yu, Y., Chen, C., Fekete, R., Hannon, G.J. (2013) A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila. *Molecular Cell*, 50 (5), pp. 736–748.
DOI: 10.1016/j.molcel.2013.04.006.

Guzzardo, P.M., **Muerdter, F.**, Hannon, G.J. (2013) The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics and Development*, 23 (1), pp. 44-52.
DOI: 10.1016/j.gde.2012.12.003.

Muerdter, F.[*], Olovnikov, I.[*], Molaro, A.[*], Rozhkov, N.V.[*], Czech, B., Gordon, A., Hannon, G.J., Aravin, A.A. (2012) Production of artificial piRNAs in flies and mice. *RNA*, 18 (1), pp. 42-52.
DOI: 10.1261/rna.029769.111.

Additional publications

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*, Article in Press.

Preall, J.B.[*], Czech, B.[*], Guzzardo, P.M., **Muerdter, F.**, Hannon, G.J. (2012) Shutdown is a component of the Drosophila piRNA biogenesis machinery. *RNA*, 18 (8), pp. 1446-1457.

Haase, A.D., Fenoglio, S., **Muerdter, F.**, Guzzardo, P.M., Czech, B., Pappin, D.J., Chen, C., Gordon, A., Hannon, G.J. (2010) Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. *Genes and Development*, 24 (22), pp. 2499-2504.

Mathieu, J., Yant, L.J., **Mürdter, F.**, Küttner, F., Schmid, M. (2009) Repression of Flowering by the miR172 Target SMZ. *PLoS Biology*, 7 (7).

[*]These authors contributed equally to this work

Contributions to publications

1. **Muerdter, F.**[*], Guzzardo, P.M.[*], Gillis, J., Luo, Y., Yu, Y., Chen, C., Fekete, R., Hannon, G.J. (2013) A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila. *Molecular Cell*, 50 (5), pp. 736–748.

   FM and PMG contributed equally to this work. FM, PMG and GJH conceived all experiments and wrote the manuscript with input from JG and YY. FM made the figures. FM and PMG executed the primary and validation screen. FM wrote the scripts controlling the liquid handlers of the primary screen. RF and CC designed primers for the primary screen qPCR assay. FM wrote the analysis pipeline for the primary screen, RNA-seq, ChIP-seq and small RNA-seq data. JG did the statistical analysis on primary and validation screen data. PMG produced RNA-seq and ChIP-seq libraries; FM constructed small RNA-seq libraries. YL and YY imaged fluorescent fusion proteins in OSS. GJH supervised the project.

2. Guzzardo, P.M., **Muerdter, F.**, Hannon, G.J. (2013) The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics and Development*, 23 (1), pp. 44-52.

   PMG and GJH conceived the concept of this review and wrote the manuscript. FM commented on the manuscript and made the figures.

3. **Muerdter, F.**[*], Olovnikov, I.[*], Molaro, A.[*], Rozhkov, N.V.[*], Czech, B., Gordon, A., Hannon, G.J., Aravin, A.A. (2012) Production of artificial piRNAs in flies and mice. *RNA*, 18 (1), pp. 42-52.

   FM, IO, AM and NVR contributed equally to this work. FM contributed data for a tagged, ectopic piRNA cluster in flies (*flamenco*), which was previously described *(Functional Dissection of Primary piRNA Biogenesis in Drosophila. Diplomarbeit der Fakultät für Biologie der Eberhard Karls Universität Tübingen, vorgelegt von Mürdter, Felix, Tübingen, März 2010)*. FM designed the analysis pipeline to detect and interpret artificial piRNA production from three modified piRNA clusters in fly: X-TAS (data contributed by NVR), *flamenco* (data contributed by FM) and the 3'UTR of *traffic jam* (data contributed by IO and AAA); as well as an ectopic piRNA cluster in mouse (data contributed by AM). Analyzing all datasets from independent authors with the same analysis workflow was essential for being able to compare these datasets. Therefore, creating this pipeline was a significant advance and all scripts are presented in this work (see appendix A.1). The perl script used for ping-pong analysis was designed by FM and implemented by AG with comments from FM (see appendix A.1.2). AAA and GJH wrote the manuscript with comments from the other authors. FM and IO made figures with comments from the other authors. AAA and GJH supervised the project.

[*]These authors contributed equally to this work

Chapter 1

Introduction

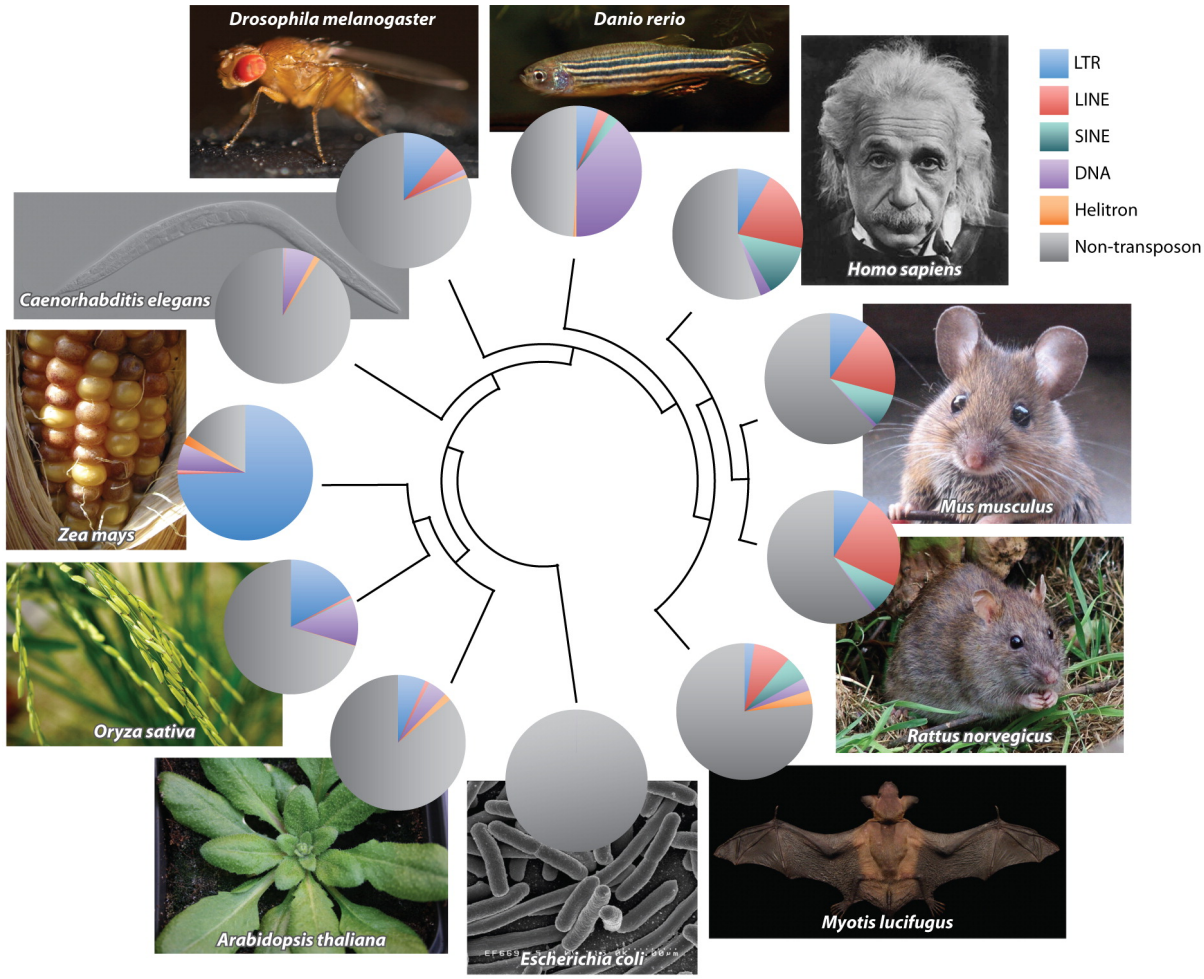
Over 60 years ago, Barbara McClintock discovered mobile genetic elements (McClintock, 1951). It has become evident that McClintock's research was not only groundbreaking, but also that her interpretations thereof were close to our current understanding of genome dynamics (McClintock, 1984). Unlike early skeptics had thought, transposable elements (TEs) are present in most observed genomes, both prokaryotic and eukaryotic. Furthermore, their genetic behavior more closely resembles that of the very control elements McClintock initially proposed, as opposed to just being 'junk DNA', as they were later referred to (McClintock, 1951; Ohno, 1972; Rebollo et al., 2012). In fact, TEs may influence every aspect of genome evolution and gene regulation. Their presence can be a rich source of heritable variability resulting in natural variation within species or even in new speciation events. On the other hand, when uncontrolled, TEs can lead to insertional mutagenesis and hence may have an impact on virtually every disease with a genetic component.

In this chapter, I will review our current understanding of transposon activity in a range of organisms, and how this activity is controlled. I will highlight the impact that small RNA (sRNA) pathways, particularly the piRNA pathway, have on transposon control. Lastly, I will outline how this thesis work aimed to uncover a general genetic framework of transposon control in the *Drosophila melanogaster* germline.

TEs are ubiquitous in all observed genomes

When the human genome was published in the early 2000s, one surprising discovery was the large amount of sequence with homology to transposable elements. More than 50% of our genome was found to be of repetitive nature, with about 45% being recognized as transposon derived (Figure 1.1) (Lander et al., 2001). In comparison, coding genes only comprise around 5% of the sequence information. This vast amount of repetitive, non-coding sequence can explain in part why genome size does not necessarily correlate with genic (and organismal) complexity (the C-value paradox) (Gregory and Hebert, 1999; Hartl, 2000). It seems that the most variable aspect of eukaryotic genomes is the percentage of repetitive or

transposon-derived sequence (Figure 1.1). For example, *Drosophila melanogaster* has a comparably low TE content with 4%-10% (Adams et al., 2000; Kaminker et al., 2002; Spradling and Rubin, 1981), whereas 85% of the maize genome consists of TE sequence (Schnable et al., 2009).



Huang CRL, et al. 2012.
Annu. Rev. Genet. 46:651–75

Figure 1.1: Transposon compositions in different species.
In clockwise order, species shown are *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Myotis lucifugus*, *Escherichia coli*, *Arabidopsis thaliana*, *Oryza sativa*, *Zea mays*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Danio rerio*. The phylogenetic tree in the center describes the evolutionary relationships among them. The pie charts illustrate the fraction of the genome accounted for by different transposon classes (Smit et al., 2010).
Image and legend from: Huang et al. (2012).

The omnipresence and sheer number of different TEs in eukaryotic genomes calls for a clear classification and nomenclature system. Traditionally, TEs have been separated according to their transposition mechanisms and the underlying differences in their coding repertoire (Jurka et al., 2005; Kapitonov and Jurka, 2008; Wicker et al., 2007).

TEs can be classified based on transposition intermediates and mechanistic or enzymatic criteria.

TEs can be divided into two main classes: Retrotransposons (class 1) and DNA transposons (class 2) (Wicker et al., 2007). This division into two classes is based on whether or not an RNA intermediate is made during transposition. DNA transposons mobilize via a 'cut-and-paste' mechanism for which typically only a single enzyme is needed: the transposase (Figure 1.2A) (Benjamin and Kleckner, 1989; Kleckner, 1990; Levin and Moran, 2011). The activity of the transposase is guided by the presence of terminal inverted repeats flanking the TE locus that are bound by it. The transposase excises the original locus and pastes it into a new genomic location, leaving behind short target site duplications at the site of cleavage.

Retrotransposons encode a reverse transcriptase (RT), which is critical for transposition through an RNA intermediate (Boeke et al., 1985; Garfinkel et al., 1985). After transcription of a class 1 TE locus, the RT uses the transposon mRNA as a template to build complementary DNA (cDNA), which is then integrated by an integrase (the functional counterpart of a transposase) into the target locus (Figure 1.2B) (Levin and Moran, 2011). Priming of reverse transcription can occur through different ways: One subclass of retrotransposons encodes long terminal repeat (LTR) sequences at the 5' and 3' ends. These LTRs not only contain the internal promoter for transcription by RNA polymerase II (Pol II), but also contain binding sites for structural RNAs (most commonly tRNAs) downstream of the 3' end of the 5' LTR (Craig et al., 2002). Binding of the tRNA primer to these sites creates a free 3' OH that can be used by the RT for cDNA synthesis. In the case of non-LTR retrotransposons, RT priming occurs at the target locus (Figure 1.2C) (Levin and Moran, 2011; Luan et al., 1993). In this case, a TE-encoded endonuclease creates a single-stranded nick in the dsDNA of the new genomic location, thereby creating a free 3'-OH.

How can TEs influence genome structure and evolution?

Given the variety of regulatory elements transposons can contain within their sequence, it is not surprising that mobilization of these elements can influence host gene regulation. For example, TEs can harbor elements that attract transcriptional machineries such as Pol II. They can introduce enhancer sequences, transcription factor binding sites, insulator sequences, polyadenylation signals, transcriptional termination sites and splice sites into novel genomic surroundings (Rebollo et al., 2012). It is estimated that around 20% of gene regulatory sequences in the human genome are exapted from transposable elements (Lindblad-Toh et al., 2011; Lowe and Haussler, 2012). Thus, the mere fact that TEs are capable of mobilization implies a great potential impact on the host's gene regulatory networks. Such changes in regulatory networks can be of adaptive nature and create natural variation within populations (González et al., 2008, 2009; González and Petrov, 2009). As mentioned before, transposon activity can also alter genome size: some hybrid sunflower species show up to 20-fold increases in retrotransposon sequences within their genomes compared to their parental species (Ungerer et al., 2006). Another case of TE-driven genome

evolution is the presence of host genes derived from domesticated transposable elements. For instance, transposases serve as molecular blueprints for a range of DNA binding proteins such as transcriptional activators and repressors (Feschotte and Pritham, 2007).

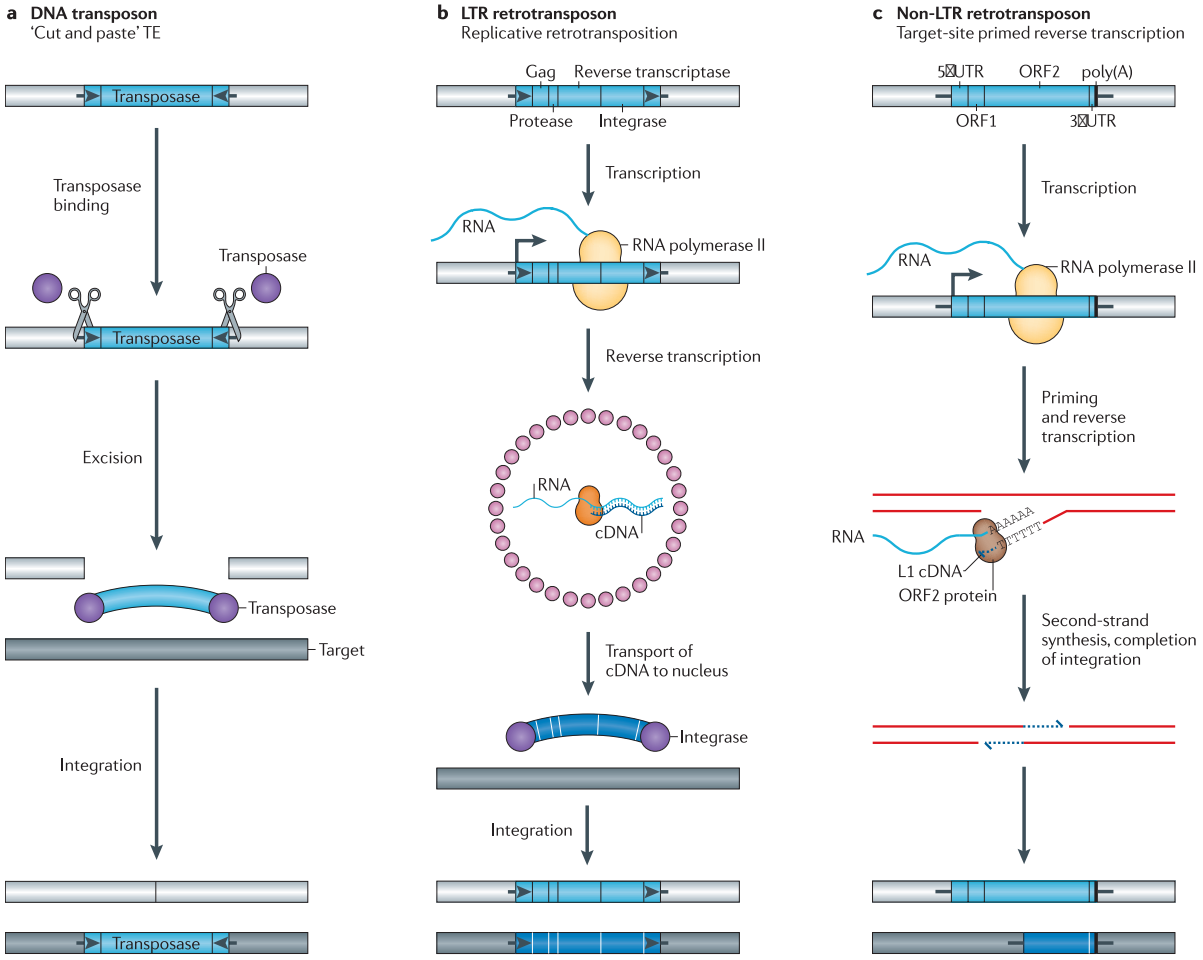


Figure 1.2: The diverse mechanisms of transposon mobilization.
a) DNA transposons. Many DNA transposons are flanked by terminal inverted repeats (TIRs; black arrows), encode a transposase (purple circles), and mobilize by a 'cut and paste' mechanism (represented by the scissors). The transposase binds at or near the TIRs, excises the transposon from its existing genomic location (light grey bar) and pastes it into a new genomic location (dark grey bar). The cleavages of the two strands at the target site are staggered, resulting in a target-site duplication (TSD) typically of 4–8 bp (short horizontal black lines flanking the transposable element (TE)) as specified by the transposase. Retrotransposons (b and c) mobilize by replicative mechanisms that require the reverse transcription of an RNA intermediate. b) LTR retrotransposons contain two long terminal repeats (LTRs; black arrows) and encode Gag, protease, reverse transcriptase and integrase activities, all of which are crucial for retrotransposition. The 5' LTR contains a promoter that is recognized by the host RNA polymerase II and produces the mRNA of the TE (the start-site of transcription is indicated by the right-angled arrow). In the first step of the reaction, Gag proteins (small pink circles) assemble into virus-like particles that contain TE mRNA (light blue), reverse transcriptase (orange shape) and integrase. The reverse transcriptase copies the TE mRNA into a full-length dsDNA. In the second step, integrase (purple circles) inserts the cDNA (shown by the wide, dark blue arc) into the new target site. Similarly to the transposases of DNA transposons, retrotransposon integrases create staggered cuts at the target sites, resulting in TSDs.
Image and legend from: Levin and Moran (2011).

(*legend continued on next page*)

Lastly, a rather straightforward form of TE exaptation is the domestication of an entire element class for host purposes in the case of telomere maintenance by *HeT-A*, *TART* and *THARE* (Zhang and Rong, 2012). In most eukaryotic organisms, the enzyme telomerase prevents incomplete replication of the telomeres by capping and elongating chromosome ends. In the telomerase deficient fruit fly *Drosophila melanogaster*, *HeT-A*, *TART* and *THARE* take over this function by transposing specifically to or within telomeric repeat arrays.

The negative impact of TE activity

Although TEs aid in shaping genome structure and gene regulatory networks, transposon activity can also lead to insertional mutagenesis and double stranded DNA breaks (McClintock, 1942, 1950, 1951). Such insertional mutagenesis can potentially contribute to any pathological phenotype or disease with a genetic basis. For example, *de novo LINE1* insertions (a non-LTR retrotransposon) were shown to cause Hemophilia A in humans (Kazazian et al., 1988). From a broader perspective, it has been shown that the accumulation of TE sequences in *Drosophila* correlates with an overall decrease in fitness (Pasyukova et al., 2004). This is not surprising, given that uncontrolled mobilization of a single TE in the *Drosophila* germline can lead to sterility in a phenomenon called hybrid dysgenesis. This genetic peculiarity is a collection of aberrant phenotypes such as an increased mutation rate and sterility in offspring of crosses between wild-caught males and laboratory-strain females (Kidwell et al., 1977; Picard, 1976). In the dysgenic offspring, it was shown that the uncontrolled activity of newly introduced transposon classes such as the *I* and *P* element cause these phenotypes (Bucheton et al., 1984; Castro and Carareto, 2004; Chambeyron and Bucheton, 2005; Kidwell, 1983; Pélisson, 1981; Rubin et al., 1982). The fact that one population (the wild-caught flies) is adapted to the presence of the TE, while the other (the laboratory-strain) is susceptible to it, illustrates that transposable elements can create sexual barriers ultimately leading the way to new speciation events. Surprisingly, the genetically identical offspring of the reciprocal cross between wild-caught females and laboratory-strain males does not exhibit these effects. This suggests that there is a maternally transmitted factor that protects the genome of the offspring against the unknown

Figure 1.2: (Continued from page 4.) c) Non-LTR retrotransposons lack LTRs and encode either one or two ORFs. As for LTR retrotransposons, the transcription of non-LTR retrotransposons generates a full-length mRNA (wavy, light blue line). However, these elements mobilize by target-site-primed reverse transcription (TPRT). In this mechanism, an element-encoded endonuclease generates a single-stranded 'nick' in the genomic DNA, liberating a 3'-OH that is used to prime reverse transcription of the RNA. The proteins that are encoded by autonomous non-LTR retrotransposons can also mobilize non-autonomous retrotransposon RNAs, as well as other cellular RNAs. The TPRT mechanism of a long interspersed element 1 (L1) is depicted in the figure; the new element (dark blue rectangle) is 5' truncated and is retrotransposition-defective. Some non-LTR retrotransposons lack poly(A) tails at their 3' ends. The integration of non-LTR retrotransposons can lead to TSDs or small deletions at the target site in genomic DNA. For example, L1s are generally flanked by 7–20 bp TSDs.
Image and legend from: Levin and Moran (2011).

invader. It is now known, that this factor is related to small RNA pathways, which are further discussed below (Brennecke et al., 2008).

Taken together, it is clear that although TEs have a considerable positive impact on genome evolution, their activity must be tightly controlled in order to prevent the detrimental effects of transposition.

How are TEs controlled?

Besides many highly conserved, host-encoded defense pathways, transposons seem to have evolved self-control mechanisms that constrain their own levels of transposition. For instance, the germline specificity of the *P* element in *Drosophila* is manifested at the splicing step: this element contains two introns which are spliced ubiquitously, while a third intron, which is necessary for a full-length transposase, is specifically spliced in the germline (Laski et al., 1986). Thus, the *P* element expression pattern ensures that the TE can be active within the only cell type that is relevant for propagation (the germ cells), while not compromising the host's overall fitness by causing deleterious somatic mutations.

Another system that is thought to govern TE activity are host factors such as the APOBEC family of cytidine deaminases, which traditionally are associated with restricting exogenous virus infections (Sheehy et al., 2002; Trono, 2004). These deaminases have also been implicated in retrotransposon control in yeast and mammals, further highlighting the potential connections between viral defense pathways and retrotransposon control (Chiu and Greene, 2008). However, it is unclear how widespread APOBEC-mediated defense mechanisms are among animals, because there seems to be no clear orthologous function in *Drosophila* (Conticello et al., 2005).

A further major control mechanism in many eukaryotes is the methylation of DNA at transposon loci, which leads to transcriptional silencing (Slotkin and Martienssen, 2007). Plants deficient in DNA methylation exhibit bursts of retrotransposon transcription (Tsukahara et al., 2009). This feature seems to be conserved in mammals, as retrotransposons are reactivated in male germ cells of DNA methyltransferase 3-like (Dnmt3L) deficient mice (Bourc'his and Bestor, 2004). These TEs normally undergo *de novo* DNA methylation mediated by Dnmt3L during a brief period in the development of spermatogonial precursors. Deletion of the methyltransferase prevents the reestablishment of methylation of transposon loci and their subsequent silencing. Interestingly, *de novo* DNA methylation is linked to sRNA pathways, which are further discussed below (Aravin et al., 2008; Okano et al., 1999; Sabin et al., 2013). It should be noted that a number of species do not encode functional *de novo* DNA methyltransferases (Dnmt2-only genomes) (Raddatz et al., 2013). An interesting question is how organisms without DNA methylation pathways, such as *Drosophila*, have evolved parallel strategies to establish epigenetic silencing of TE loci in a heritable manner (Guzzardo et al., 2013).

Small RNA pathways and TE control

One of the foremost challenges in transposon control is to distinguish TE-related targets from non-coding or protein-coding genes (Malone and Hannon, 2009). Whether or not this differentiation and subsequent silencing is achieved at the DNA or RNA level, all classes of transposable elements must undergo transcription in order to mobilize (not to be confused with the reverse transcription of retrotransposons). Thus, it is possible that the presence and recognition of TE transcripts is the single unifying factor in transposon control pathways.

A major breakthrough related to this idea occurred when Andrew Fire, Craig Mello and colleagues discovered in the late 1990s that the expression of endogenous mRNAs can be disrupted by introducing exogenous double stranded RNA into worms (Fire et al., 1998). At the core of this mechanism, termed RNA interference (RNAi), is a class of noncoding RNAs named small interfering RNAs (siRNAs) (Hamilton et al., 2002; Hamilton and Baulcombe, 1999). These siRNAs are processed from double stranded RNA precursors by the activity of a ribonuclease called Dicer (Bernstein et al., 2001; Ketting et al., 2001; Knight and Bass, 2001; Lee et al., 2004; Liu et al., 2003). Once matured to their length of 21nt, siRNAs can act as guides in ribonucleoprotein (RNP) complexes, based on sequence complementarity, in order to silence target mRNAs (Elbashir et al., 2001; Hammond et al., 2000; Zamore et al., 2000). Given the ability of small RNAs to target (and thereby distinguish) any given transcript, these molecular guides are well suited for TE silencing. Experimental evidence for this hypothesis first emerged in worms: endogenous double stranded RNA in the *C. elegans* germline triggers RNAi to silence transposable elements (Ketting et al., 1999; Sijen and Plasterk, 2003; Tabara et al., 1999). Later, this form of natural RNAi as a means of transposon control was shown to be conserved in somatic cells of mammals and flies, and is mediated by endogenous siRNAs (Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; Soifer et al., 2005; Tam et al., 2008; Yang and Kazazian, 2006).

The protein partners of small RNAs in the RNP complexes responsible for transposon recognition are called Argonaute proteins (AGOs) (Hammond et al., 2001). In flies, siRNAs interact with Argonaute2 (AGO2) (Okamura et al., 2004; Tomari et al., 2007). Structural and biochemical analysis of orthologous AGO2 proteins revealed that this enzyme exhibits ribonuclease activity, consistent with its proposed role in target mRNA degradation (Liu et al., 2004; Song et al., 2004). Interestingly, Argonaute proteins are highly conserved and can be grouped into several clades based on multiple sequence alignments (Tolia and Joshua-Tor, 2007). AGO2 is part of the Ago-like clade, which can be found in a diverse spectrum of eukaryotes such as plants (*Arabidopsis thaliana*), fungi (*Schizosaccharomyces pombe*) and a wide range of animals (*C. elegans*, *Drosophila melanogaster*, mice and humans). Another prominent member of the Ago-like clade is AGO1, which usually binds another class of Dicer-dependent sRNAs, the microRNAs (miRNAs). These slightly longer regulatory RNAs (22-23nt) are involved in translational repression of endogenous protein-coding genes (Ghildiyal and Zamore, 2009). miRNA directed transposon silencing is not a generally accepted theme, but cannot be ruled out entirely. Several human mRNAs were shown to contain TE

derived sequences that can be targeted by miRNAs (Smalheiser and Torvik, 2006). Superficially, this may look like transposon repression through miRNAs. However, it seems more likely that these transposon derived sequences have been co-opted to serve as regulatory elements, and that the targeted mRNAs are the principal subject of miRNA control. One argument supporting this interpretation is that these TE derived sequences were found in the 3' UTR of cellular mRNAs, the preferred seed location of miRNA targeting, and not within intergenic regions as distinct TE copies. The question remains, however, why these TE derived sequences show complementarity to miRNAs. Interestingly, there is evidence for TE or repeat derived miRNAs, providing an explanation for how these two entities may be connected (Piriyapongsa and Jordan, 2007, 2008; Piriyapongsa et al., 2006).

In regard to TE control, the most interesting clade of AGOs is the Piwi-clade. This subset of proteins can be grouped together based on sequence homology to the *Drosophila melanogaster* Piwi protein (Tolia and Joshua-Tor, 2007). Piwi is required for proper germline stem cell division and was previously implicated in transcriptional and post-transcriptional gene silencing and transposon control (Cox et al., 1998, 2000; Kalmykova et al., 2005; Lin and Spradling, 1997; Pal-Bhadra et al., 2002, 2004; Sarot et al., 2004). Piwi is specifically expressed in germ cells and thus is a perfect fit for the necessity of a specialized defense pathway against TEs in these cells. Indeed, most small RNAs that are bound to Piwi, the Piwi-interacting RNAs (piRNAs), are homologous to repeats, making them the appropriate guides for TE targeting (Aravin et al., 2001; Saito et al., 2006; Vagin et al., 2006).

piRNA directed post-transcriptional transposon silencing in flies

A comprehensive analysis of piRNAs bound by all three *Drosophila* Piwi proteins, Piwi, Aubergine (AUB) and Argonaute3 (AGO3), revealed that the subpopulations of piRNAs specific to each protein differ in several ways: Piwi and AUB bind piRNAs that are mainly anti-sense to TEs, whereas AGO3 binds sense species (Brennecke et al., 2007; Gunawardane et al., 2007). This implies that each protein may have distinct roles in TE silencing. Differential expression and localization patterns add to this hypothesis: *ago3* and *aub* are only expressed in germ cells of the ovary, while *piwi* is expressed in germ cells and follicle cells (Brennecke et al., 2007; Gunawardane et al., 2007). Furthermore, AGO3 and AUB localize to a cytoplasmic, peri-nuclear structure called 'nuage', whereas Piwi exhibits a nuclear localization (Brennecke et al., 2007; Cox et al., 2000; Findley, 2003; Lim and Kai, 2007; Mahowald, 1971a,b; Nagao et al., 2011).

Intriguingly, transposable elements reflect the divide into follicle cell and germ cell specificity in their expression. For example, follicle cells express TEs of the *gypsy* family of retrotransposons (Bucheton, 1995; Pelisson et al., 1994). Piwi-bound piRNAs homologous to *gypsy* display a striking clustering effect when mapped back to the genome; they originate from a discrete locus on chromosome X called *flamenco* (Brennecke et al., 2007), a locus that has long been implicated in *gypsy* control (Bucheton, 1995; Chalvet et al., 1999; Mével-Ninio et al., 2007; Pelisson et al., 1994; Pélisson et al., 1997; Sarot et al., 2004). Analysis of P element insertions into the putative promoter region of this locus suggests that *flamenco* is

transcribed as a single, 180kb long precursor that gives rise to the population of mature piRNAs targeting *gypsy* elements. When *flamenco* transcription is disrupted by the insertion, these piRNAs are lost, and *gypsy* mRNA is strongly upregulated (Brennecke et al., 2007). Because of the clustering effect of piRNAs mapping to these discrete loci, they have been named piRNA clusters (Brennecke et al., 2007).

In contrast to *flamenco*-derived piRNAs in follicle cells, piRNAs bound to the germ cell specific Piwi proteins AUB or AGO3 do not show homology to *gypsy*. Instead, they show biases towards reads sense and anti-sense to a number of germ cell specific TEs. The presence of sense species of piRNAs, which are preferentially loaded into AGO3, does not fit the simple model of TE targeting through complementary, anti-sense guides. However, their existence can be explained when considered together with other molecular characteristics of these piRNA populations. Species bound by Piwi and AUB tend to have a strong bias towards a uridine at their 5' end. In contrast, AGO3-bound piRNAs exhibit a strong preference for an adenosine at position 10. This coincides with an unusually high number of piRNA 'pairs' that overlap by 10 nt. These observations led to the proposed model of secondary piRNA biogenesis termed the ping-pong cycle (Figure 1.3) (Brennecke et al., 2007; Gunawardane et al., 2007).

In this model, AUB, loaded with a piRNA bearing a 5' uridine, can target and nucleolytically 'slice' a transcript of an active transposon. This cleavage creates the new 5' end of a secondary piRNA, which bears an adenosine at position 10, by definition of sequence complementarity. After the secondary piRNA is loaded into AGO3, the complex would then be able to slice transcripts of piRNA generating loci, creating a piRNA with a 5' U bias, which can be loaded into Piwi and AUB, thereby closing the cycle. This model not only explains the observed sequence and strand biases of mature piRNA populations in germ cells, but also implies that active transcription of a TE leads to an amplification of the silencing response, adding an adaptive component to small RNA-directed transposon control. The cell type specific expression patterns of Piwi proteins suggest that the ping-pong cycle is exclusive to germ cells, possibly because these cells have a higher burden of active transposon transcription. Taken together, these data indicate that the differential load of TEs in different cells of the *Drosophila* ovary led to the evolution of two specialized piRNA pathways, one primary pathway in germ cells and somatic follicle cells, and a more complex pathway exclusive to germ cells (Figure 1.4A) (Li et al., 2009a; Malone et al., 2009).

Primary piRNA biogenesis in *Drosophila*

Unlike siRNAs and miRNAs, piRNAs are processed from precursors independently of Dicer (Houwing et al., 2007; Vagin et al., 2006). These precursors are derived from discrete genomic loci or piRNA clusters, which harbor an abundance of TE fragments. piRNAs originating from these clusters can be considered a catalogue of transposons that a population has been exposed to. Insertion of a new TE into one of these 'transposon graveyards' leads to *de novo* piRNA production and subsequent resistance to the invader (Khurana et al., 2011).

Data from mice and flies suggests that piRNA clusters are transcribed as a single precursor by Pol II and exhibit a poly-adenosine (polyA) tail, like other conventional mRNAs (Brennecke et al., 2007; Li et al., 2013). In mice, Pol II recruitment is driven by an ancient transcription factor (TF), A-MYB (Li

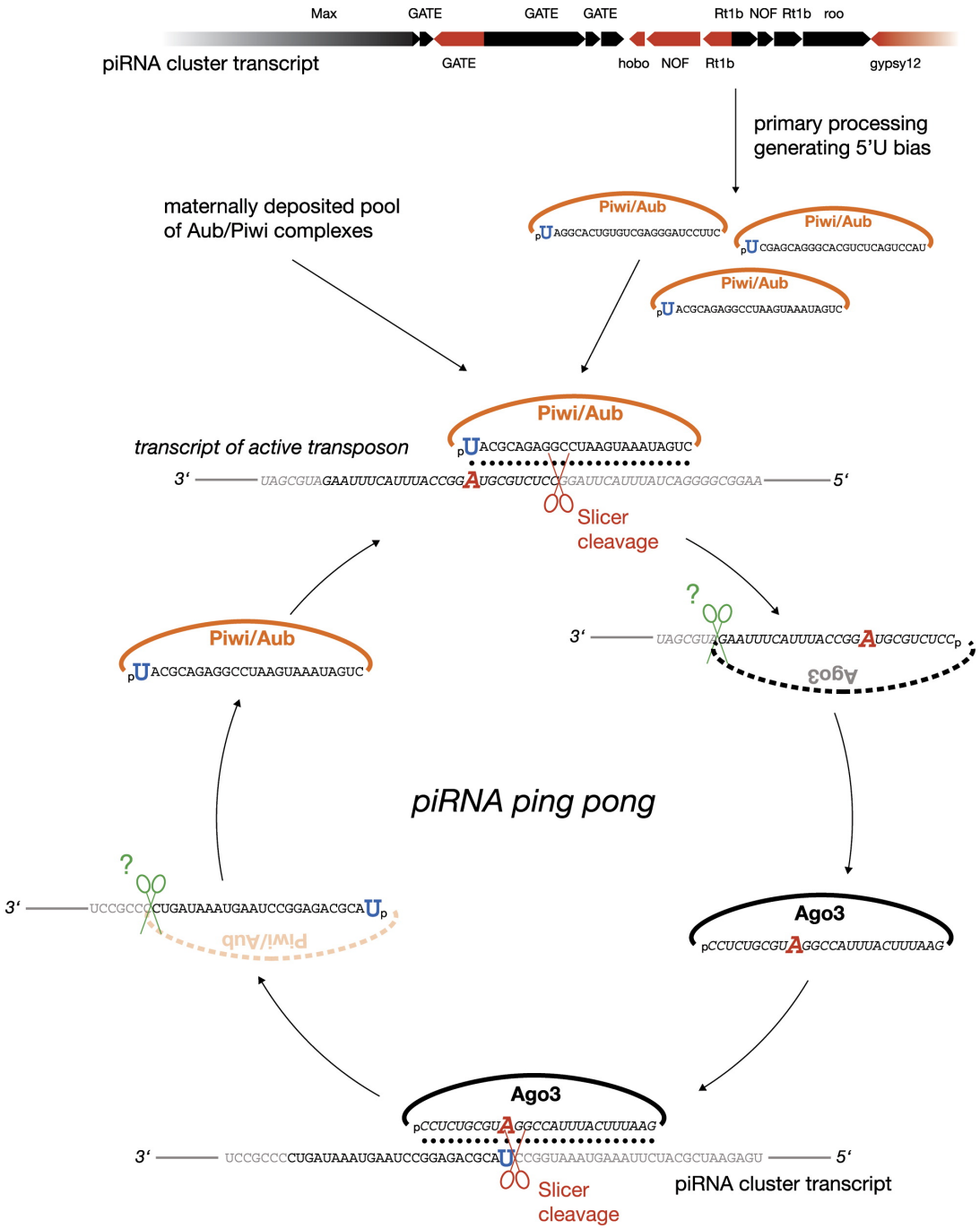


Figure 1.3: The piRNA Ping-Pong Model
Illustrated is the amplification loop consisting of Piwi/Aub complexes, Ago3 complexes, piRNA cluster transcripts, and transcripts of active transposons. Nucleotide cleavage events are shown as scissors. Potential sources of primary piRNAs are piRNA cluster transcripts and maternally inherited piRNA complexes.
Image and legend from: Brennecke et al. (2007).

et al., 2013). This TF also promotes expression of several key piRNA pathway components. To what extent this factor is conserved in flies remains an open question.

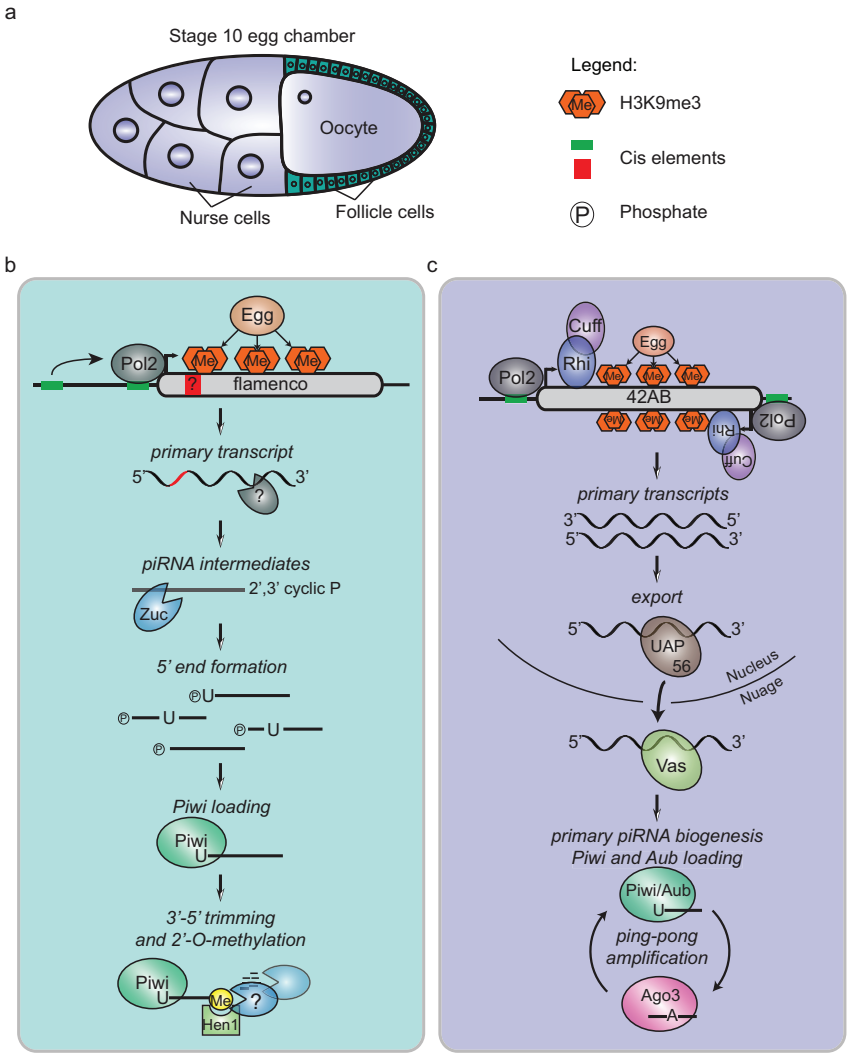


Figure 1.4: A model for piRNA biogenesis in the *Drosophila* ovary.
a) Two distinct piRNA pathways are active in a stage 10 egg chamber of the *Drosophila* ovary. The nurse cells that provide nutrients to the oocyte and the oocyte itself make up the germ cells of the ovary, shown in blue. The monolayer of somatic follicle cells surrounding the oocyte is shown in green. Nuclei are indicated as circles within each cell. b) In follicle cells, primary piRNAs arise from *flamenco* and are processed through a cascade of enzymatic cuts. Transcription by RNA polymerase II (Pol II) depends on deposition of Histone 3 Lysine 9 trimethyl marks (H3K9me3) by Eggless (Egg). Regulatory cis-acting elements, indicated as green boxes, upstream of the transcriptional start site could affect Pol II recruitment and transcription. Additionally, clusters could carry cis elements within themselves, shown in red, that affect downstream processing. After processing of the primary cluster transcript by unknown activities, piRNA intermediates are cleaved by the nuclease, Zucchini (Zuc). After 5' end formation, transcripts with a U at the first position are preferentially loaded into Piwi. Trimming activity, which could be carried out by redundant nucleases, shortens the transcript to its mature length. This process is coupled to 2'-O-methylation by Hen1.
Image and legend from: Guzzardo et al. (2013)

(*legend continued on next page*)

Despite their conventional appearance, piRNA cluster transcripts somehow escape the fate of their protein-coding counterparts, and instead are funneled towards the piRNA biogenesis machinery. Little is known about other factors essential for cluster transcription, except for the requirement of heterochromatin formation and histone 3 lysine 9 trimethylation along these loci through Eggless (Figure 1.4B-C) (Rangan et al., 2011). Interestingly, the heterochromatin protein Rhino (Rhi, a HP1 homolog) binds to this repressive chromatin mark and together with Cutoff (Cuff) is required for transcription of *42AB*, one of the most abundant piRNA clusters in germ cells of the *Drosophila* ovary (Figure 1.4C) (Chen et al., 2007; Klattenhoff et al., 2009; Pane et al., 2011). However, it is likely that there are additional factors that remain unknown, since the proteins mentioned above do not have an effect on uni-directionally transcribed piRNA clusters such as *flamenco*.

One protein that could potentially function at the interface of transcription of piRNA clusters and their export to the cytoplasm is UAP56, a nuclear DEAD box RNA helicase (Figure 1.4C) (Zhang et al., 2012). UAP56 associates with germ cell specific piRNA cluster transcripts and may be involved in a fate decision towards the piRNA biogenesis machinery located at the nuage. To what extent UAP56 also binds to somatic clusters remains enigmatic. After export to the cytoplasm, primary precursor transcripts are thought to be processed by a number of nucleolytic cleavage events in order to create a mature piRNA (Figure 1.4B) (Ishizu et al., 2012). Genetic, structural and biochemical data suggests that the RNA nuclease Zucchini is involved in the first cut, creating a mature 5' end (Haase et al., 2010; Ipsaro et al., 2012; Malone et al., 2009; Nishimasu et al., 2012; Olivieri et al., 2010; Pane et al., 2007; Saito et al., 2009, 2010; Voigt et al., 2012).

Following 5' end formation, piRNA intermediates are loaded into Piwi within cytoplasmic structures dedicated to the piRNA biogenesis machinery. In germ cells, this occurs in the previously mentioned perinuclear clouds (the nuage), whereas in follicle cells, the core components of biogenesis locate to Yb-bodies (Siomi et al., 2011). These foci are named after their main component: FS(1)YB. Other prominent members of the piRNA biogenesis pathway are present in Yb bodies, including Piwi itself, Armitage (ARMI, an RNA helicase), Vreteno (VRET, a TUDOR domain protein) and Shutdown (SHU, a co-chaperone) (Handler et al., 2011; Olivieri et al., 2012, 2010; Preall et al., 2012; Qi et al., 2011; Saito et al., 2010; Szakmary et al., 2009).

How Piwi loading is accomplished in the nuage is less clear. VRET interacts with two Tudor domain containing proteins, Brother of Yb and Sister of Yb, which are essential for primary piRNA biogenesis

Figure 1.4: (Continued from page 10.) c) The transcription of clusters in germ cells can occur bidirectionally. In addition to Egg, the HP1 homolog Rhino (Rhi) and Cutoff (Cuff) are essential for transcription. Subsequently, the helicase UAP56 binds the primary transcript and escorts it to the nuclear periphery. There, it is handed over to another RNA helicase Vasa (Vas) and arrives at its site of biogenesis, the nuage. After primary processing by similar machinery as in a), primary piRNAs are loaded into Piwi and Aub, and potentially Ago3. These primary piRNAs can be used to kick-start the ping-pong amplification cycle, which silences transposons post-transcriptionally. Image and legend from: Guzzardo et al. (2013).

in germ cells and localize to the nuage (Handler et al., 2011). This could be an indication that these proteins carry out similar functions to Yb in somatic follicle cells. Based on molecular phenotypes and mutant analysis, many other proteins have been implicated in biogenesis of mature piRNAs; however, their functions remain unknown (Ishizu et al., 2012). The situation is further complicated by the fact that AUB and AGO3 locate to the nuage, suggesting that this is also the site of secondary piRNA biogenesis through the ping-pong cycle. In any case, additional research is needed to elucidate the individual functions of these key players in order to separate Piwi loading from piRNA processing steps.

In vitro analysis on *Bombyx mori* lysates has shown that, after successful loading into Piwi, the 3' end of piRNA intermediates is further resected through an unknown exonuclease (Figure 1.4B) (Kawaoka et al., 2011). The trimmed RNAs are then 2'-O-methylated by HEN1 to generate mature piRNAs (Figure 1.4B) (Horwich et al., 2007; Saito et al., 2007). The newly assembled Piwi-RISC now enters the nucleus with the help of unknown factors. Interestingly, mutated Piwi lacking its N-terminal nuclear localization signal can be loaded with piRNAs, but is unable to silence TEs (Klenov et al., 2011; Saito et al., 2009, 2010). This observation strongly suggests that Piwi loading is indeed a cytoplasmic process. Moreover, the data also imply that Piwi mediated silencing is accomplished in the nucleus independently of its catalytic slicer domain (Saito et al., 2009, 2010). Three recent publications strengthen this hypothesis. Brennecke and colleagues showed that Piwi, together with Maelstrom (MAEL), is required for transcriptional gene silencing (TGS) of transposon loci (Sienski et al., 2012). Depletion of Piwi leads to loss of H3K9me3 marks over euchromatic transposon insertions, followed by recruitment of Pol II and increased transcriptional activity (Figure 1.5). Transcriptional activation is potentially linked to loss of HP1, which usually associates with H3K9me3 marks and was shown to physically interact with Piwi (Brower-Toland et al., 2007; Le Thomas et al., 2013). TGS is the major force in piRNA directed transposon silencing in follicle cells, given that these cells only express Piwi, but not AUB or AGO3. In germ cells, both TGS and PTGS differentially act to silence transposons (Rozhkov et al., 2013). Which classes of transposons and what mechanisms trigger either of the two silencing pathways remains largely unclear.

What are the missing links?

Even though the main principles of piRNA directed transposon silencing have been outlined, there are still many unanswered questions in the field. For example, the genetic identities of several key enzymes are unknown. These missing links may be key to our understanding of the molecular details of the pathway. For instance, it has become clear that Piwi-piRNA complexes can target transposon loci for transcriptional gene silencing through deposition of repressive histone marks (i.e. H3K9me3). Nonetheless, since Piwi is not a methyltransferase, there must be a cascade of proteins that act after target recognition by Piwi (guided by its bound piRNA) to silence the locus. A recent study suggests that Su(var)3-9 is involved in the deposition of methyl marks; however, the existing data leave space for a potential redundancy in this function, possibly involving EGG (Huang et al., 2013). Moreover, it remains unclear whether other histone modifications,

such as H3K4 de-methylation, play a role in transposon silencing too. A combination of genetic screens, evolutionary approaches, and structural and biochemical analyses of key components of the pathway may further our current understanding of piRNA-mediated genome defense. Some of the important challenges and future endeavors of the piRNA field are reviewed in Guzzardo et al. (2013).

1.1 Aims of this work

Aim 1: To understand how piRNA clusters act as a memory of transposon activity. The first aim of my thesis was to experimentally examine the molecular events that take place after *de novo* insertion of a sequence into a piRNA cluster. I investigated the following questions: Are piRNA clusters and the biogenesis machinery able to produce artificial piRNAs from *de novo* insertions, and if so, do these artificial piRNAs engage in secondary biogenesis mechanisms? What are the rules that govern piRNA production along precursor sequences, and have these rules evolved similarly in flies and mice? Lastly, are the triggers that determine piRNA production inherent in piRNA clusters, and is the genomic surrounding of critical importance? I used and developed a range of molecular techniques, transgenesis approaches and bioinformatic tools in order to attack these problems.

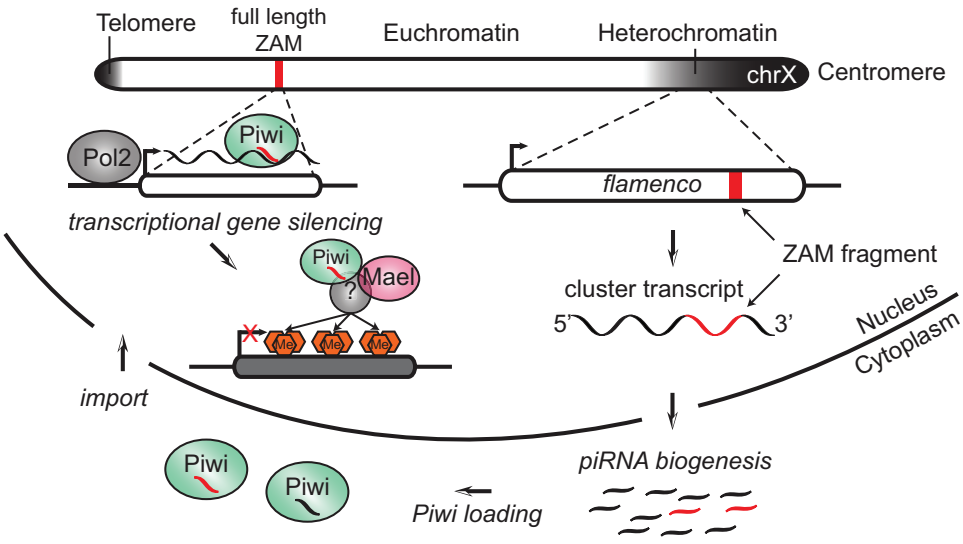


Figure 1.5: Transcriptional silencing of transposable elements by Piwi-piRNA complexes in the soma.
The X chromosome of *Drosophila melanogaster* (chrX) is shown. A simplistic view of its chromatin state is indicated in shades of gray. The transcriptionally active euchromatin in white harbors a full-length copy of the retroelement, ZAM (indicated as a red box). An inactive remnant of the same element (in red) can be found within the *flamenco* piRNA cluster in pericentromeric heterochromatin. After transcription and processing of *flamenco*, this fragment gives rise to antisense piRNAs that are loaded into Piwi in the cytoplasm (indicated as red piRNA species). Upon reimport into the nucleus, these Piwi-piRNA complexes recognize active transcription of the full-length *ZAM* copy by RNA polymerase II (Pol II) based on sequence complementarity. This recognition leads to the recruitment of additional factors such as Maelstrom (MAEL) and unknown chromatin remodelers. Ultimately, the deposition of H3K9me3 marks leads to loss of Pol II occupancy and the transcriptional silencing of *ZAM*.
Image and legend from: Guzzardo et al. (2013).

Aim 2: To uncover the genetic basis of transposon control in the *Drosophila* germline. The second aim of my thesis was to identify a comprehensive set of genetic components of transposon control in the *Drosophila* ovary. I focused on somatic follicle cells as an experimental system, in order to emphasize factors that may be involved in the primary piRNA pathway. Such factors could be part of both biogenesis and effector steps. By choosing an unbiased, reverse-genetic and genome-wide screening strategy, I intended to discover unforeseen links between already known components, or even pathways not yet implicated in transposon control. The results of this screen provide a foundation for the scientific community to build upon and shed light on some of the remaining mysteries of TE control and the piRNA pathway.

Chapter 2

Results

2.1 Production of artificial piRNAs in flies and mice

Original Article:
Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B., Gordon, A., Hannon, G.J., Aravin, A.A. (2012) Production of artificial piRNAs in flies and mice. *RNA*, 18 (1), pp. 42-52.
DOI: 10.1261/rna.029769.111.

2.1.1 Overview

In animals, small RNAs are used to identify and silence transposon-derived transcripts (Malone and Hannon, 2009). piRNAs in particular comprise a dynamic catalogue of all transposons a host population has experienced throughout its existence. When loaded into Piwi-clade Argonaute proteins, these Piwi-piRNA effector complexes can recognize active transposon transcription through sequence complementarity between the piRNA and a transposon mRNA (Brennecke et al., 2007; Gunawardane et al., 2007). Subsequently, they can trigger the destruction of the target mRNA (i.e. through PTGS) or silence the locus of origin altogether (i.e. through TGS) (Guzzardo et al., 2013).

piRNAs are derived from long, single-stranded precursor transcripts, originating from distinct genomic loci called piRNA clusters (Aravin et al., 2006, 2007; Brennecke et al., 2007). Each precursor transcript can give rise to thousands of individual piRNAs through a cascade of biogenesis steps (Ishizu et al., 2012). Our hypothesis at the time of publication of this manuscript was that the invasion of a host genome by a transposon would eventually lead to insertion of its sequence into a piRNA cluster. This insertion event would then lead to *de novo* piRNA production and consequently to resistance against the transposable element. Yet, little was known about the rules that govern the acquisition of new sequence

*These authors contributed equally to this work

information. We sought to experimentally test our hypothesis by inserting exogenous sequences into piRNA generating loci in mice and flies. According to our supposition, insertion should lead to *de novo* piRNA production of these artificial sequences. We further tested if the piRNA generation was dependent on the genomic context by introducing the tagged clusters into atypical (i.e. euchromatic) chromosomal locations.

By comparing flies and mice carrying tagged clusters we showed that artificial piRNAs could be produced irrespective of the origin and sequence of the tags. In addition, we demonstrated that both signals within clusters themselves, as well as surrounding and long-range sequence contexts influenced the generation of artificial piRNAs.

2.1.2 Results

To profile the molecular events that are triggered upon insertion of an exogenous sequence into a piRNA cluster, we took advantage of the PlArB transgene insertion into the subtelomeric TAS repeat on the X chromosome of *Drosophila melanogaster* (X-TAS) (Figure 2.1.1A) (Roche and Rio, 1998; Wilson et al., 1989). The PlArB transgene consists of fragments coming from endogenous *Drosophila melanogaster* genes (*hsp70*, *adh* and *rosy*), as well as a bacterial *lacZ* sequence. The X-TAS locus was previously shown to produce abundant piRNAs (Brennecke et al., 2007). We sought out to test if integration of PlArB into X-TAS led to *de novo* production of artificial piRNAs (apiRNAs) and if so, if these apiRNAs would exhibit molecular characteristics of bona-fide piRNAs. To this end, we sequenced small RNAs from ovaries of flies containing the transgene (P-1152 strain). Indeed, we saw production of small RNAs mapping to both genomic strands of PlArB, with most sequences mapping to the sense strand (Figure 2.1.1B). This closely resembled patterns of endogenous piRNAs mapping to the regions of X-TAS surrounding the transgene insertion (data not shown). Most small RNAs homologous to PlArB were 23 to 27nt long, a size profile commonly attributed to genuine piRNAs (Figure 2.1.1C). We did clone a small amount of RNAs 21nt in length, which presumably were part of the endo-siRNA fraction, a likely product of bi-directionally transcribed piRNA loci (Figure 2.1.1C) (Czech et al., 2008). Intriguingly, PlArB derived small RNAs exhibited a strong bias for a uridine at their 5' end and showed molecular signatures of secondary piRNA biogenesis (ping-pong signature, Figure 2.1.1D-E). It was previously shown that insertion of PlArB into X-TAS led to silencing of euchromatic lacZ transgenes (Ronsseray et al., 1991). We confirmed these observations, therefore implying that artificial piRNAs against *lacZ* are functional and able to silence (see Appendix A.2 for Supplemental Figure S1). One interesting observation we made, was that all parts of PlArB (i.e. *lacZ*, *adh*, *rosy* and *pBL*) produced apiRNAs, and did so to a similar extent (Figure 2.1.1B). The *adh* and *rosy* genes, which are present in wildtype *Drosophila melanogaster* in their native context, normally do not produce piRNAs. Thus, we concluded that all signals that lead to apiRNA production in P-1152 flies must lie within the surrounding piRNA cluster transcript.

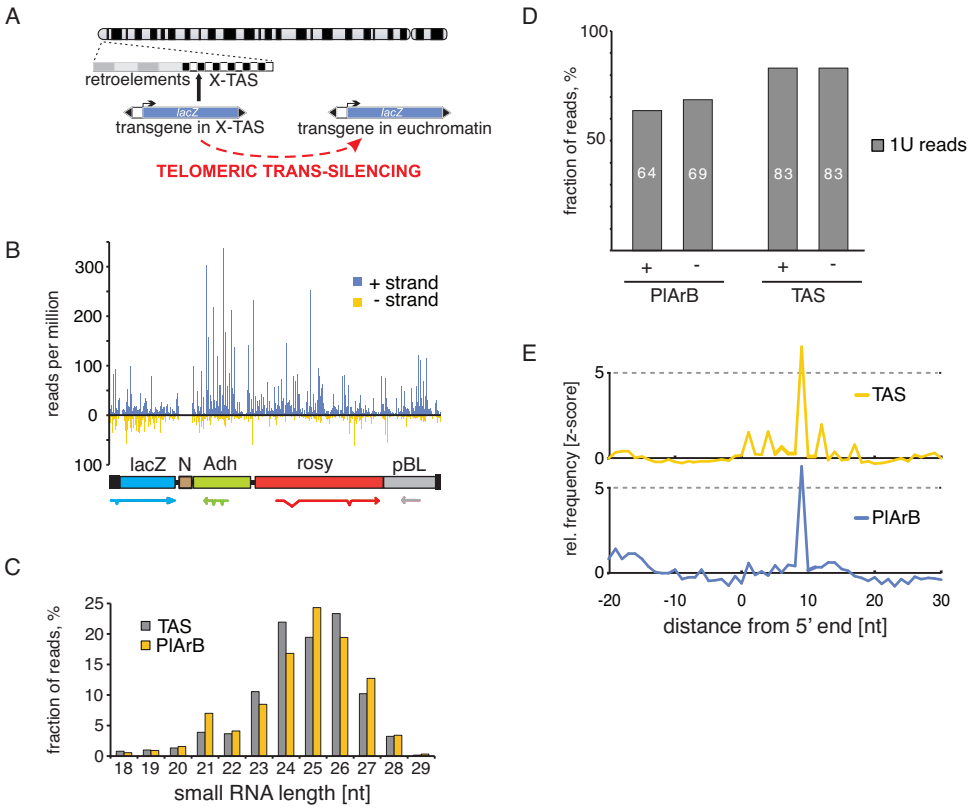


Figure 2.1.1: Production of artificial piRNAs (apiRNAs) from the *Drosophila* X-TAS cluster. (A) The PIArB insertion into the X-TAS cluster is shown schematically along with an illustration of trans-silencing. (B) Below is a schematic of the P{IArB} insert with the inferred structures of the transcripts it can produce (see text). N is an area where the sequence is unknown. Above is a plot of piRNA read frequencies along the plus and minus strands (indicated) of the element. (C) Small RNA lengths are plotted as a fraction of reads for TAS and for the inserted element. (D) Fractions of reads beginning with a 5' U are plotted for the P{IArB} and TAS plus and minus strands. (E) The degree of 5' overlap for piRNAs from the plus and minus strands for P{IArB} and TAS were quantified and plotted as relative frequencies (Z-scores). The spike at position 9 is a signature of the ping-pong amplification cycle.

To build upon these results, we decided to tag ectopic piRNA clusters in flies and mice with a *gfp* transgene insertion. This strategy allowed us not only to observe the production of anti-*gfp* artificial piRNAs, but also to test whether the native genomic context of the tagged clusters is essential for piRNA production. In flies, we used targeted integration to introduce a tagged version of *flamenco* into a euchromatic region on chromosome 3L of *Drosophila melanogaster* (Figure 2.1.2A). *flamenco* normally resides within pericentromeric heterochromatin of chromosome X. Because most piRNA clusters are found within such heterochromatic regions, we wanted to test if divorcing *flamenco* from its typical genomic context would interfere with production of artificial piRNAs against the tag. Small RNA sequencing from animals carrying one transgenic allele of *flamenco* revealed apiRNAs mapping to the entirety of the cassette, indicating that this was not the case (Figure 2.1.2C). These small RNAs further exhibited all molecular signatures that we would expect from small RNAs produced by the piRNA biogenesis machinery: we detected a strong bias for a 5' uridine and the typical size profile reminiscent of genuine piRNAs bound to Piwi (Fig-

ure 2.1.2E). It should be noted that apiRNA production from the cassette was only detected for one of the two genomic strands, which is in concordance with the strand bias reported for the endogenous *flamenco* locus (Brennecke et al., 2007).

In parallel, we tested for artificial piRNA production from the same tag in mice. We used random integration of an ectopic, bi-directionally transcribed piRNA cluster natively found on chromosome 17, and sequenced small RNAs from two independent transgenic lines carrying the tagged versions (Figure 2.1.2B). Similar to our findings in flies, the entirety of the cassette produced abundant small RNAs, which as a population exhibited the same strand bias as the native cluster (Figure 2.1.2D). Furthermore, we compared nucleotide biases of small RNAs mapping to the cassette with those mapping to a native, bi-directionally transcribed cluster on chromosome 9, and found the same preference for a 5' uridine (Figure 2.1.2F). While the anti-*gfp* apiRNAs in flies were virtually indistinguishable from genuine *flamenco* piRNAs, mouse apiRNAs showed slightly different length distributions compared to their endogenous counterparts: we cloned apiRNAs in sizes characteristic for MILI and MIWI-bound fractions, however, the ratio between the two changed drastically. RNAs with the size expected for MIWI-bound populations were much more abundant than the longer, MILI-bound species (Figure 2.1.2F). Why the artificially derived piRNA population exhibited this bias towards one of the two potential binding partners of the Piwi-family proteins remains an open question.

In conclusion, our data from two ectopic, tagged piRNA clusters in flies and mice demonstrate that these loci can be detached from their native genomic environment and still produce piRNAs. It has recently been shown that the heterochromatic context of piRNA clusters is indispensible for piRNA generation (Rangan et al., 2011). While our results do not contradict this finding, the signals that trigger heterochromatin formation seem to lie within the piRNA clusters themselves, rather than being inherited from the general genomic neighborhood.

We had sequenced small RNAs from *gfp* insertions in ectopic piRNA clusters in flies and mice, which gave us the opportunity to compare the patterns of artificial piRNA production along the same tag (i.e. the *gfp* coding sequence) but within different contexts. As mentioned above, artificial and genuine piRNAs exhibit a strong preference for a uridine at their 5' end. This already implies that piRNA production is not uniform along a given piRNA-producing transcript. Instead, certain positions are sampled more often than others, which is precisely what we observed for mouse and fly apiRNAs (Figure 2.1.3A). The top 1% of all possible *gfp* piRNAs from the ectopic *flamenco* insertion made up 19% of all *gfp* mapping reads (Figure 2.1.3B). The top 10% accounted for 70% of all reads. Furthermore, this bias towards certain positions was far from random: two independent lines of *gfp* insertions into ectopic mouse piRNA clusters favored the exact same positions along the *gfp* sequence and showed a high correlation of abundance of apiRNAs per position (R^2=0.99, Figure 2.1.3C). Nevertheless, the correlation between *gfp* processing in flies versus mice was far lower (R^2=0.01, Figure 2.1.3D). Intriguingly, this argues against the possibility that biases in our cloning procedure produced the high correlation we saw between the two mouse lines.

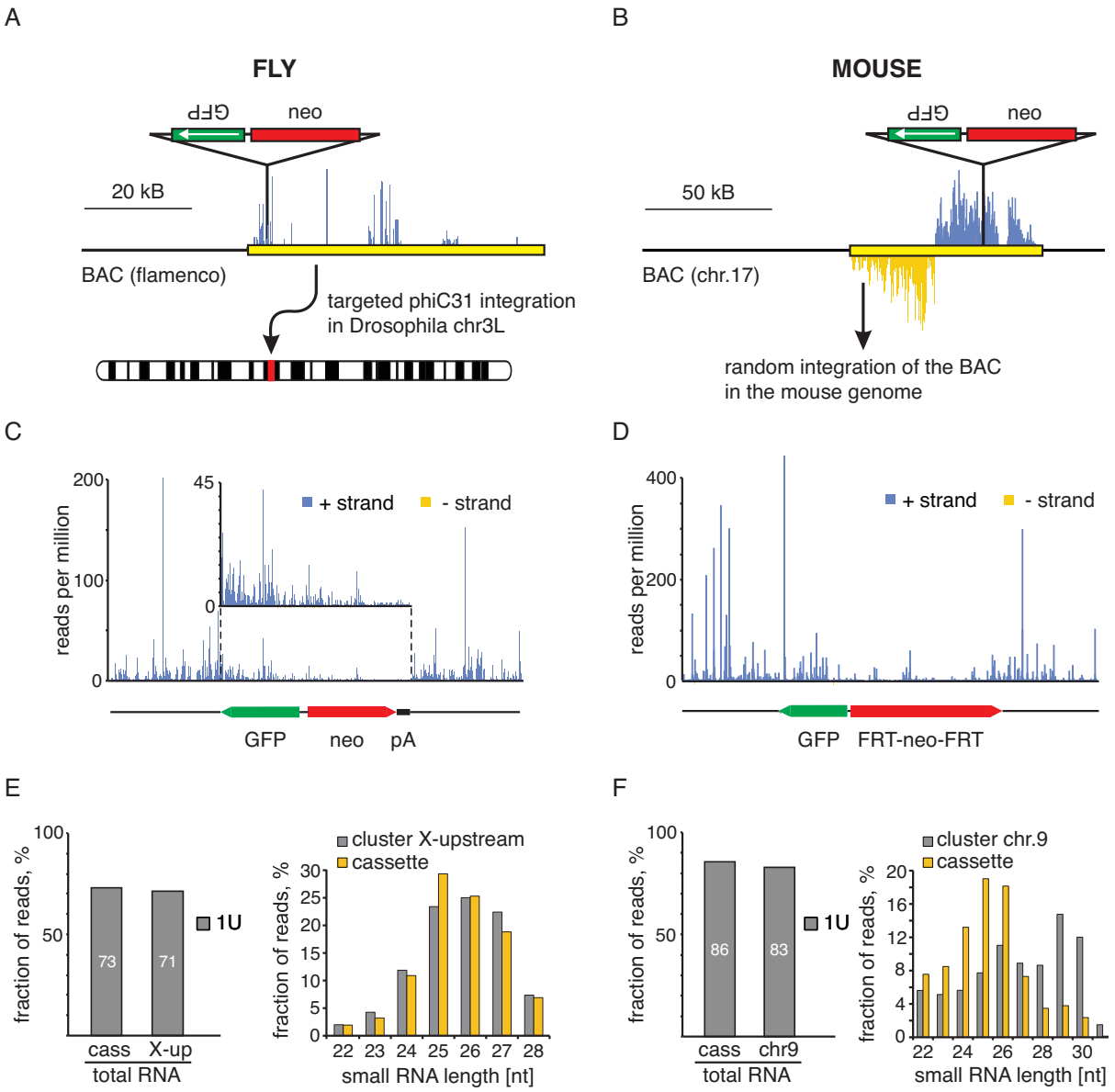


Figure 2.1.2: Generation of apiRNAs from ectopic clusters in flies and mice.
(A) A schematic representation of the GFP/Neo cassette is shown along a diagram of the *flamenco* locus (in yellow, piRNA densities in blue) in the BAC used for transgenesis. Below is a schematic indicating that the transgene is inserted into chromosome 3L. (B) The GFP/neo insertion into the mouse chromosome 17 cluster is diagrammed as in A. (C) The structure of the *flamenco* GFP/Neo insertion is diagrammed below a plot of piRNA frequencies along the insert on the plus and minus strands (indicated). For reference, piRNAs are also mapped to flanking regions, though these represent a mixture of RNAs derived from the two native and one ectopic *flamenco* cluster. (D) A scheme of the GFP/Neo insert into the mouse chromosome 17 cluster is shown below piRNAs mapping to the insert and its context as in C. Again, piRNAs that flank the insert can be derived from the two native or inserted ectopic loci. (E) The 1U bias (left) and size distributions (right) of apiRNAs from the ectopic *flamenco* cluster are compared with another piRNA cluster (X-upstream) that also produces piRNAs from one genomic strand in follicle cells. (F) As in E, apiRNAs from the ectopic ch17 cluster in mice are compared with a similarly structured cluster on chromosome 9.

Instead, it seems that local signals exist that determine what positions of the precursor RNA is processed preferentially, yet, the two experimental systems (i.e. the piRNA biogenesis machinery in flies and mice) seem to interpret these signals differently.

To discriminate local from long-range signals, we inserted the *gfp* coding sequence into an ectopic copy of the *traffic jam* 3' UTR, a locus which is expressed in follicle cells and produces abundant piRNAs (Figure 2.1.4A) (Saito et al., 2009). As with the other ectopic clusters, we cloned apiRNAs against the entire *gfp* sequence (Figure 2.1.4B), resembling the length and molecular signature of genuine piRNAs (Figure 2.1.4C). When we compared the processing patterns of *gfp* within *flamenco* versus *traffic jam*, the correlations of biogenesis patterns were considerably lower than between the two independent mouse lines (R^2=0.24, Figure 2.1.4D), implying that not only local signals play a role in creating biases towards certain positions. Nevertheless, the correlation was substantially higher than for patterns produced in flies vs. mice (R^2=0.01). In fact, uridine positions that produced abundant piRNAs in the *flamenco* transgene, almost always produced abundant piRNAs in the *traffic jam* transgene as well (Figure 2.1.4E). Conversely, positions that did not generate piRNAs from *flamenco* did not generate piRNAs from *traffic jam* either.

2.1.3 Discussion

Describing piRNAs as a catalogue of transposon sequence information is an elegant way of explaining both the inheritability and adaptability of the system. Nonetheless, the *de novo* production of piRNAs upon integration of their target into a piRNA cluster had not been vigorously tested at the time of publication of this manuscript. In an effort to do so, we sequenced small RNA populations of flies carrying transgenes inserted in native and ectopic piRNA producing loci. We detected *de novo* production of what we call artificial piRNAs, which were derived from transgenes, yet closely resembled the appearance of their native counterparts. In the case of *lacZ*-derived piRNAs, we also confirmed previous reports that these telemoric apiRNAs are capable of silencing a euchromatic insertion of *lacZ* (Roche and Rio, 1998; Ronsseray et al., 1998). One month after publication of our manuscript, Theurkauf and colleagues demonstrated that these findings extend to how flies actually adapt to transposon invasion (Khurana et al., 2011). In a *P-M* hybrid dysgenesis system, young flies are affected by the activity of the *P* element and activation of resident transposons, yet fertility is restored as these flies age. This can be attributed to transposon insertion into piRNA clusters, consequently *de novo* piRNA production and *P* element silencing.

While the overall biology behind the adaptation to transposon invasion seems clear given these results, the molecular events of piRNA biogenesis from long precursor transcripts are not well understood. Our observations of putative local and long-range signals, which inform the piRNA biogenesis machinery, point towards two conclusions: First, while piRNA clusters depend on heterochromatin formation at least for their transcription, they seem to have inherent signals that ensure that the proper epigenetic state is set, even when divorced from their native context. Given that maternal deposition of piRNAs bound to Piwi can lead to the emergence of novel piRNA producing loci (a process called paramutation), one could

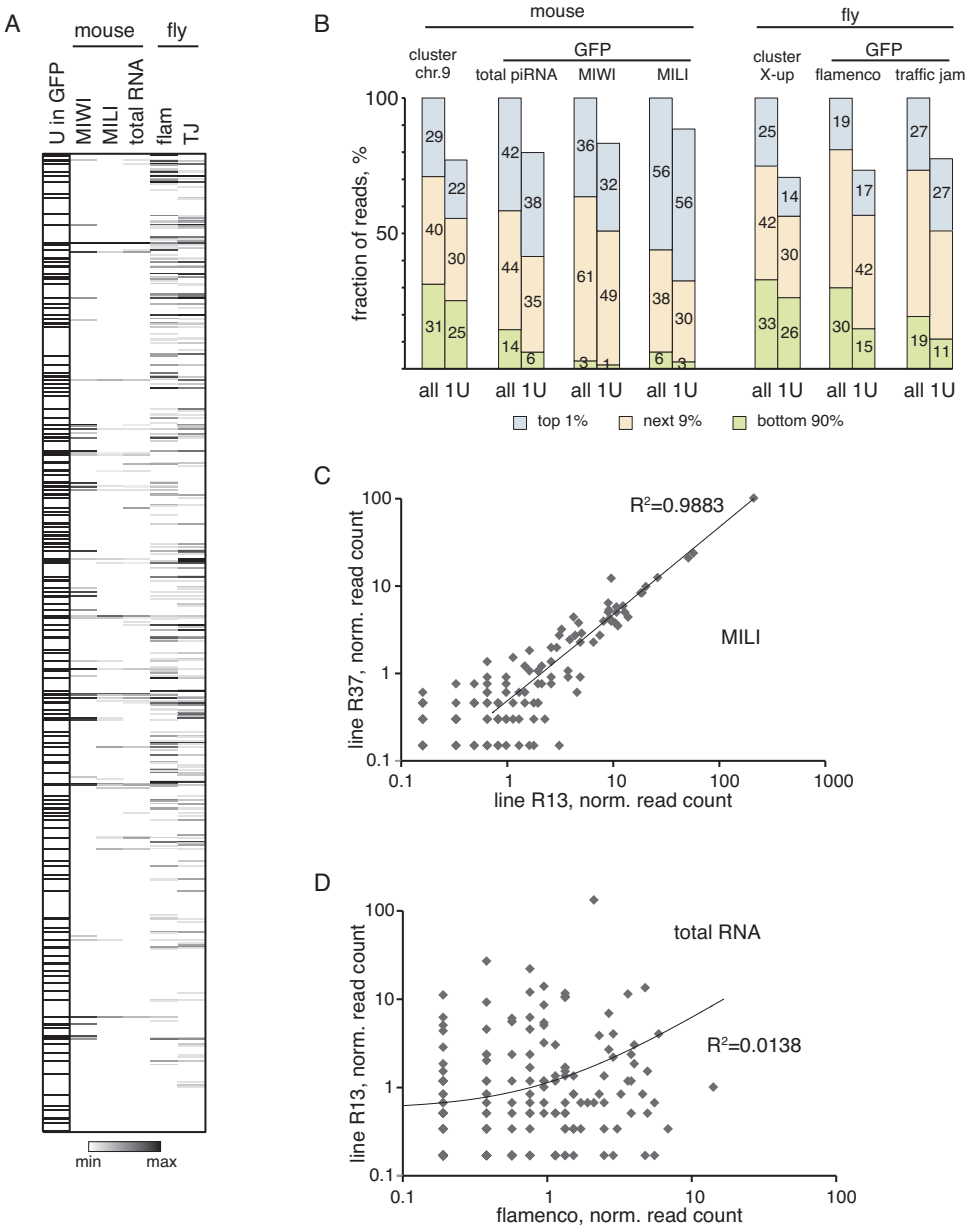


Figure 2.1.3: apiRNA production is not uniform along inserted sequences.
(A) A heatmap of piRNA abundance is displayed for all positions in the GFP insert carried in ectopic piRNA clusters as indicated. Sequence measurements were from total RNAs except in mouse, where MIWI and MILI immunoprecipitates (indicated) were also analyzed. The first column simply indicates U positions relative to the heatmaps. (B) All possible positions for piRNA production from GFP sequences inserted into ectopic clusters (all sites or only U positions, indicated) were ranked by their contribution to actual piRNA populations. The fraction of piRNAs contributed by the top 1%, the next 9%, or the remaining 90% were measured and indicated. Native clusters (indicated) were similarly analyzed for reference. (C) MILI-bound piRNAs were quantified by sequencing from two independent lines carrying the ectopic ch17 cluster. Correlations between read counts for GFP-derived piRNAs are shown. (D) A similar analysis was performed for GFP-derived piRNAs in total reads, comparing the R13 mouse line carrying the ectopic ch17 cluster and the fly strain carrying the ectopic *flamenco* cluster.

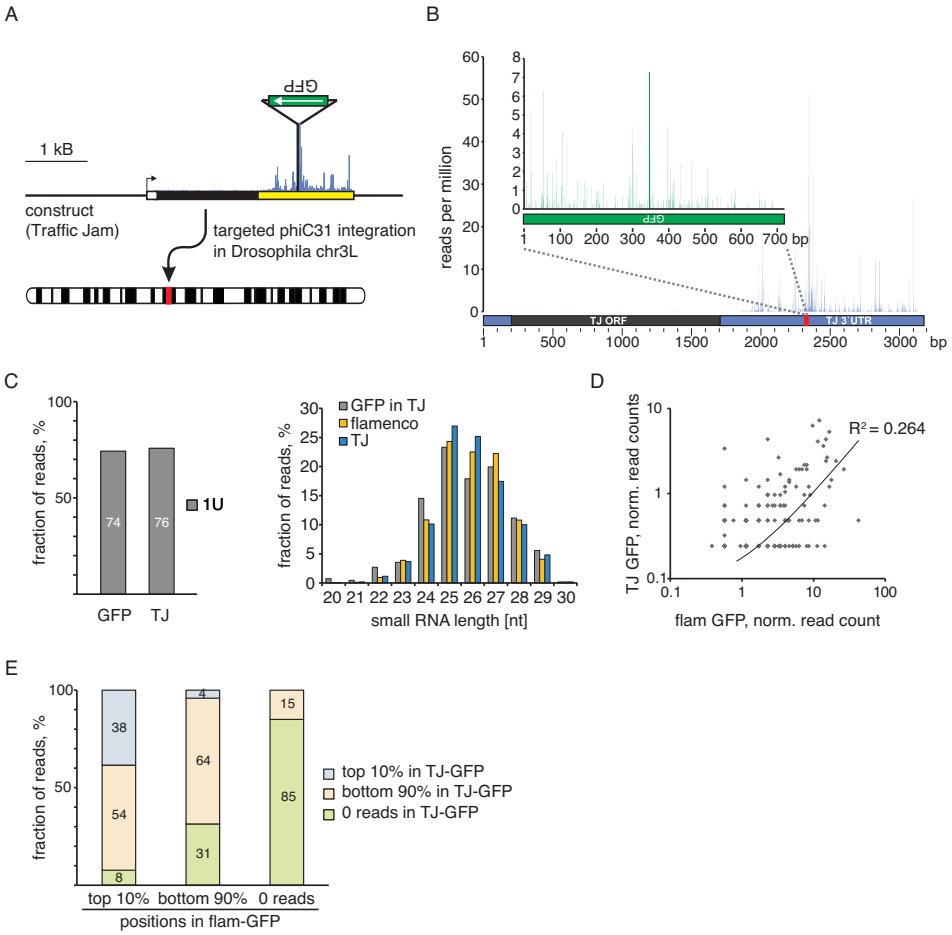


Figure 2.1.4: apiRNA production from the 3' UTR of traffic jam.
(A) A schematic of the GFP insertion into the 3' UTR of the *traffic jam* gene indicates the transcriptional start site (arrow), the coding sequence (black box), and the 3' UTR (yellow box). Below, a diagram indicates site-specific insertion into 3L. (B) piRNA read counts are plotted along the inserted GFP sequence (green inset) and the surrounding areas of the *tj* 3' UTR. Note that sequences mapping outside of GFP could be produced from the ectopic insert or the two endogenous copies of *tj*. (C) The 1U bias (left) and the size distribution of piRNAs mapping to the GFP insert are shown with reference to piRNAs from the *flamenco* cluster. (D) Normalized piRNA read counts were compared for the GFP insertions into the ectopic *flamenco* or *tj* piRNA clusters. (E) Read counts are calculated for all possible piRNAs that start with uridine derived from the GFP insertion into *flamenco*. These were divided into the top 10%, the next 90%, and the subset that contributed no reads. For each subset, the number that were present in the top 10%, the next 90%, or the non-contributors for the GFP insertion into *tj* were plotted.

speculate that these inherent signals lead to DNA binding by Piwi in an small RNA dependent manner, and subsequently to establishment of heterochromatin (de Vanssay et al., 2012).

Secondly, the preferential processing of certain positions within a piRNA-producing transcript cannot only be a simple reflection of loading biases of the Piwi protein (i.e. a bias towards 5' uridine). Non-uridine positions created abundant piRNAs from the *gfp* coding sequence, and these positions were identical in two independent mouse lines carrying the same ectopic piRNA clusters in presumably different genomic locations. This implies that the local sequence environment of each piRNA can change the accessibility of

the piRNA biogenesis machinery. In addition, the secondary or tertiary structure of the piRNA-producing transcript, a product of its sequence, seems to influence processing patterns as well. Until now, structural motifs that are shared between piRNA producing loci have escaped discovery, probably because of the difficulty of RNA structure prediction for such long transcripts. Hopefully, advances in methodologies for determining RNA structure will aid in improving our understanding of piRNA biogenesis.

Finally, it has not escaped our attention that these observations open up the possibility of using artificial piRNAs as a tool for gene silencing in a number of *in vivo* and *in vitro* systems. Taking any given target sequence and using ectopic drivers of piRNA production in euchromatic loci (e.g. through targeted integration), would make it possible to silence target loci at the transcriptional level, rather than through PTGS.

2.2 A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in *Drosophila*

Original Article:
Muerdter, F.*, Guzzardo, P.M.*, Gillis, J., Luo, Y., Yu, Y., Chen, C., Fekete, R., Hannon, G.J. (2013) A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in *Drosophila*. *Molecular Cell*, 50 (5), pp. 736–748.
DOI: 10.1016/j.molcel.2013.04.006.

2.2.1 Overview

Transposable elements are ubiquitous in virtually all known eukaryotic genomes. Their activity and presence can be an important source of inheritable variation, however, when unrestrained can threaten their host's genomic integrity. Especially within the germline genome, which gives rise to future generations, governing transposons in order to prevent insertional mutagenesis and double stranded DNA breaks is of paramount importance (McClintock, 1951).

In higher animals, the piRNA pathway is thought to enforce control over these elements (Malone and Hannon, 2009). However, many of the molecular details and key players of this pathway remain undefined. For example, what factors control the transcription of piRNA precursors and how these precursors are directed towards piRNA processing is unknown. Work by Zamore and colleagues provided some insight into this process when they identified an ancient transcription factor in mice, A-MYB, as the protein responsible for the transcriptional output of pachytene piRNA precursors as well as piRNA pathway factors themselves (Li et al., 2013). However, the orthologous function in flies remains elusive. The enzymes that process the precursors into mature sRNA are also unknown. A big step forward in this matter has been the characterization of an RNA nuclease, Zucchini, as the best candidate for 5' end maturation (Ipsaro et al., 2012; Nishimasu et al., 2012). However, what enzyme is responsible for subsequent trimming of the 3' end and if there are additional nucleolytic cuts remains undetermined.

Transposon silencing in somatic cells of the ovary is a nuclear phenomenon and occurs through transcriptional gene silencing (TGS) (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). While the principles of small RNA directed heterochromatin formation is well established in fission yeast (Cam, 2010), proteins that are responsible for chromatin remodeling at target loci after recognition through Piwi are yet to be identified in *Drosophila melanogaster*.

*These authors contributed equally to this work

Rather than focusing on individual open questions following a hypothesis driven approach, we decided to perform an unbiased, hypothesis generating genome-wide screen. We used RNAi knockdown strategies to identify genetic factors responsible for transposon control in OSS cells *in vitro*, and validated the identified hit candidates *in vivo*. Knockdown of 87 genes in follicle cells of the *Drosophila* ovary led to derepression of transposon transcription and in many cases to sterility. We found several candidates to be essential for biogenesis of primary piRNAs, whereas other genes did not have an effect on piRNA populations while still exhibiting the same derepression phenotypes. One such factor was *CG3893*. Loss of this gene led to decreases in H3K9me3 marks on certain transposons resulting in their mobilization and subsequently sterility of the affected flies.

2.2.2 Results

An RNAi screen for elements of the somatic piRNA pathway

One molecular phenotype of gonads defective in the piRNA pathway is massive transcriptional activation of transposable elements. For example, flies with mutant alleles of *flamenco* exhibit much higher levels of *gypsy* full length mRNA (Pelisson et al., 1994). Interestingly, *gypsy* additionally encodes for a subgenomic transcript, giving rise to its envelope protein. The output of this subgenomic transcript is very sensitive to disruption of the piRNA pathway, even more so than its full-length counterpart (Pelisson et al., 1994). With the aim of discovering novel components of the piRNA pathway and genes needed for transposon control, our strategy was to knock down each gene in the *Drosophila* genome transfecting OSS cells with long dsRNAs and subsequently test for *gypsy* subgenomic transcript derepression (Figure 2.2.1A). Assaying for a transcript with a 5kB intron allowed us to use crude lysates from transfected cells as direct input into reverse transcription, without the need for complete removal of genomic DNA. Knockdown of *piwi* in this setting led to derepression of the *gypsy* subgenomic transcript by up to 70-fold (see Appendix A.2 for all screen results and supplemental information). We made use of two genome-wide libraries totaling in 41,342 dsRNAs, targeting 13,914 genes with valid Flybase IDs (McQuilton et al., 2012). After analysis of all primary qPCR results, 33,780 dsRNAs met our criteria for inclusion in further analysis. Out of these, 320 dsRNAs made the cutoff for primary hit candidate selection (Figure 2.2.1B). All genes previously implicated in the control of *gypsy* were part of this list.

In vivo validation of primary hits

We attempted to validate our primary hit candidates *in vivo* utilizing long hpRNAs from the Vienna *Drosophila* RNAi Center collection (Figure 2.2.1A) (Dietzl et al., 2007). We were able to obtain flies harboring hairpins against 288 of our hit candidates and crossed them to virgin females expressing a *Gal4* driver under the follicle-cell specific *traffic jam* promoter. Out of all tested primary hit candidates, 87 genes could be validated using this approach (Figure 2.2.2A). Knockdown of 52 genes led to severe developmental defects,

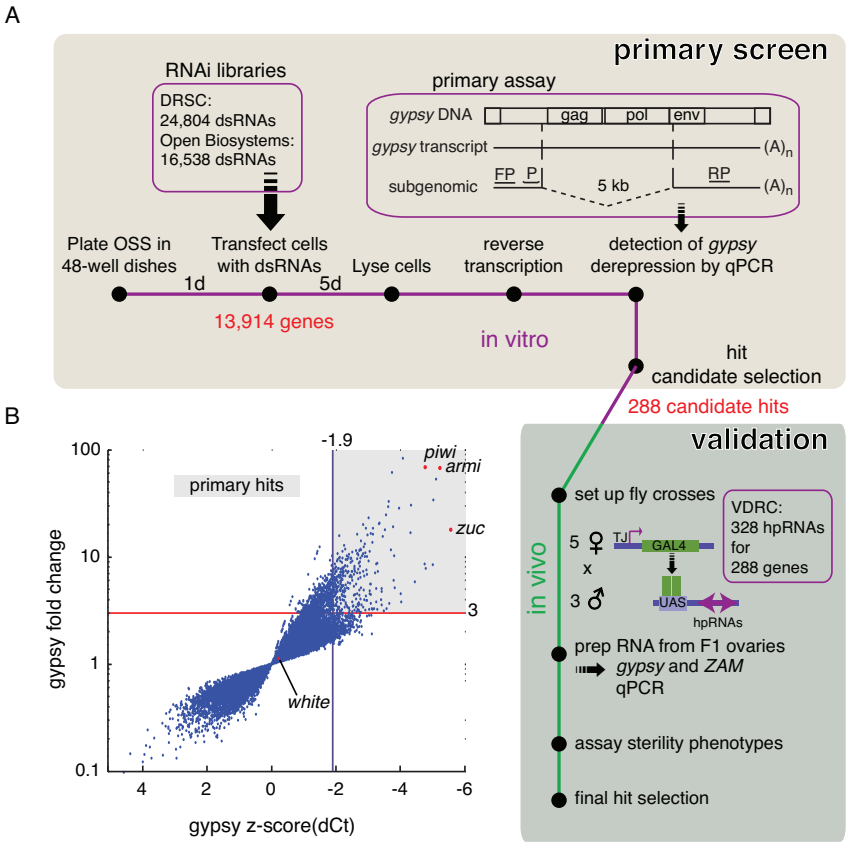


Figure 2.2.1: A Genome-wide RNAi Screen for piRNA Pathway Components Acting in the Somatic Compartment of *Drosophila* Ovaries.
(A) A workflow of the primary RNAi screen in ovarian somatic sheet (OSS) cells and validation of primary hit candidates *in vivo* is shown. Each gene in the *Drosophila* genome was knocked down with one or more dsRNAs. At 5 days after transfection, cells were tested for increased levels of the *gypsy* retrotransposon. The primers and the hydrolysis probe used for the qPCR are shown (FP, forward primer; P, hydrolysis probe; RP, reverse primer). The dashed line indicates the ~5 kb segment not present in the subgenomic transcript. We further tested 288 genes *in vivo* using the *Gal4*/UAS system to drive hairpin RNAs (hpRNAs) within the *traffic jam* (TJ) expression domain. (B) All transfected wells were assayed for levels of *gypsy* and one reference gene for normalization. Levels of *gypsy* expression are displayed as Z scores and fold change. The cutoffs for both Z score (<-1.9) and fold change (>3) are indicated as red lines. The shaded area shows the selection of primary hit candidates. Three positive controls (*piwi*, *armi*, *zuc*) and one negative control (*white*) are marked as red dots. Only wells that passed the filter for primary data point selection are shown.

rendering *in vivo* validation technically infeasible. The validation categories exhibited some remarkable properties: We saw a strong enrichment for genes preferentially expressed in ovarian tissues, both for validated genes and for the developmental defect category (Figure 2.2.2B). Furthermore, knockdown of many of these genes led to sterility in daughter generations: 16% of all validated strains produced no fertile offspring (Figure 2.2.2C).

When looking at the top 20 validated hits, an expected yet remarkable observation was that all known piRNA pathway components except for *piwi* (which was part of the developmental defects category) scored strongly (Table 2.2.1). This further supports our current hypothesis that the piRNA pathway is the major control mechanism of transposon expression in gonads (Malone and Hannon, 2009).

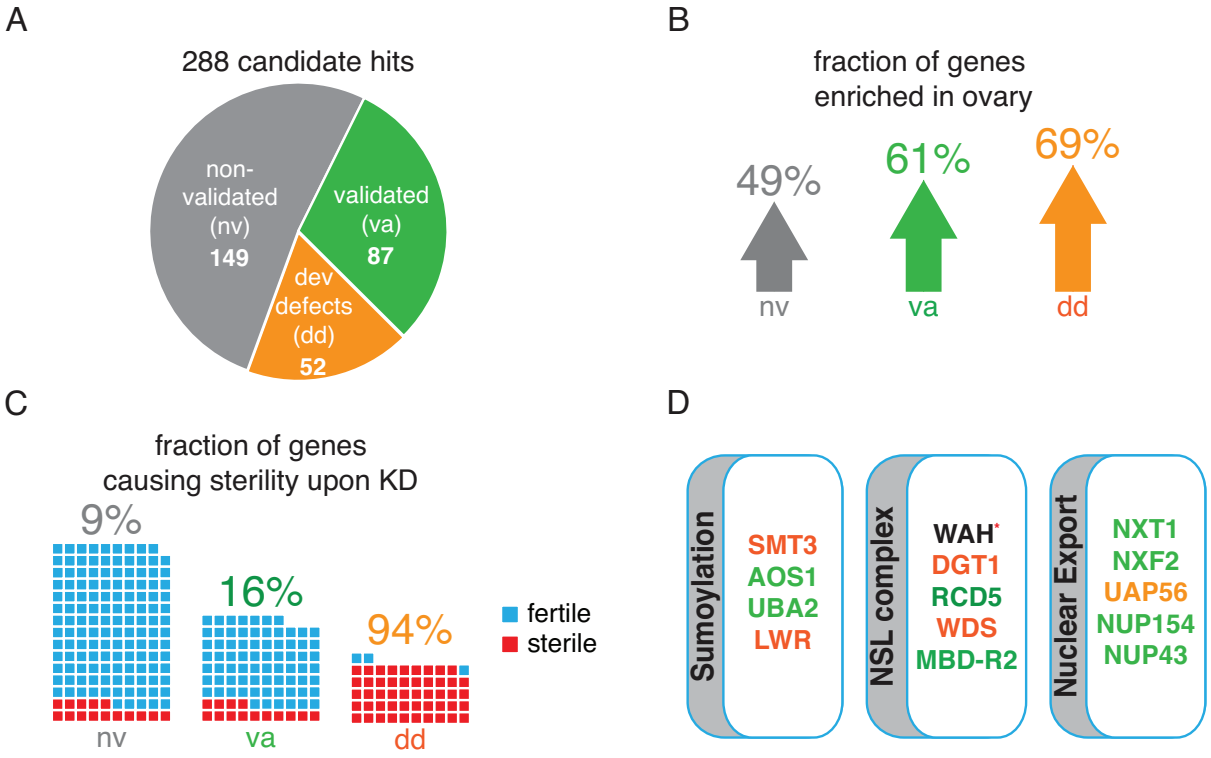


Figure 2.2.2: Primary Candidates Were Validated *In Vivo*.
(A) The number of hit candidates that validated (va) or did not validate (nv) *in vivo* is shown. Genes that caused severe developmental defects upon knockdown and therefore could not be assayed are also indicated (dd). (B) Validated hits are preferentially expressed in ovaries. The percentage of genes that are enriched in ovaries compared to whole fly is shown for the three categories. These data are based on mRNA signals on Affymetrix expression arrays available from FlyAtlas (Chintapalli et al., 2007). (C) The fraction of genes causing sterility upon knockdown is shown. Each small box represents one gene, with blue and red indicating if flies were fertile or sterile, respectively, upon knockdown. (D) Components of the *Drosophila* sumoylation pathway, the nonspecific lethal complex, and proteins involved in nuclear export are primary hits that validate *in vivo*. WAH could not be validated *in vivo* because no RNAi fly was available at the time of submission (red asterisk). The text coloring of each gene indicates the result of the validation screen and is consistent with the categories in (A).

Table 2.2.1: Top 20 Validated Hits

Symbol	Primary screen fold change	Validation screen fold change			Fertility	Comments
		Gypsy	*ZAM*	*Gypsy3*		
nxt1	2	2452	3566	41	-	mRNA export from nucleus
fs(1)Yb	11	96	700	335	+	piRNA pathway component
armi	48	197	846	112	+	piRNA pathway component
zuc	19	809	549	9	+	piRNA pathway component
vret	4	74	315	22	+	piRNA pathway component
CG3893	42	80	207	10	+	CHHC zinc fingers
mael	3	159	452	16	+	piRNA pathway component
CG2183	4	173	158	11	+	Fly homolog of GASZ
lin-52	5	153	153	8	+	dREAM complex subunit
MBD-R2	16	85	48	4	+	NSL complex subunit
uba2	3	84	12	8	+	Sumoylation E1 ligase

Table 2.2.1: Top 20 Validated Hits (continued)

Symbol	Primary screen fold change	Validation screen fold change			Fertility	Comments
		Gypsy	*ZAM*	*Gypsy3*		
CG9754	2	26	61	14	+	No conserved domains
wde	3	40	120	12	+	Co-factor of Eggless
nup154	6	30	186	3	-	nuclear pore
Su(var)2-10	2	9	20	4	-	dPIAS, putative SUMO E3 ligase
dlg1	2	16	2	1	+	Guanylate kinase
shu	7	14	416	1	+	piRNA pathway component
CG4686	4	13	1	1	+	Part of ribokinase/pfkB
nup43	3	12	3	1	+	WD40-repeat-containing domain
cchl	0	12	1	2	-	Cytochrome C heme lyase

Another emerging pattern within the validated hits was the presence of several nuclear export factors and nucleoporins. *nxt1*, *nxf2*, *nup154* and *nup43*, which all act in nuclear RNA export pathways, were confirmed *in vivo* (Figure 2.2.2D) (Herold et al., 2001; Lévesque et al., 2001).

Two other validated genes scoring strongly were as of yet uncharacterized: *CG3893* and *CG2183*. CG3893 shows sequence and structural homology to mammalian Gtsf1, a germline specific factor implicated in transposon control in mice (Yoshimura et al., 2009). CG2183 is predicted to be the ortholog of GASZ, which was previously implicated in retrotransposon repression in the male mouse germline (Ma et al., 2009).

windei (*wde*), which is a cofactor of *eggless* (*egg*), was another strong hit that validated *in vivo*. Eggless, a H3K9 methyltransferase, is essential for piRNA cluster transcription in the *Drosophila* germline (Koch et al., 2009; Rangan et al., 2011). If *wde* has a function independent of *egg* remains to be seen, given that *egg* was not a hit in the primary screen.

Zamore and colleagues recently demonstrated the involvement of A-MYB in transcriptional regulation of both piRNA cluster transcript as well as piRNA pathway components (Li et al., 2013). Another strong hit that validated *in vivo*, *lin-52*, is a member of the **D**rosophila **R**BF, **E**2F, and **M**yb-interacting proteins complex (dREAM complex) and co-purifies with *Drosophila* Myb complex components, therefore making it a likely candidate for an orthologous function in *Drosophila* (Lewis et al., 2004).

Several other genes implicated in transcriptional regulation besides *lin-52* were hits. All members of the **N**on-**S**pecific **L**ethal complex (NSL complex) except for *rcd1* could be identified as outliers in the primary screen, though only *rcd5* and *MBD-R2* were subsequently validated *in vivo* (Figure 2.2.2D) (Raja et al., 2010).

Characterization of newly described piRNA pathway components

To gain a deeper understanding of the mechanism by which some of the top validated hits control transposon expression, we analyzed transposon mRNA levels and small RNA populations (i.e. piRNAs) on a global scale. To this end we constructed RNA-seq and sRNA-seq libraries from ovaries of *tj*-Gal4 knockdown of five of the top 20 scoring genes (Figure 2.2.3A-D). Overall, our analysis of TE mRNA levels strikingly resembled our previous qPCR results: knockdown of *CG3893*, *nxt1*, *uba2*, *wde* and *lin-52* led to derepression of soma specific TEs to a similar extent as knockdown of known piRNA pathway components (*armi*, a biogenesis factor and *mael*, a gene involved in piRNA directed TGS) (Figure 2.2.3A). At the same time, transposons that were expected to be germline dominant in their expression bias, and therefore should not change expression levels in a follicle-cell specific knockdown of target genes, behaved similar to a negative control knockdown (*aub*).

One way to place each of the above genes at a certain step within the piRNA pathway was to examine mature piRNA populations. Knockdown of effector proteins such as *mael* should strongly affect transposon mRNA levels, as shown by RNA-seq, but should not impact mature piRNA levels (Sienski et al., 2012). Knockdown of a canonical biogenesis factor such as *armi*, should lead to strong effects on piRNA populations while similarly causing transposon derepression (Saito et al., 2010). Disruption of *CG3893* resulted in patterns resembling a *mael* knockdown (Figure 2.2.3B-D). Thus, this gene can hypothetically be placed at the effector step of transposon silencing. *nxt1*, *uba2* and to a lesser extent *wde* and *lin-52* knockdowns exhibited patterns closer to *armi*, and may possibly occupy biogenesis functions.

CG3839 is indispensible for transposon silencing in the germline

Because *CG3893* knockdown resulted in molecular phenotypes similar to *mael* knockdowns, we hypothesized this gene may be involved in the effector step of the pathway. CG3893, a ~20kD protein, is a member of a protein family with unknown function (UPF0224). The only structural characteristic unifying all members of this family is a set of two highly conserved zinc finger domains at the N-terminus (Figure 2.2.4A). All members of this family show weak expression in *Drosophila* gonads (Figure 2.2.4B), but only knockdown of *CG3893* led to substantial *gypsy* derepression in OSS cells (Figure 2.2.4C).

For a more complete picture of its function, we analyzed a mutant allele of *CG3893* in which a *P* element transgene is inserted in the 5' UTR disrupting *CG3893* expression ($CG3893^{[204406]}$). Homozygous females for this insertion showed severe defects in ovarian morphology, reminiscent of, albeit not as severe as, *piwi* null mutant females (Figure 2.2.4D) (Lin and Spradling, 1997).

Transposon control through TGS in somatic cells of the ovary is a nuclear process (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). Intriguingly, we found CG3893 protein tagged with GFP to localize to the nucleus when overexpressed in OSS cells (Figure 2.2.4E). Furthermore, RNA-seq and sRNA-seq of RNA extracted from females heterozygous and homozygous for $CG3893^{[204406]}$ confirmed

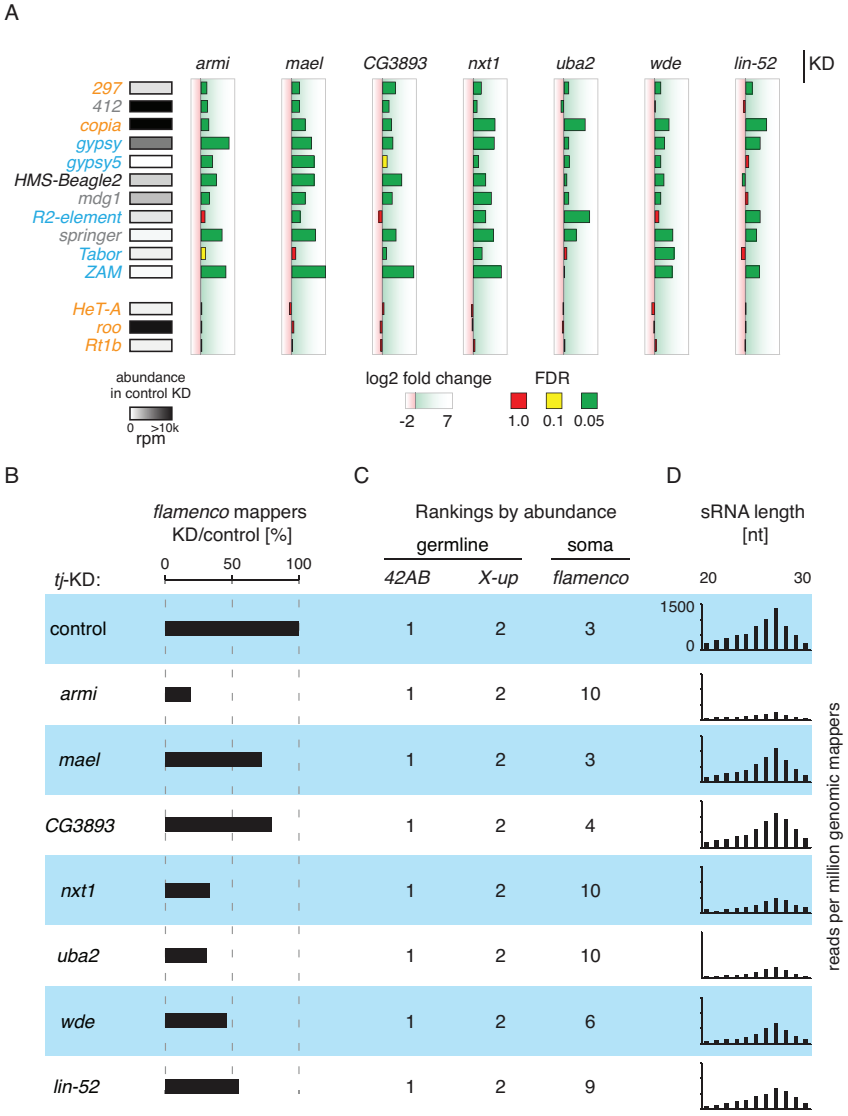


Figure 2.2.3: RNA-Seq and sRNA-seq Shows Changes in Transposon Expression and *flamenco* derived piRNAs upon Knock-down of Top Candidates In Vivo.
(A) A subset of somatically expressed transposons is derepressed in the indicated KD. The classification of transposons according to Malone et al. (2009) is indicated in orange (germline dominant), gray (intermediate), blue (soma dominant), and black (unclassified). The absolute abundance of reads in control knockdown mapping to each transposon is shown in shades of gray. The log2 fold change of each target gene versus a negative control (*aub*) is shown. Color of the bars represents the significance of these fold changes and is indicated as an adjusted p value (FDR). Green indicates highly significant differences ($p \leq 0.05$), yellow indicates moderately significant changes ($0.05 < p \leq 0.1$), and red indicates non-significant changes ($0.1 < p \leq 1$) based on two biological replicates. Each knockdown is normalized to *aub* knockdown controls from their corresponding library (GD or KK). (B) Percentages of total unique mappers (sense species, >23 nt) to *flamenco* in each knockdown (as indicated) in relation to the control knockdown are shown. (C) The internal rankings for three representative piRNA clusters based on their representation in piRNA populations are displayed. Expression bias toward either domain (soma or germline) is indicated. Cluster definitions are in concordance with Brennecke et al. (2007). (D) The size profiles of piRNAs mapping in sense orientation to *flamenco* in each knockdown (as in [B]) are plotted as total read count per million genomic mappers.

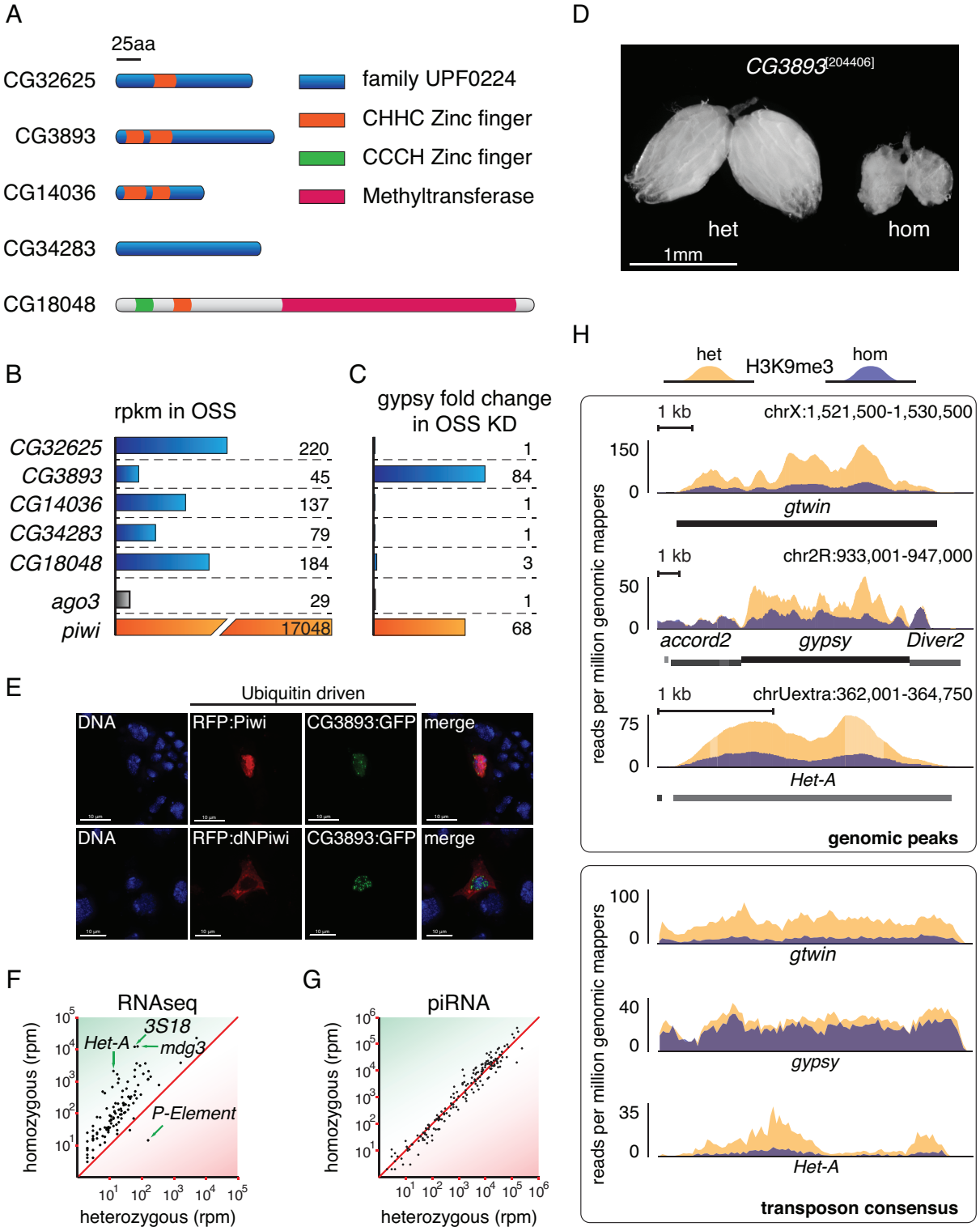


Figure 2.2.4: Disruption of CG3893 Function Has a Severe Impact on Transposon Silencing.
(*legend continued on next page*)

our initial results from RNAi experiments: all classes of transposon colonizing the *Drosophila melanogaster* genome were massively upregulated in homozygous animals when compared to their heterozygous sisters, while mature piRNA populations remained unchanged (Figure 2.2.4F-G). Because piRNA directed TGS is thought to act through the deposition of H3K9me3 marks over transposable element loci, we tested genome-wide patterns of this mark in heterozygous and homozygous animals through ChIP-seq. And indeed, we saw a substantial drop of H3K9me3 over certain classes of transposons (i.e. retroelements) (Figure 2.2.4H). This observation was restricted to full-length insertions and both somatic and germline dominant TE classes were equally affected. Because of its small size, yet powerful role in transposon control, we decided to name *CG3893 asterix* (*arx*).

2.2.3 Discussion

There are several ways to tackle open questions in the piRNA field. One important route is to dissect each step of primary piRNA biogenesis with biochemical experiments *in vitro*. Yet, to do so, all proteins involved in these steps have to be known. Another way to gain insight in piRNA directed silencing mechanisms is to screen collections of mutagenized fly strains for phenotypes indicative of a disrupted piRNA pathway (i.e. transposon derepression, fused dorsal appendages or sterility). However, even after successfully finding such phenotypes, one has to identify the underlying genotype using laborious recombination mapping strategies.

Here, we took an unbiased, reverse-genetic approach to identify the genetic framework for transposon control, pinpoint key enzymatic players and generate new and unexpected hypotheses of how Piwi-piRNAs

Figure 2.2.4: (Continued from page 32.) (A) The five members of the *Drosophila* uncharacterized protein family UPF0224 and their domain structures are diagrammed. The conserved domains are highlighted as colored boxes. (B) All five family members are weakly expressed in OSS cells. *piwi* and *ago3* expression levels are shown for comparison. Expression levels are based on the modENCODE cell line expression data and are displayed as reads per kilobase per million mapped reads (rpkm). (C) CG3893, but no other members of its protein family, has a strong impact on transposon silencing upon knockdown in OSS cells. Effects of knockdown of *ago3* and *piwi* are shown for comparison. Numbers represent fold changes of *gypsy* levels with respect to the median fold change of the corresponding plate in the primary screen. (D) The ovarian morphology of flies heterozygous or homozygous for a P element insertion in *CG3893* is shown (204406, Kyoto DGRC). (E) Tagged CG3893 colocalizes with Piwi in the nucleus of OSS cells when overexpressed in transient transfections. Nuclear Hoechst staining is blue, GFP-tagged CG3893 is green, and red fluorescent protein (RFP)-tagged Piwi or DNT-Piwi is shown in red (Saito et al., 2009). (F) Transposons are highly upregulated upon disruption of *CG3893* in the P element insertion line. A scatter plot of reads per million (rpm) is shown for RNA-seq of heterozygous versus homozygous flies. Each dot represents one transposon consensus sequence. Only sequences mapping in the sense orientation are taken into account. (G) piRNA levels are not affected by CG3893 disruption. The number of piRNA reads mapped to the same transposon consensus sequences as in (F) is expressed in reads per million. (H) Levels of H3K9me3 decrease dramatically on a subset of transposons upon depletion of CG3893. Density plots for normalized H3K9me3 ChIP-seq reads over three transposons, *gtwin*, *gypsy*, and *Het-A*, are shown. Yellow distributions correspond to levels in heterozygous flies, and blue distributions correspond to the homozygous state. The upper box shows three distinct genomic peaks over transposon insertions; the lower box shows the corresponding consensus sequences.

complexes function. To make this resource accessible to the scientific community, we developed a web-based query platform with all primary and validation data points (somatic-pirnascreen.cancan.cshl.edu). Among the strongest hits of our RNAi screen, we find likely candidates for many of the gaps in our understanding of the piRNA pathway. For example, with Lin-52 and the NSL complex, we uncovered proteins putatively affecting transcription of primary piRNA precursors as well as key pathway components (Figure 2.2.5). Nuclear export factors such as NXT1, NXF2 or the identified nucleoporins could be responsible for subsequent export of these precursors into the cytoplasm and therefore the sites of further enzymatic processing. We strengthen the current hypothesis that ZUC is responsible for creating the 5' end of mature piRNAs, given that no other annotated endonucleases could be validated *in vivo*. With members of the exosome (RRP6 and CSL4), we uncover two exonucleases that seem to be essential for proper transposon silencing. Unfortunately, knockdown of these essential factors *in vivo* led to severe developmental defects, rendering their validation technically challenging. Whether these proteins exert their function co-transcriptionally at TE loci, or through potential involvement in piRNA 3' end formation remains to be seen.

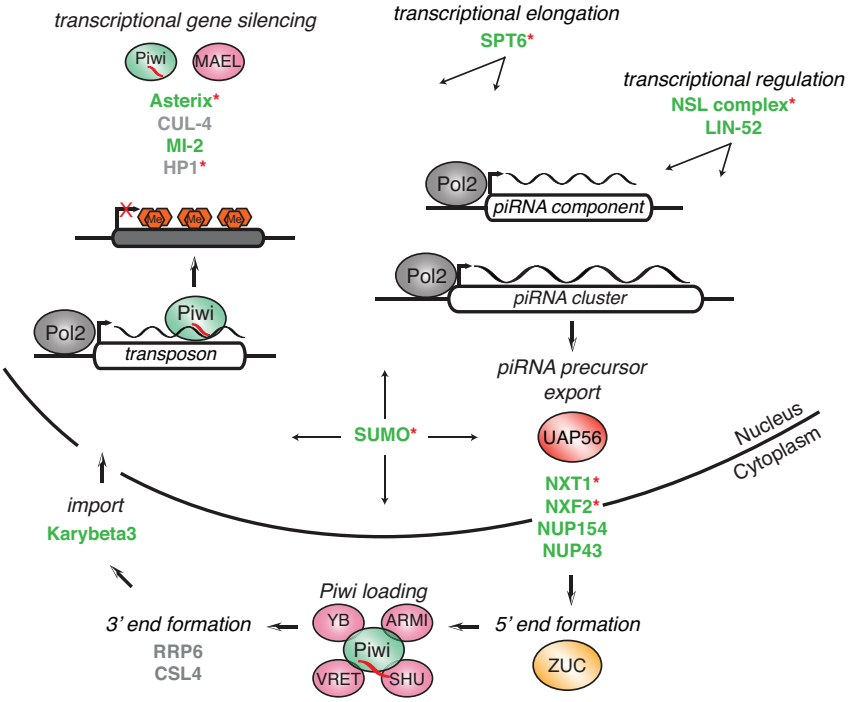


Figure 2.2.5: Potential Roles for Newly Identified piRNA Pathway Components.
Known piRNA components are shown as bubbles. The newly identified genes are shown in bold text colored according to their validation status (color code as in figure 2; Pol2, RNA polymerase 2; red hexagons represent H3K9me3). Red asterisks denote genes that validated in the germline screen done in parallel (Czech et al., 2013).

Regarding piRNA directed TGS, we present *asterix*, a gene affecting the deposition of H3K9me3 marks over full-length transposon insertion, thereby governing their transcriptional activity. Asterix is not predicted to have histone methyltransferase activity itself, but has two highly conserved zinc finger domains likely involved in RNA binding (Andreeva and Tidow, 2008). Where exactly Asterix is located

within the cascade of molecular events between transposon locus transcription and the locus' eventual silencing will be the subject of exciting future research.

Chapter 3

Concluding Remarks

Describing mobile genetic elements as selfish parasites invading a host organism provides a conceptual framework for highlighting two important aspects of TE biology. First, transposon activity can be highly deleterious to the host when not properly controlled (Levin and Moran, 2011; Slotkin and Martienssen, 2007). Second, in order to restrain TE activity, surveillance machineries must distinguish between transposon transcripts and protein coding mRNAs (Malone and Hannon, 2009). The latter aspect is reminiscent of 'self versus non-self' discrimination, which is common in host-parasite interactions. With this in mind, one could state a simple hypothesis: organisms that rid themselves of these genomic parasites altogether should have a considerable advantage in the evolutionary arms race. Nevertheless, this is in stark contrast to what we observe in most genomes. Transposons seem to be ubiquitous within all domains of life, and a genome without any signs of TE activity is considered a peculiarity (Huang et al., 2012). In fact, transposons are not only omnipresent, but also seem to influence all aspects of genome evolution and gene regulation (Levin and Moran, 2011; Rebollo et al., 2012; Slotkin and Martienssen, 2007). Several decades ago, Doolittle and Sapienza gave a possible explanation for this seeming paradox: If a genomic element, even without proven phenotypic function, has evolved a strategy to propagate within a genome, then no other explanation for its presence is necessary (Doolittle and Sapienza, 1980). They argue that transposition is such a strategy, and that the multiplicative nature of transposition alone is enough to ensure genomic survival, because "a single copy of a DNA sequence of no phenotypic benefit to the host risks deletion" (Doolittle and Sapienza, 1980). This argument, however, contains one flaw: it does not account for transposition of class 2 DNA elements through a cut-and-paste mechanism, which is conservative in copy number. According to Doolittle and Sapienza's argument, this subclass of elements should therefore have ceased to exist long ago, which is clearly not the case.

The idea that TEs exist as genomic parasites without any beneficial impact on the host was put into a population genetics model by Hickey (Hickey, 1982). This model, however, was based on the same assumption that "for simplicity, a single transposable element (…) replicates in the process of transposition." (Hickey, 1982) Again, this may be an oversimplification as it does not account for the presence of

all types of TEs. Nonetheless, Hickey's model provides a reasonable framework for many of the observed interactions between TEs and their hosts. Consequently, it was later put into the context of the genomics era as more genome sequences became available (Bestor, 2003). Bestor argues that any influence of TEs on gene regulation or genome evolution is merely a coincidence that can be attributed to their omnipresence. If TEs were truly a source of inheritable variability, then asexual organisms that are in a greater need of such variability in order to adapt, would be expected to display a higher percentage of active TEs, which is not the case. However, if the opposite view of TEs as parasites were true, active transposons would be more abundant in obligate sexual organisms. According to Bestor, this is in fact what can be observed in nature. He uses the premise of TEs as genomic parasites to further argue that host-encoded defense mechanisms (such as cytosine methylation) must have evolved in response to these parasites. Yet, the question remains why these defense mechanisms have not evolved in a way that eliminates transposable elements from eukaryotic genomes.

Some of these considerations are based on the supposition that the activity of TEs has neutral or only negative phenotypic outcomes. Nonetheless, there are a number of examples of TEs and their activity having a positive effect on their host's fitness (Brookfield, 2005; Kidwell and Lisch, 2010). For example, *P* element insertion into the *methuselah* gene in *Drosophila melanogaster* led to an increase in life-span and stress resistance (Lin et al., 1998). Thus, while the models presented above may explain the persistence of a genomic parasite with no beneficial impact on their host, they do not account for the full complexity of TE-host interactions. In fact, a single instance of a transposon having a positive evolutionary effect would render the need for such an explanation obsolete. There may still be a benefit to the depiction of TEs as either selfish or helpful, however, only if these designations are used as a proxy for a certain aspect of TE biology. If they are used to describe the nature of TEs as a whole, one risks promoting a dispute over semantics instead of biology. To avoid such a discord, TEs should be considered building blocks of genomes rather than selfish parasites or helpful elements. As such they can be expected to have the same impact on genome evolution as any other regulatory element: sometimes beneficial, sometimes deleterious (Feschotte, 2008; Rebollo et al., 2010).

One could argue that the impact of TE activity on the host in terms of evolutionary fitness depends entirely on the genomic context in which they exist. A helpful allegory to explain this proposal is the presence of introns in eukaryotic genes. While still subject to debate, current hypotheses state that introns likely originated from a bacterial endosymbiont (Cavalier-Smith, 1991; Cech, 1986; Martin and Koonin, 2006). Group II introns exhibit properties similar to those that led Doolittle and Sapienza to propose their theory on the persistence of TEs within genomes (Cousineau et al., 2000; Lambowitz and Zimmerly, 2004). Thus, introns could be considered selfish parasites, which invaded a intronless host genome, followed by their mutational decay (Martin and Koonin, 2006). This resulted in the emergence of spliceosome-dependent introns, which are a hallmark of eukaryotic genes. If one would consider the presence of these spliceosome-dependent introns while at the same time ignoring the presence of a spliceosome, they would

have to be considered 'junk', disrupting most of our coding sequences, creating an incredible mutational burden on the host. Yet, referring to introns as selfish junk DNA would be considered a great underappreciation of their cellular function.

The same argument can be made for TEs: Their presence and activity should always be judged in the context of host defense pathways controlling their activity. Bestor's reasoning that these pathways have evolved in response to a threat imposed by genomic parasites, could consequently lead to an alternate conclusion. The very nature of these host defense pathways illustrates that they have likely evolved in response to the presence of TEs, which can be beneficial in times, but also deleterious when unrestrained. Instead of using sRNAs for targeted elimination of transposable elements (a mechanism that has evolved in ciliates, reviewed in Sabin et al. (2013)), TE control pathways offer the appropriate tools to govern transposon activity while simultaneously allowing them to persist in the host genome.

In the *Drosophila* germline, Piwi-piRNA complexes do not alter TE loci at the DNA sequence level, but instead target them for epigenetic silencing at the chromatin level (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). This 'gentle' way of silencing is effective enough for transposon control in follicle cells, which do not exhibit PTGS through the ping-pong cycle (Malone et al., 2009). Intriguingly, active transcription is required for successful Piwi-piRNA directed TGS (Sienski et al., 2012). The superficial leakiness of this system can be interpreted as a way to allow for low-level transposition, thereby creating genetic variance. This variance could potentially be advantageous to the host as long as transposition levels do not exceed a certain threshold. If, for any reason, transcriptional levels of TEs rise above these low levels in germ cells, thereby threatening the integrity of the inheritable genome, the adaptive portion of the piRNA pathway takes over; when a transposon mRNA makes its way to the cytoplasm of a germ cell, secondary biogenesis of piRNAs is triggered and the target mRNA can be silenced through PTGS (Brennecke et al., 2007; Gunawardane et al., 2007). The duality of this system highlights how this pathway may have evolved in response to the dual effects of transposition.

Viewing transposons as endogenous genomic building blocks implies that the same rules that govern gene expression are also applicable to TEs. In fact, transposon control and gene regulatory networks operate using a parallel set of core mechanisms. For example, sRNAs bound by Argonaute proteins control gene and transposon expression in the miRNA and the piRNA pathway, respectively. It is therefore conceivable that additional mechanisms that govern RNA expression, processing and surveillance are also shared between these two entities. For example, transposon and piRNA cluster transcription is influenced by the same cellular factors that are involved in epigenetic gene regulation, such as histone modification by methyltransferases and recruitment of HP1 (Kawaoka et al., 2013; Le Thomas et al., 2013; Luteijn and Ketting, 2013; Ritland Politz et al., 2013; Sienski et al., 2012). Similar to genes, piRNA clusters and some transposons are transcribed by Pol II and are likely 5' capped and 3' polyadenylated (Li et al., 2013; Weiner et al., 1986). Given these similarities, it then follows that these transcripts should be governed by the same entities that affect any other Pol II transcript. Nonetheless, downstream processing of piRNA

producing transcripts and coding mRNAs is very different. Therefore, the question arises how the fate of these transcripts is decided. Specific DNA binding proteins could initially target these loci and induce certain modifications of histones and the overall chromatin structure (e.g. H3K9me3 by *egg* (Rangan et al., 2011)). In the case of bi-directionally transcribed piRNA clusters, RNA binding through Piwi-piRNA complexes could actually replace this DNA binding event. Given the strand bias of Piwi loaded piRNAs, transcription of the opposite strand would be necessary for this mechanism, since Piwi is not a DNA binding protein itself (de Vanssay et al., 2012). Targeting of the locus and subsequent changes in the chromatin state would then induce the recruitment of downstream histone binding proteins such as *rhi* (Klattenhoff et al., 2009).

Based on the observation that mRNAs usually form ribonucleoprotein (RNP) complexes (Müller-McNicoll and Neugebauer, 2013), it seems likely that specific proteins bound to each class of transcripts then govern their fate, rather than an intrinsic property of the transcript itself, such as secondary structure or modifications. Transcription by Pol II might trigger the attraction of specific RNA binding proteins that attach to the newly emerging transcript (e.g. UAP56 (Zhang et al., 2012)). If RNA binding proteins are critical to distinguish piRNA-producing transcripts from regular mRNAs, an important question is what other proteins may play a role in this step. Given that piRNA clusters are collections of transposon fragments, one could imagine that a single TE locus, incapable of transposition, could have been the predecessor of a piRNA cluster. If this premise were true, one could expect to find overlap between proteins that bind piRNA cluster transcripts and those that bind to transposon transcripts or associate with their respective RNPs. Supporting this hypothesis, a recent study on the LINE1 ORF1 interactome revealed that the human ortholog of UAP56 co-precipitates with tagged ORF1, or more likely with the LINE1 RNP (Goodier et al., 2013). Besides UAP56, homologs of a number of other hits of the screen described in chapter 2.2 can be found within the LINE1 RNP interactome: Armi (a piRNA biogenesis factor), Larp (an mRNA binding protein) and several uncharacterized proteins. PolyA binding protein (pAbp), which interacts with Larp and was found within the ORF1 co-immunoprecipitate, also scored weakly in the primary screen. If one could test binding of these proteins to piRNA clusters, new insights into piRNA cluster biology would be within reach.

In conclusion, RNA binding proteins specific to the piRNA pathway may protect a piRNA-producing transcript whenever it could be subject to general RNA processing, and this processing would disrupt piRNA biogenesis. One example for this principle is co-transcriptional splicing of pre-mRNAs. The splicing machinery ignored splice sites within the artificial sequences introduced into ectopic piRNA clusters described in chapter 2.1. Suppression of processing signals such as splice sites seems critical from an evolutionary perspective: *de novo* insertion of a transposon into a piRNA cluster will likely introduce at least a polyadenylation signal or other regulatory elements, which would disrupt or alter the locus' transcription and processing. Thus, binding of proteins specific to piRNA cluster transcripts could be suppressing the recruitment of SR proteins and the spliceosome. This could be coupled to a shuttling mechanism that en-

sures the correct localization of precursors. Compartmentalization of downstream biogenesis steps would then ensure proper processing. Nonetheless, the suppression of general RNA processing is not always necessary. There should be no selective pressure to evolve specialized parts of a biogenesis pathway, if access of the general RNA processing machinery would aid in such biogenesis, rather than disrupt it.

Intriguingly, the involvement of general cellular pathways in piRNA biogenesis is reflected in the list of genes that scored in the genome-wide screen. For example, LIN-52 and the RNA export factors NXT1 and NXF2 are likely to be involved in piRNA biogenesis. The putative ortholog of LIN-52, A-MYB, is an ancient transcription factor involved in piRNA cluster transcription in mice and is highly conserved (Lewis et al., 2004; Li et al., 2013). Instead of being restricted to this role, A-MYB serves as a master regulator of transcription during the pachytene stage of meiosis (Bolcun-Filas et al., 2011; Trauth et al., 1994). The second example, NXT1, can be considered a general RNA export factor (Herold et al., 2001). NXF2 on the other hand, shows an expression profile restricted to the germline (Herold et al., 2001). Assuming that NXF2's expression implies a potential specificity to the piRNA pathway, it could be considered the more interesting factor. Nonetheless, knockdown of NXT1 *in vivo* had much stronger effects on transposon expression than knockdown of NXF2 (see chapter 2.2). Besides being the strongest hit in the *in vivo* validation, knockdown of NXT1 also led to a severe drop in mature piRNA levels (see Figure 2.2.3). A straightforward hypothesis would be that NXT1 is involved in the export of piRNA cluster transcripts to the cytoplasm, highlighting how these transcripts can be a substrate of general RNA binding factors. Another possibility could be that the strong transposon de-repression is a cumulative phenotype: Knockdown of *nxt1* may additionally affect export of genic mRNAs, such as those of piRNA biogenesis factors. Whether or not NXT1 acts as a dimer with NXF1 in these processes remains to be investigated.

A hypothetical model summarizing these considerations is presented in figure 3.1.

Taken together, our results indicate that the piRNA pathway and other more general RNA processing pathways are more interconnected than initially anticipated. Understanding these connections may prove invaluable for our understanding of transposon control. A reasonable starting point to this end may be the many cellular pathways outside of the usual realm of a small RNA biologist that emerged from our studies.

Even almost 75 years after their discovery, transposable elements continue to excite researchers from all fields. Given their immense impact on genome evolution and their potential implication in human disease, this interest is likely to continue. There are many open questions remaining to be answered, particularly regarding the mechanisms that keep these elements in check. Understanding these mechanisms may be the key to settling the dispute over whether TEs should be considered 'junk' or the quintessence of genomes to which we owe our very existence.

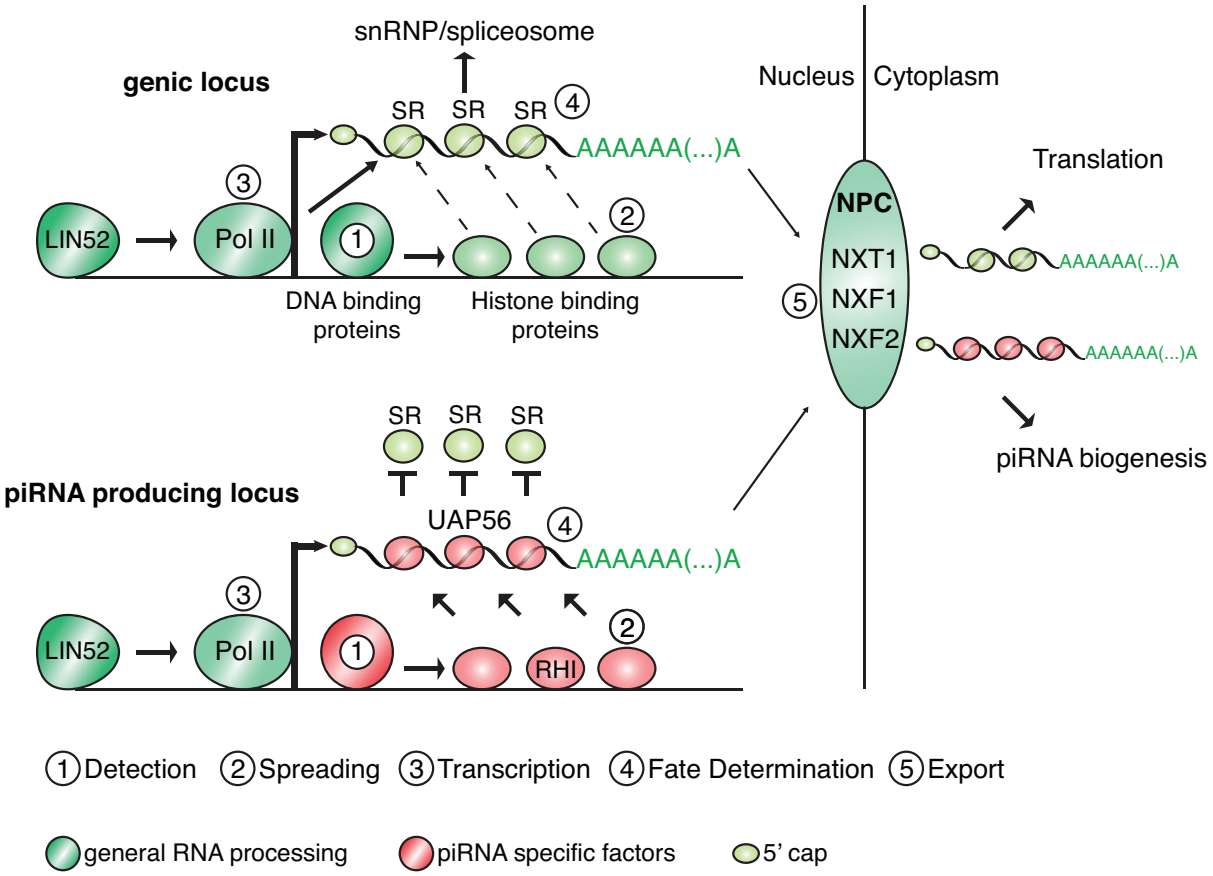


Figure 3.1: A hypothetical model for fate decisions in RNA processing pathways.
General RNA expression and processing factors are shown in green colors, piRNA specific factors in red colors. (1) piRNA producing loci are distinguished from regular genic loci through the action of DNA binding proteins. These proteins are specific to each class of locus and recruit different downstream effectors. (2) Recruitment of chromatin modifiers and histone binding proteins (e.g. RHI) leads to formation of higher order chromatin structures, thereby establishing the identity of the locus. (3) Pol II transcribes both types of loci and is driven by the same general transcription factors (e.g. LIN-52). (4) Ribonucleoprotein complexes are formed co-transcriptionally. The proteins involved in this process are recruited by histone bound specificity factors. In the case of genic loci, SR proteins and subsequently spliceosomal effectors are recruited to the mRNA. In the case of piRNA producing loci, this is suppressed by the presence of piRNA specific RNA binding proteins (e.g. UAP56). (5) Both types of RNPs might be bound and exported through the nuclear pore complex (NPC) by similar export factors. Their ultimate fate is decided in the cytoplasm based on proteins bound to the respective RNA.

List of Figures

List of Tables

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H. C., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor Miklos, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Michael Cherry, J., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Deslattes Mays, A., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Harley Gorrell, J., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D. C., Scheeler, F., Shen, H., Christopher Shue, B., Siden-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., Woodage, T., Worley, K. C., Wu, D., Yang, S., Alison Yao, Q., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Craig Venter, J. (2000). The genome sequence of Drosophila melanogaster. *Science*, 287(5461):2185–2195.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106.

Andreeva, A. and Tidow, H. (2008). A novel CHHC Zn-finger domain found in spliceosomal proteins and tRNA modifying enzymes. *Bioinformatics*, 24(20):2277–2280.

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(7099):203–207.

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(13):1017–1027.

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. *Molecular Cell*, 31(6):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(5825):744–747.

Begun, D. J. and Aquadro, C. F. (1993). African and North American populations of Drosophila melanogaster are very different at the DNA level. *Nature*, 365(6446):548–550.

Benjamin, H. W. and Kleckner, N. (1989). Intramolecular transposition by Tn10. *Cell*, 59(2):373–383.

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(6818):363–366.

Bestor, T. H. (2003). Cytosine methylation mediates sexual conflict. *Trends in Genetics*, 19(4):185–190.

Boeke, J. D., Garfinkel, D. J., Styles, C. A., and Fink, G. R. (1985). Ty elements transpose through an RNA intermediate. *Cell*, 40(3):491–500.

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(15):3319–3330.

Bourc'his, D. and Bestor, T. H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, 431(7004):96–99.

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(6):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(5906):1387–1392.

Brookfield, J. F. Y. (2005). The ecology of the genome - mobile DNA elements and their hosts. *Nature Publishing Group*, 6(2):128–136.

Brower-Toland, B., Findley, S. D., Jiang, L., Liu, L., Yin, H., Dus, M., Zhou, P., Elgin, S. C. R., and Lin, H. (2007). Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes & Development*, 21(18):2300–2311.

Bucheton, A. (1995). The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. *Trends in Genetics*, 11(9):349–353.

Bucheton, A., Paro, R., Sang, H. M., Pélisson, A., and Finnegan, D. J. (1984). The molecular basis of I-R hybrid dysgenesis in Drosophila melanogaster: identification, cloning, and properties of the I factor. *Cell*, 38(1):153–163.

Cam, H. P. (2010). Roles of RNAi in chromatin regulation and epigenetic inheritance. *Epigenomics*, 2(5):613–626.

Castro, J. P. and Carareto, C. M. A. (2004). Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation. *Genetica*, 121(2):107–118.

Cavalier-Smith, T. (1991). Intron phylogeny: a new hypothesis. *Trends in Genetics*, 7(5):145–148.

Cech, T. R. (1986). The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell*, 44(2):207–210.

Chalvet, F., Teysset, L., Terzian, C., Prud'homme, N., Santamaria, P., Bucheton, A., and Pélisson, A. (1999). Proviral amplification of the Gypsy endogenous retrovirus of Drosophila melanogaster involves env-independent invasion of the female germline. *The EMBO Journal*, 18(9):2659–2669.

Chambeyron, S. and Bucheton, A. (2005). I elements in Drosophila: in vivo retrotransposition and regulation. *Cytogenet Genome Res*, 110(1-4):215–222.

Chen, Y., Pane, A., and Schüpbach, T. (2007). Cutoff and aubergine mutations result in retrotransposon upregulation and checkpoint activation in Drosophila. *Curr Biol*, 17(7):637–642.

Chintapalli, V. R., Wang, J., and Dow, J. A. T. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nature Genetics*, 39(6):715–720.

Chiu, Y.-L. and Greene, W. C. (2008). The APOBEC3 cytidine deaminases: an innate defensive network opposing exogenous retroviruses and endogenous retroelements. *Annual review of immunology*, 26:317–353.

Chung, W.-J., Okamura, K., Martin, R., and Lai, E. C. (2008). Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Current Biology*, 18(11):795–802.

Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., and Neuberger, M. S. (2005). Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. *Molecular biology and evolution*, 22(2):367–377.

Cousineau, B., Lawrence, S., Smith, D., and Belfort, M. (2000). Retrotransposition of a bacterial group II intron. *Nature*, 404(6781):1018–1021.

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 & development*, 12(23):3715–3727.

Cox, D. N., Chao, A., and Lin, H. (2000). piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development*, 127(3):503–514.

Craig, N. L., Craigie, R., Gellert, M., and Lambowitz, A. M. (2002). Mobile DNA II.

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(7196):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. *Molecular Cell*, 50(5):749–761.

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(7418):112–115.

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(7150):151–156.

Doolittle, W. F. and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, 284(5757):601–603.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics*, 9(5):397–405.

Feschotte, C. and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet*, 41:331–368.

Findley, S. D. (2003). Maelstrom, a Drosophila spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage. *Development*, 130(5):859–871.

Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., and Mello, C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811.

Garfinkel, D. J., Boeke, J. D., and Fink, G. R. (1985). Ty element transposition: reverse transcriptase and virus-like particles. *Cell*, 42(2):507–517.

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., Xu, J., Kittler, E. L. W., Zapp, M. L., Weng, Z., and Zamore, P. D. (2008). Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, 320(5879):1077–1081.

Ghildiyal, M. and Zamore, P. (2009). Small silencing RNAs: an expanding universe. *Nature Reviews Genetics*, 10(2):94–108.

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., and Petrov, D. A. (2008). High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *PLoS Biology*, 6(10):e251–2129.

González, J., Macpherson, J. M., and Petrov, D. A. (2009). A recent adaptive transposable element insertion near highly conserved developmental loci in Drosophila melanogaster. *Molecular biology and evolution*, 26(9):1949–1961.

González, J. and Petrov, D. A. (2009). The adaptive role of transposable elements in the Drosophila genome. *Gene*, 448(2):124–133.

Goodier, J. L., Cheung, L. E., and Kazazian, H. H. (2013). Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res*.

Gregory, T. R. and Hebert, P. D. (1999). The modulation of DNA content: proximate causes and ultimate consequences. *Genome Research*, 9(4):317–324.

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 Drosophila. *Science*, 315(5818):1587–1590.

Guzzardo, P. M., Muerdter, F., and Hannon, G. J. (2013). The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics & Development*, 23(1):44–52.

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 & development*, 24(22):2499–2504.

Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. (2002). Two classes of short interfering RNA in RNA silencing. *The EMBO Journal*, 21(17):4671–4679.

Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952.

Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature*, 404(6775):293–296.

Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R., and Hannon, G. J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science*, 293(5532):1146–1150.

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. *The EMBO Journal*, 30(19):3977–3993.

Hartl, D. L. (2000). Molecular melodies in high and low C. *Nature Publishing Group*, 1(2):145–149.

Herold, A., Klymenko, T., and Izaurralde, E. (2001). NXF1/p15 heterodimers are essential for mRNA nuclear export in Drosophila. *RNA*, 7(12):1768–1780.

Hickey, D. A. (1982). Selfish DNA: a sexually-transmitted nuclear parasite. *Genetics*, 101(3-4):519–531.

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. *Current Biology*, 17(14):1265–1272.

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D. V., Blaser, H., Raz, E., Moens, C. B., Plasterk, R. H. A., Hannon, G. J., Draper, B. W., and Ketting, R. F. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, 129(1):69–82.

Huang, C. R. L., Burns, K. H., and Boeke, J. D. (2012). Active Transposition in Genomes. *Annu Rev Genet*, 46(1):651–675.

Huang, X. A., Yin, H., Sweeney, S., Raha, D., Snyder, M., and Lin, H. (2013). A major epigenetic programming mechanism guided by piRNAs. *Developmental Cell*, 24(5):502–516.

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(7423):279–283.

Ishizu, H., Siomi, H., and Siomi, M. C. (2012). Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. *Genes & Development*, 26(21):2361–2373.

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*, 110(1-4):462–467.

Kalmykova, A. I., Klenov, M. S., and Gvozdev, V. A. (2005). Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. *Nucleic Acids Research*, 33(6):2052–2059.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., Ashburner, M., and Celniker, S. E. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome biology*, 3(12):RESEARCH0084.

Kapitonov, V. V. and Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, 9(5):411–2– author reply 414.

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., Okada, T. N., Siomi, M. C., and Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, 453(7196):793–797.

Kawaoka, S., Hara, K., Shoji, K., Kobayashi, M., Shimada, T., Sugano, S., Tomari, Y., Suzuki, Y., and Katsuma, S. (2013). The comprehensive epigenome map of piRNA clusters. *Nucleic Acids Res*, 41(3):1581–1590.

Kawaoka, S., Izumi, N., Katsuma, S., and Tomari, Y. (2011). 3' end formation of PIWI-interacting RNAs in vitro. *Molecular Cell*, 43(6):1015–1022.

Kazazian, H. H., Wong, C., Youssoufian, H., Scott, A. F., Phillips, D. G., and Antonarakis, S. E. (1988). Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, 332(6160):164–166.

Ketting, R. F., Fischer, S. E. J., Bernstein, E., Sijen, T., Hannon, G. J., and Plasterk, R. H. A. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. 15(20):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(2):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(7):1551–1563.

Kidwell, M. G. (1983). Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 80(6):1655–1659.

Kidwell, M. G., Kidwell, J. F., and Sved, J. A. (1977). Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics*, 86(4):813–833.

Kidwell, M. G. and Lisch, D. R. (2010). Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution*, 55(1):1–24.

Klattenhoff, C., Xi, H., Li, C., Lee, S., Xu, J., Khurana, J. S., Zhang, F., Schultz, N., Koppetsch, B. S., Nowosielska, A., Seitz, H., Zamore, P. D., Weng, Z., and Theurkauf, W. E. (2009). The Drosophila HP1 Homolog Rhino Is Required for Transposon Silencing and piRNA Production by Dual-Strand Clusters. *Cell*, 138(6):1137–1149.

Kleckner, N. (1990). Regulation of transposition in bacteria. *Annual review of cell biology*, 6:297–327.

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. *Proceedings of the National Academy of Sciences*, 108(46):18760–18765.

Knight, S. W. and Bass, B. L. (2001). A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science*, 293(5538):2269–2271.

Koch, C. M., Honemann-Capito, M., Egger-Adam, D., and Wodarz, A. (2009). Windei, the Drosophila Homolog of mAM/MCAF1, Is an Essential Cofactor of the H3K9 Methyl Transferase dSETDB1/Eggless in Germ Line Development. *PLoS Genetics*, 5(9):e1000644.

Lambowitz, A. M. and Zimmerly, S. (2004). Mobile group II introns. *Annu Rev Genet*, 38:1–35.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R. A., Muzny, D. M., Scherer, S. E., Bouck, J. B., Sodergren, E. J., Worley, K. C., Rives, C. M., Gorrell, J. H., Metzker, M. L., Naylor, S. L., Kucherlapati, R. S., Nelson, D. L., Weinstock, G. M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D. R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H. M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R. W., Federspiel, N. A., Abola, A. P., Proctor, M. J., Myers, R. M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D. R., Olson, M. V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G. A., Athanasiou, M., Schultz, R., Roe, B. A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W. R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J. A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D. G., Burge, C. B., Cerutti, L., Chen, H. C., Church, D., Clamp, M., Copley, R. R., Doerks, T., Eddy, S. R., Eichler, E. E., Furey, T. S., Galagan, J., Gilbert, J. G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L. S., Jones, T. A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W. J., Kitts, P., Koonin, E. V., Korf, I., Kulp, D., Lancet, D., Lowe, T. M., McLysaght, A., Mikkelsen, T., Moran, J. V., Mulder, N., Pollara, V. J., Ponting, C. P., Schuler, G., Schultz, J., Slater, G., Smit, A. F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y. I., Wolfe, K. H., Yang, S. P., Yeh, R. F., Collins, F., Guyer, M. S., Peterson, J., Felsenfeld, A., Wetterstrand, K. A., Patrinos, A., Morgan, M. J., de Jong, P., Catanese, J. J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y. J., Szustakowki, J., and International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

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 biology*, 10(3):R25.

Laski, F. A., Rio, D. C., and Rubin, G. M. (1986). Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell*, 44(1):7–19.

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, J. K., Aravin, A. A., and Toth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes & development*, 27(4):390–399.

Lee, Y. S., Nakahara, K., Pham, J. W., Kim, K., He, Z., Sontheimer, E. J., and Carthew, R. W. (2004). Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 117(1):69–81.

Lévesque, L., Guzik, B., Guan, T., Coyle, J., Black, B. E., Rekosh, D., Hammarskjöld, M. L., and Paschal, B. M. (2001). RNA export mediated by tap involves NXT1-dependent interactions with the nuclear pore complex. *The Journal of biological chemistry*, 276(48):44953–44962.

Levin, H. L. and Moran, J. V. (2011). Dynamic interactions between transposable elements and their hosts. *Nature Reviews Genetics*, 12(9):615–627.

Lewis, P. W., Beall, E. L., Fleischer, T. C., Georlette, D., Link, A. J., and Botchan, M. R. (2004). Identification of a Drosophila Myb-E2F2/RBF transcriptional repressor complex. *Genes & development*, 18(23):2929–2940.

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. W., Zapp, M. L., Klattenhoff, C., Schulz, N., Theurkauf, W. E., Weng, Z., and Zamore, P. D. (2009a). Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals Somatic piRNAs in Flies. *Cell*, 137(3):509–521.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009b). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

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. *Molecular Cell*, 50(1):67–81.

Lim, A. 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(16):6714–6719.

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(12):2463–2476.

Lin, Y. J., Seroude, L., and Benzer, S. (1998). Extended life-span and stress resistance in the Drosophila mutant methuselah. *Science*, 282(5390):943–946.

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., Kheradpour, P., Ernst, J., Jordan, G., Mauceli, E., Ward, L. D., Lowe, C. B., Holloway, A. K., Clamp, M., Gnerre, S., Alföldi, J., Beal, K., Chang, J., Clawson, H., Cuff, J., Di Palma, F., Fitzgerald, S., Flicek, P., Guttman, M., Hubisz, M. J., Jaffe, D. B., Jungreis, I., Kent, W. J., Kostka, D., Lara, M., Martins, A. L., Massingham, T., Moltke, I., Raney, B. J., Rasmussen, M. D., Robinson, J., Stark, A., Vilella, A. J., Wen, J., Xie, X., Zody, M. C., Broad Institute Sequencing Platform and Whole Genome Assembly Team, Baldwin, J., Bloom, T., Chin, C. W., Heiman, D., Nicol, R., Nusbaum, C., Young, S., Wilkinson, J., Worley, K. C., Kovar, C. L., Muzny, D. M., Gibbs, R. A., Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, Cree, A., Dihn, H. H., Fowler, G., Jhangiani, S., Joshi, V., Lee, S., Lewis, L. R., Nazareth, L. V., Okwuonu, G., Santibanez, J., Warren, W. C., Mardis, E. R., Weinstock, G. M., Wilson, R. K., Genome Institute at Washington University, Delehaunty, K., Dooling, D., Fronik, C., Fulton, L., Fulton, B., Graves, T., Minx, P., Sodergren, E., Birney, E., Margulies, E. H., Herrero, J., Green, E. D., Haussler, D., Siepel, A., Goldman, N., Pollard, K. S., Pedersen, J. S., Lander, E. S., and Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482.

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(5689):1437–1441.

Liu, Q., Rand, T. A., Kalidas, S., Du, F., Kim, H.-E., Smith, D. P., and Wang, X. (2003). R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science*, 301(5641):1921–1925.

Lowe, C. B. and Haussler, D. (2012). 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS ONE*, 7(8):e43128.

Luan, D. D., Korman, M. H., Jakubczak, J. L., and Eickbush, T. H. (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, 72(4):595–605.

Luteijn, M. J. and Ketting, R. F. (2013). PIWI-interacting RNAs: from generation to transgenerational epigenetics. *Nature Reviews Genetics*.

Ma, L., Buchold, G. M., Greenbaum, M. P., Roy, A., Burns, K. H., Zhu, H., Han, D. Y., Harris, R. A., Coarfa, C., Gunaratne, P. H., Yan, W., and Matzuk, M. M. (2009). GASZ Is Essential for Male Meiosis and Suppression of Retrotransposon Expression in the Male Germline. *PLoS Genetics*, 5(9):e1000635.

Mahowald, A. P. (1971a). Polar granules of Drosophila. 3. The continuity of polar granules during the life cycle of Drosophila. *The Journal of experimental zoology*, 176(3):329–343.

Mahowald, A. P. (1971b). Polar granules of drosophila. IV. Cytochemical studies showing loss of RNA from polar granules during early stages of embryogenesis. *The Journal of experimental zoology*, 176(3):345–352.

Malone, C. and Hannon, G. (2009). Small RNAs as guardians of the genome. *Cell*, 136(4):656–668.

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(3):522–535.

Martin, W. and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. *Nature*, 440(7080):41–45.

McClintock, B. (1942). The Fusion of Broken Ends of Chromosomes Following Nuclear Fusion. *Proc Natl Acad Sci U S A*, 28(11):458–463.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*, 36(6):344–355.

McClintock, B. (1951). Chromosome organization and genic expression. *Cold Spring Harbor symposia on quantitative biology*, 16:13–47.

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science*, 226(4676):792–801.

McQuilton, P., St Pierre, S. E., Thurmond, J., and FlyBase Consortium (2012). FlyBase 101–the basics of navigating FlyBase. *Nucleic Acids Res*, 40(Database issue):D706–14.

Mével-Ninio, M., Pelisson, A., Kinder, J., Campos, A. R., and Bucheton, A. (2007). The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. *Genetics*, 175(4):1615–1624.

Moriyama, E. N. and Powell, J. R. (1996). Intraspecific nuclear DNA variation in Drosophila. *Molecular biology and evolution*, 13(1):261–277.

Müller-McNicoll, M. and Neugebauer, K. M. (2013). How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nature Reviews Genetics*, 14(4):275–287.

Nagao, A., Sato, K., Nishida, K. M., Siomi, H., and Siomi, M. C. (2011). Gender-Specific Hierarchy in Nuage Localization of PIWI-Interacting RNA Factors in Drosophila. *Frontiers in genetics*, 2:55.

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(7423):284–287.

Ohno, S. (1972). So much "junk" DNA in our genome. *Brookhaven symposia in biology*, 23:366–370.

Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M. C. (2004). Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes & development*, 18(14):1655–1666.

Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257.

Olivieri, D., Senti, K.-A., Subramanian, S., Sachidanandam, R., and Brennecke, J. (2012). The cochaperone shutdown defines a group of biogenesis factors essential for all piRNA populations in Drosophila. *Molecular Cell*, 47(6):954–969.

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. *The EMBO Journal*, 29(19):3301–3317.

Pal-Bhadra, M., Bhadra, U., and Birchler, J. A. (2002). RNAi related mechanisms affect both transcriptional and posttranscriptional transgene silencing in Drosophila. *Molecular Cell*, 9(2):315–327.

Pal-Bhadra, M., Leibovitch, B. A., Gandhi, S. G., Rao, M., Bhadra, U., Birchler, J. A., and Elgin, S. C. R. (2004). Heterochromatic Silencing and HP1 Localization in Drosophila Are Dependent on the RNAi Machinery. *Science*, 303(5658):669–672.

Pane, A., Jiang, P., Zhao, D. Y., Singh, M., and pbach, T. S. u. (2011). The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline. *The EMBO Journal*, 30(22):4601–4615.

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(6):851–862.

Pasyukova, E. G., Nuzhdin, S. V., Morozova, T. V., and Mackay, T. F. C. (2004). Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. *The Journal of heredity*, 95(4):284–290.

Pélisson, A. (1981). The I-R system of hybrid dysgenesis in Drosophila Melanogaster: Are I factor insertions responsible for the mutator effect of the I-R interaction? *Molecular Genetics and Genomics*, 183(1):123–129.

Pelisson, A., Song, S., Prud'homme, N., Smith, P., Bucheton, A., and Corces, V. (1994). Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The EMBO Journal*, 13(18):4401.

Pélisson, A., Teysset, L., Chalvet, F., Kim, A., Prud'homme, N., Terzian, C., and Bucheton, A. (1997). About the origin of retroviruses and the co-evolution of the gypsy retrovirus with the Drosophila flamenco host gene. *Genetica*, 100(1-3):29–37.

Picard, G. (1976). Non-mendelian female sterility in Drosophila melanogaster: hereditary transmission of I factor. *Genetics*, 83(1):107–123.

Piriyapongsa, J. and Jordan, I. K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*, 2(2):e203.

Piriyapongsa, J. and Jordan, I. K. (2008). Dual coding of siRNAs and miRNAs by plant transposable elements. *RNA*, 14(5):814–821.

Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I. K. (2006). Origin and Evolution of Human microRNAs From Transposable Elements. *Genetics*, 176(2):1323–1337.

Preall, J. B., Czech, B., Guzzardo, P. M., Muerdter, F., and Hannon, G. J. (2012). shutdown is a component of the Drosophila piRNA biogenesis machinery. *RNA*, 18(8):1446–1457.

Qi, H., Watanabe, T., Ku, H.-Y., Liu, N., Zhong, M., and Lin, H. (2011). The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. *The Journal of biological chemistry*, 286(5):3789–3797.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Raddatz, G., Guzzardo, P. M., Olova, N., Fantappié, M. R., Rampp, M., Schaefer, M., Reik, W., Hannon, G. J., and Lyko, F. (2013). Dnmt2-dependent methylomes lack defined DNA methylation patterns. *Proceedings of the National Academy of Sciences*, 110(21):8627–8631.

Raja, S. J., Charapitsa, I., Conrad, T., Vaquerizas, J. M., Gebhardt, P., Holz, H., Kadlec, J., Fraterman, S., Luscombe, N. M., and Akhtar, A. (2010). The nonspecific lethal complex is a transcriptional regulator in Drosophila. *Molecular Cell*, 38(6):827–841.

Rangan, P., Malone, C. D., Navarro, C., Newbold, S. P., Hayes, P. S., Sachidanandam, R., Hannon, G. J., and Lehmann, R. (2011). piRNA production requires heterochromatin formation in Drosophila. *Curr Biol*, 21(16):1373–1379.

Rebollo, R., Horard, B., Hubert, B., and Vieira, C. (2010). Jumping genes and epigenetics: Towards new species. *Gene*, 454(1-2):1–7.

Rebollo, R., Romanish, M. T., and Mager, D. L. (2012). Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. *Annu Rev Genet*, 46(1):21–42.

Ritland Politz, J. C., Scalzo, D., and Groudine, M. (2013). Something Silent This Way Forms: The Functional Organization of the Nuclear Repressive Compartment. *Annual review of cell and developmental biology*.

Roche, S. E. and Rio, D. C. (1998). Trans-silencing by P elements inserted in subtelomeric heterochromatin involves the Drosophila Polycomb group gene, Enhancer of zeste. *Genetics*, 149(4):1839–1855.

Ronsseray, S., Lehmann, M., and Anxolabéhère, D. (1991). The maternally inherited regulation of P elements in Drosophila melanogaster can be elicited by two P copies at cytological site 1A on the X chromosome. *Genetics*, 129(2):501–512.

Ronsseray, S., Marin, L., Lehmann, M., and Anxolabéhère, D. (1998). Repression of hybrid dysgenesis in Drosophila melanogaster by combinations of telomeric P-element reporters and naturally occurring P elements. *Genetics*, 149(4):1857–1866.

Rozhkov, N. V., Hammell, M., and Hannon, G. J. (2013). Multiple roles for Piwi in silencing Drosophila transposons. *Genes & development*, 27(4):400–412.

Rubin, G. M., Kidwell, M. G., and Bingham, P. M. (1982). The molecular basis of P-M hybrid dysgenesis: The nature of induced mutations. *Cell*, 29(3):987–994.

Sabin, L. R., Delás, M. J., and Hannon, G. J. (2013). Dogma derailed: the many influences of RNA on the genome. *Molecular Cell*, 49(5):783–794.

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(7268):1296–1299.

Saito, K., Ishizu, H., Komai, M., Kotani, H., Kawamura, Y., Nishida, K. M., Siomi, H., and Siomi, M. C. (2010). Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. *Genes & development*, 24(22):2493–2498.

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 & Development*, 20(16):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 & Development*, 21(13):1603–1608.

Sarot, E., Payen-Groschêne, G., Bucheton, A., and Pélisson, A. (2004). Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. *Genetics*, 166(3):1313–1321.

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., Abbott, R. M., Cotton, M., Levy, A., Marchetto, P., Ochoa, K., Jackson, S. M., Gillam, B., Chen, W., Yan, L., Higginbotham, J., Cardenas, M., Waligorski, J., Applebaum, E., Phelps, L., Falcone, J., Kanchi, K., Thane, T., Scimone, A., Thane, N., Henke, J., Wang, T., Ruppert, J., Shah, N., Rotter, K., Hodges, J., Ingenthron, E., Cordes, M., Kohlberg, S., Sgro, J., Delgado, B., Mead, K., Chinwalla, A., Leonard, S., Crouse, K., Collura, K., Kudrna, D., Currie, J., He, R., Angelova, A., Rajasekar, S., Mueller, T., Lomeli, R., Scara, G., Ko, A., Delaney, K., Wissotski, M., Lopez, G., Campos, D., Braidotti, M., Ashley, E., Golser, W., Kim, H., Lee, S., Lin, J., Dujmic, Z., Kim, W., Talag, J., Zuccolo, A., Fan, C., Sebastian, A., Kramer, M., Spiegel, L., Nascimento, L., Zutavern, T., Miller, B., Ambroise, C., Muller, S., Spooner, W., Narechania, A., Ren, L., Wei, S., Kumari, S., Faga, B., Levy, M. J., McMahan, L., Van Buren, P., Vaughn, M. W., Ying, K., Yeh, C.-T., Emrich, S. J., Jia, Y., Kalyanaraman, A., Hsia, A.-P., Barbazuk, W. B., Baucom, R. S., Brutnell, T. P., Carpita, N. C., Chaparro, C., Chia, J.-M., Deragon, J.-M., Estill, J. C., Fu, Y., Jeddeloh, J. A., Han, Y., Lee, H., Li, P., Lisch, D. R., Liu, S., Liu, Z., Nagel, D. H., McCann, M. C., SanMiguel, P., Myers, A. M., Nettleton, D., Nguyen, J., Penning, B. W., Ponnala, L., Schneider, K. L., Schwartz, D. C., Sharma, A., Soderlund, C., Springer, N. M., Sun, Q., Wang, H., Waterman, M., Westerman, R., Wolfgruber, T. K., Yang, L., Yu, Y., Zhang, L., Zhou, S., Zhu, Q., Bennetzen, J. L., Dawe, R. K., Jiang, J., Jiang, N., Presting, G. G., Wessler, S. R., Aluru, S., Martienssen, R. A., Clifton, S. W., McCombie, W. R., Wing, R. A., and Wilson, R. K. (2009). The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., and Malim, M. H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature*, 418(6898):646–650.

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(5):964–980.

Sijen, T. and Plasterk, R. H. A. (2003). Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature*, 426(6964):310–314.

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(4):246–258.

Slotkin, R. and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nature Reviews Genetics*, 8(4):272–285.

Smalheiser, N. R. and Torvik, V. I. (2006). Alu elements within human mRNAs are probable microRNA targets. *Trends in Genetics*, 22(10):532–536.

Smit, A., Hubley, R., and Green, P. (2010). REPEATMASKER OPEN-3.0.

Soifer, H. S., Zaragoza, A., Peyvan, M., Behlke, M. A., and Rossi, J. J. (2005). A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. *Nucleic Acids Res*, 33(3):846–856.

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(5689):1434–1437.

Spradling, A. C. and Rubin, G. M. (1981). Drosophila genome organization: conserved and dynamic aspects. *Annu Rev Genet*, 15:219–264.

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. *The Journal of Cell Biology*, 185(4):613–627.

Tabara, H., Sarkissian, M., Kelly, W. G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C. C. (1999). The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell*, 99(2):123–132.

Tam, O., Aravin, A., Stein, P., Girard, A., Murchison, E., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R., and Hannon, G. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–538.

Tolia, N. H. and Joshua-Tor, L. (2007). Slicer and the argonautes. *Nature chemical biology*, 3(1):36–43.

Tomari, Y., Du, T., and Zamore, P. D. (2007). Sorting of Drosophila small silencing RNAs. *Cell*, 130(2):299–308.

Trauth, K., Mutschler, B., Jenkins, N. A., Gilbert, D. J., Copeland, N. G., and Klempnauer, K. H. (1994). Mouse A-myb encodes a trans-activator and is expressed in mitotically active cells of the developing central nervous system, adult testis and B lymphocytes. *The EMBO Journal*, 13(24):5994–6005.

Trono, D. (2004). Retroviruses under editing crossfire: a second member of the human APOBEC3 family is a Vif-blockable innate antiretroviral factor. *EMBO reports*, 5(7):679–680.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A., and Kakutani, T. (2009). Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, 461(7262):423–426.

Ungerer, M. C., Strakosh, S. C., and Zhen, Y. (2006). Genome expansion in three hybrid sunflower species is associated with retrotransposon proliferation. *Current Biology*, 16(20):R872–3.

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(5785):320–324.

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(12):2128–2134.

Weiner, A. M., Deininger, P. L., and Efstratiadis, A. (1986). Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annual review of biochemistry*, 55:631–661.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., and Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, 8(12):973–982.

Wilson, C., Pearson, R. K., Bellen, H. J., O'Kane, C. J., Grossniklaus, U., and Gehring, W. J. (1989). P-element-mediated enhancer detection: an efficient method for isolating and characterizing developmentally regulated genes in Drosophila. *Genes & Development*, 3(9):1301–1313.

Yang, N. and Kazazian, H. H. (2006). L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nature structural & molecular biology*, 13(9):763–771.

Yoshimura, T., Toyoda, S., Kuramochi-Miyagawa, S., Miyazaki, T., Miyazaki, S., Tashiro, F., Yamato, E., Nakano, T., and Miyazaki, J.-i. (2009). Gtsf1/Cue110, a gene encoding a protein with two copies of a CHHC Zn-finger motif, is involved in spermatogenesis and retrotransposon suppression in murine testes. *Developmental biology*, 335(1):216–227.

Zamore, P. D., Tuschl, T., Sharp, P. A., and Bartel, D. P. (2000). RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, 101(1):25–33.

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(4):871–884.

Zhang, L. and Rong, Y. S. (2012). Retrotransposons at Drosophila telomeres: Host domestication of a selfish element for the maintenance of genome integrity. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1819(7):771–775.

FELIX MÜRDTER

1 Bungtown Road
Cold Spring Harbor, 11724 NY, U.S.A.

Phone: 516-367-8459

email: fmuerdte@cshl.edu

Education

03/2010 – 10/2013 (expected)	Cold Spring Harbor Laboratory, USA University of Tübingen, Germany	Phd, Natural Sciences
10/2004 – 02/2010	University of Tübingen, Germany	Diploma, Biology

Research Experience

04/2009 – present	Cold Spring Harbor Laboratory, USA Lab of Dr. Greg Hannon "Transposon activity and control in the *Drosophila* germline"	Graduate Researcher
10/2006 – 03/2009	MPI for Developmental Biology, Germany Lab of Dr. Detlef Weigel Group Leader: Dr. Markus Schmid "Analysis of regulatory cis-elements in the FT-Promoter of *Arabidopsis thaliana* and related Brassicaceae"	Undergraduate Researcher
06/2008 – 07/2008	WSI Computer Science Dept. and ZMBP University of Tübingen, Germany Supervisor: Dr. Kenneth Berendzen and Dr. Andreas Zell "Statistical Background Models based on Genome-wide Analysis"	Research internship
04/2008 – 06/2008	MPI for Developmental Biology, Germany Lab of Dr. Detlef Weigel Supervisor: Dr. Markus Schmid "Purification of FT, TFL1 and FD protein expressed in E. coli"	Research internship
02/2006 – 09/2006	MPI for Developmental Biology, Germany Lab of Dr. Christiane Nüsslein-Volhard EU-Project: "Zebrafish models for human development and disease"	Undergraduate Researcher

Teaching Experience

2011	Cold Spring Harbor Laboratory, USA Mentor in the Undergraduate Research Program (CSHL) and Exceptional Research Opportunities Program (HHMI)
2007	University of Tübingen and MPI for Developmental Biology, Germany Tutor in Molecular Biology courses Supervisor: Dr. Markus Schmid

Honors & awards

2013 Best Talk (PhD Workshop) at the 8th Microsymposium on Small RNAs
IMBA, Vienna

2011 Outstanding Graduate Student Poster Prize, In-House Symposium XXV
Cold Spring Harbor Laboratory

2009 Evolution and Ecology Award, EvE Förderpreis für Evolutionsbiologie
Volkswagen Foundation

Publications

*contributed equally

2013 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.
Minotaur is critical for primary piRNA biogenesis. *RNA*, Article in Press.

2013 Muerdter, F.*, Guzzardo, P.M.*, Gillis, J., Luo, Y., Yu, Y., Chen, C., Fekete, R., Hannon, G.J.
A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila. *Molecular Cell*, 50 (5), pp. 736–748.

2013 Guzzardo, P.M., Muerdter, F., Hannon, G.J.
The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics and Development*, 23 (1), pp. 44-52.

2012 Preall, J.B.*, Czech, B.*, Guzzardo, P.M., Muerdter, F., Hannon, G.J.
Shutdown is a component of the Drosophila piRNA biogenesis machinery. *RNA*, 18 (8), pp. 1446-1457.

2012 Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B., Gordon, A., Hannon, G.J., Aravin, A.A.
Production of artificial piRNAs in flies and mice. *RNA*, 18 (1), pp. 42-52.
F1000Prime Recommendation, 19 Jun 2012; DOI: 10.3410/f.716747814.792102818.

2010 Haase, A.D., Fenoglio, S., Muerdter, F., Guzzardo, P.M., Czech, B., Pappin, D.J., Chen, C., Gordon, A., Hannon, G.J.
Probing the initiation and effector phases of the somatic piRNA pathway in Drosophila. *Genes and Development*, 24 (22), pp. 2499-2504.

2009 Mathieu, J., Yant, L.J., Mürdter, F., Küttner, F., Schmid, M.
Repression of Flowering by the miR172 Target SMZ. *PLoS Biology*, 7 (7).

Appendix

A.1 Shell and perl scripts used in this work

A.1.1 Analysis pipeline for small RNA data

In order to meet basic programming standards, each shell script was designed to have an introductory section describing the purpose of the script. Furthermore, a help screen (option –h) explains the command line usage and all available options. Certain variables, such as the small RNA cloning adapter, can be set at runtime (option -a). The script operates on compressed fastq files (input), which is the standard output of the next-generation sequencing facility at Cold Spring Harbor Laboratories. In order to account for the immense volume of these sequencing files, all scripts are able to invoke mapping algorithms such as bowtie using multiple processors or cores (option -p).

Listing A.1: Preamble of a basic small RNA analysis script

```
#!/bin/sh

##
##  small RNA mapping and annotation Pipeline:
##     After mapping, annotate reads
##
##     Save all printed messages to a file, mail it to yourself
#
#   Copyright (C) 2011 F. Muerdter (fmuerdte at cshl dot edu)
#
#   License: Free for personal / academic / non-commercial use.
#
#   See related publication:
#      [ Production of artificial piRNAs in flies and mice ]
#      [ Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech,
    B., Gordon, A., Hannon, G.J., Aravin, A.A. ]
#      [ RNA (2012) ]
#      [ doi: 10.1261/rna.029769.111 ]
#
#   For questions, please contact the corresponding authors:
#   A. A. Aravin, E-mail aaa@caltech.edu.
#   G. J. Hannon, E-mail hannon@cshl.edu.
#
#   When using this script (or derivatives there-of),
```

```
#   please cite: [ Muerdter et al., 2012 ]

## set exit status to last command within a pipeline
set -o pipefail

##
## Script variables, with default options
##
VERSION=0.2 ## important to increment when fixing bugs
ADAPTER="CTGTAGGCACCATCAATC" ## miRNA cloning linker 1 [IDT]
THREADS=16 ## default number of threads for parallel mapping
VERBOSE=
SANGER=false ## are sequences stored in sanger quality scores?

## Handle optional parameters
ARGS=$(getopt -o "p:a:vsh" -- "$@")
if [ $? -ne 0 ]; then
  echo "Error: invalid command-line parameters." >&2
  exit 1
fi
eval set -- "$ARGS"

## Iterate all command line options
while [ $# -gt 0 ]; do
  case "$1" in
    ## Show help and exit
    -h)
        SCRIPT=$(basename "$0")
        echo "
$SCRIPT - A small RNA mapping and annotation Pipeline

version $VERSION
Copyright (C) 2011 - Felix Muerdter ( fmuerdte@cshl.edu )

Usage: $BASENAME [OPTIONS] INPUT-FILE

Options are:
  -h       this help screen.
  -v       verbose, email summary at the end of the run.
  -s       fastq file in Sanger quality scores (default = false)
  -p N     Use N threads (default = $THREADS).
  -a X     Clip 3' cloning adapter X (default = $ADAPTER).

Example:

  # Use 5 threads and ACGT adapter, analyze file 'foo.txt.gz'
  \$ $SCRIPT -p 5 -a ACGT foo.txt.gz

        "
        exit
        ;;

    ## "-v" doesn't take any values - we only "shift" one argument.
    -v) VERBOSE="yes"
        shift
```

```
        ;;

    ## "-s" doesn't take any values - we only "shift" one argument.
    -s) SANGER=true
        shift
        ;;

    ## "-p" requires a value, we "shift" two arguments
    -p) THREADS=$2
        shift 2
        ;;
    ## "-a" requires a value, we "shift" two arguments
    -a) ADAPTER=$2
        shift 2
        ;;

    ## Last optional argument
    --) shift
        break
        ;;
  esac
done
```

Before starting to process the input files, several 'safety checks' have to be passed. The script only executes if the input files exist, and warns if too many or the wrong parameters were used.

Listing A.2: Safety checks

```
## When we get here, "$1" is the first filename to handle
INPUT=$(readlink -f "$1")
## check for correct command line
if [ -z "$INPUT" ]; then
  echo "Error: missing file name. use -h for help." >&2
  exit 1
fi

## check for existence of input file
if [ ! -e "$INPUT" ]; then
  echo "Error: input file '$INPUT' not found." >&2
  exit 1
fi

## Warn if there are too many parameters
if [ ! -z "$2" ]; then
  echo "Error: too many file names given. Use -h for help." >&2
  exit 1
fi
```

At this point, the script has to make variables that were set at runtime visible to subscripts. Creating a unique target directory, consisting of the input file name and a time stamp, ensures the safe storage of all results.

Listing A.3: Create a unique target directory

```
# make adapter sequence and number of threads visible to subscripts
export ADAPTER
```

```
export THREADS
export SANGER

# pass on input ID to subscripts
export INPUT

## create the output file name
#  determine input file compression type (gz or bz2)
MIME=$(file -bi "$INPUT" | sed 's/;.*//') || exit 1
if [ "$MIME" = "application/x-gzip" ]; then
  BASENAME=$(basename "$INPUT" .txt.gz)
elif [ "$MIME" = "application/x-bzip2" ]; then
  BASENAME=$(basename "$INPUT" .txt.bz2)
else
   echo "Error: Unknown MIME type for file ($INPUT) got '$MIME'" >&2
   exit 1
fi
export BASENAME

## Add today's date and time
DATE=$(date "+%Y-%m-%d_%H%M%S")

## start script
echo "Processing file '$BASENAME', using "$THREADS" thread(s) and "$ADAPTER
    " adapter..."

## create directory with unique name
NEWDIR="$BASENAME-$DATE"
mkdir "$NEWDIR" || exit 1
cd "$NEWDIR" || exit 1
```

After setting the genome variable to the appropriate value, the input fastq file is processed. The actual code that executes the necessary external programs is stored in a subscript. This practice makes the main script more readable and allows for easier management of subscripts shared among several analysis pipelines.

Listing A.4: Set genome and call first subscript

```
# set genome to drosophila melanogaster
GENOME="dm3"
export GENOME

## fastq  preprocessing:
# convert from fastq to fasta, clip adapter, collapse
~/bin/piRNA_pipeline_sub/fastq_processing.sh || exit 1
echo "...done"
```

Fastq processing prior to mapping requires conversion of the input file into fasta format. Most recent mapping algorithms can process fastq files directly, however, this step allows for more control over what exactly is being mapped. In the case of small RNA libraries, the number of sequencing cycles often exceeds the length of the actual RNA molecule. Therefore, the end of the sequencing read might be part of the 3'

end cloning adapter that was used. This could interfere with mapping, wherefore any part of the cloning adapter (specified by –a) has to be removed. This process is called clipping.

The structure of a typical small RNA library is very different from other libraries with high complexity, such as genomic DNA libraries. Single RNA reads may get sequenced multiple times. For example, certain miRNA reads can be represented more than 100,000 times in a library. Therefore, it is reasonable to collapse the fasta file into a non-redundant list, while preserving the read count as part of the header line for each sequence. The read or multiplicity count can be recovered later and used for normalization purposes. Collapsing the reads can substantially accelerate the subsequent mapping step.

All these fastq processing steps can be accomplished with a software package called FASTX Toolkit (v0.0.13.2 by A. Gordon, gordon@cshl.edu).

piRNA_pipeline_sub/fastq_processing.sh

```
#!/bin/sh

set –o pipefail

echo ”Converting fastq file to fasta format...”
## fastq to fasta, rename sequence identifiers to numbers, be verbose

# use xxcat
#xxcat – prints a (possibly compressed) file to STDOUT
#Version 0.2
#Copyright (C) 2012    A. Gordon ( gordon at cshl dot edu )

if $SANGER ; then
    /home/gordon/bin/xxcat ”$INPUT” |
    fastq_to_fasta –Q 33 –r –v –o ”$BASENAME”.fa \
    1> genome_fastq_to_fasta.txt || exit 1
    else
    /home/gordon/bin/xxcat ”$INPUT” |
    fastq_to_fasta –r –v –o ”$BASENAME”.fa \
    1> genome_fastq_to_fasta.txt || exit 1
fi

echo ”Collapse fasta file...”
# collapse
fastx_collapser –v –i ”$BASENAME”.fa –o ”$BASENAME”_collapsed.fa \
  1> genome_fasta_collapser_1.txt || exit 1

# remove unclipped fasta file
rm ”$BASENAME”.fa || exit 1

echo ”Clipping adapter...”
# Clipping illumina sequencing adapter,
# keep only sequences with adapter which are
# longer than 15 nt after clipping
fastx_clipper –i ”$BASENAME”_collapsed.fa –a $ADAPTER \
  –o ”$BASENAME”_clipped.fa –c –v –l 15 \
  1> genome_fasta_clipper.txt || exit 1

```

```
# remove unclipped fasta file
rm "$BASENAME"_collapsed.fa || exit 1

echo "Collapse clipped fasta file..."
# collapse
fastx_collapser -v -i "$BASENAME"_clipped.fa \
  -o "$BASENAME"_collapsed.fa \
  1> genome_fasta_collapser_2.txt || exit 1

# remove uncollapsed fasta file
rm "$BASENAME"_clipped.fa || exit 1
```

In the following step, the library is mapped against the reference genome.

Listing A.5: Call genome mapping subscript

```
## genome mapping
# map with up to two mismatches to genome
~/bin/piRNA_pipeline_sub/genome_mapping.sh || exit 1
echo "...done"
```

A number of mapping algorithms suitable for short read mapping are available and can produce different outcomes leading to different conclusions. Even using the same mapping algorithm with different parameters can significantly alter the final mapping results. Therefore, one has to pay close attention as to how the available options change the behavior of the used alignment tool.

One such option is to allow for mismatches between the sequenced RNA molecule and a putative location of origin in the reference genome. Since the interstrain level of sequence polymorphisms may be high in the given model organism, which is the case for *Drosophila melanogaster* (Begun and Aquadro, 1993; Moriyama and Powell, 1996), it might be reasonable to allow for up to 10% of a read's nucleotides to differ from the reference sequence in a valid alignment.

Another key aspect of the mapping step is the level of uniqueness of a mapping location. Depending on the biological question, only reads with one unique mapping location may be considered in order to avoid unwanted artifacts. This proved to be critical in the discovery of piRNA clusters in D. melanogaster (Brennecke et al., 2007), but may be of lesser importance in other applications. The standard small RNA analysis pipeline presented herein employed bowtie, a memory-efficient short read aligner (Langmead et al., 2009), and only considered mapping events valid if the read had less than or equal to two mismatches and one unique mapping location.

The clipped and collapsed sequence files may still contain a considerable amount of molecular contamination, such as ribosomal RNA, synthetic cloning markers or viral RNA. This contamination can differ in amounts between two given sRNA libraries making it impossible to properly normalize and compare reads between libraries. For this reason, the mapping strategy was to map the entire library to these contaminations and consequently only map non-mapping reads to the genome.

piRNA_pipeline_sub/genome_mapping.sh

```
#!/bin/sh

```

```
# set temporary directory for genome mapping
DIR=$(mktemp -d --tmpdir=. -t gen-map.XXXXXXX) || exit 1

set -o pipefail

# set index to synthetic
INDEX_GENOME=~/genomes/synthetic

echo "Mapping with up to one mismatch to synthetic..."
# Map with bowtie, up to one mismatches, keep the non-mappers
# The synthetic mapping index is a collection of synthetic sequences commonly
# used in the lab environment or during the library construction
bowtie -f -v 1 -k 1 --best -p $THREADS --sam --un $DIR/synthetic_non_mappers_v1.txt \
  "$INDEX_GENOME" "$BASENAME"_collapsed.fa > $DIR/synthetic_output_v1.sam 2> synthetic_mapping_results.txt || exit 1

# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/synthetic_output_v1.sam > synthetic_output.sam

# set index to viruses
INDEX_GENOME=~/genomes/dm3_viruses

echo "Mapping with up to three mismatches to dm3_viruses..."
# Map with bowtie, up to three mismatches, keep the non-mappers
bowtie -f -v 3 -k 1 --best -p $THREADS --sam --un $DIR/dm3_viruses_non_mappers_v3.txt \
  "$INDEX_GENOME" $DIR/synthetic_non_mappers_v1.txt > $DIR/dm3_viruses_output_v3.sam 2> dm3_viruses_mapping_results.txt || exit 1

# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/dm3_viruses_output_v3.sam > dm3_viruses_output.sam

# set index to tRNAs
INDEX_GENOME=~/genomes/dmel-tRNA

echo "Mapping with up to 2 mismatches to dmel-tRNA..."
# Map with bowtie, up to two mismatches, keep the non-mappers
bowtie -f -v 2 -k 1 --best -p $THREADS --sam --un $DIR/dmel-tRNA_non_mappers_v2.txt \
  "$INDEX_GENOME" $DIR/dm3_viruses_non_mappers_v3.txt > $DIR/dmel-tRNA_output_v2.sam 2> dmel-tRNA_mapping_results.txt || exit 1

# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/dmel-tRNA_output_v2.sam  > dmel-tRNA_output.sam

# set index to tRNAs
INDEX_GENOME=~/genomes/dmel-miscRNA

echo "Mapping with up to three mismatches to dmel-miscRNA..."
# Map with bowtie, up to three mismatches, keep the non-mappers
bowtie -f -v 3 -k 1 --best -p $THREADS --sam --un $DIR/dmel-miscRNA_non_mappers_v3.txt \
```

```
  "$INDEX_GENOME" $DIR/dmel-tRNA_non_mappers_v2.txt > $DIR/dmel-
     miscRNA_output_v3.sam 2> dmel-miscRNA_mapping_results.txt || exit 1

# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/dmel-miscRNA_output_v3.sam  > dmel-miscRNA_output.sam

# set index to genome
INDEX_GENOME=~/genomes/"$GENOME"_genome/"$GENOME"_genome

echo "Mapping with zero mismatches to "$GENOME"..."
# Map with bowtie, zero mismatches, keep the non-mappers
bowtie -f -v 0 -a -m 1 -p $THREADS --sam --un $DIR/genome_non_mappers_v0.
    txt \
  "$INDEX_GENOME" $DIR/dmel-miscRNA_non_mappers_v3.txt > $DIR/
     genome_output_v0.sam 2> genome_mapping_results_v0.txt || exit 1

if [ -e $DIR/genome_non_mappers_v0.txt ]; then
echo "Mapping with 1 mismatch to "$GENOME"..."
# Map with bowtie, allow one mismatch, keep the non-mappers
# NOTE: we don't need the SAM header for the second file, so use "--sam-
    nohead"
bowtie -f -v 1 -a -m 1 -p $THREADS --sam --sam-nohead --un $DIR/
    genome_non_mappers_v1.txt\
  "$INDEX_GENOME" $DIR/genome_non_mappers_v0.txt > $DIR/genome_output_v1.
     sam 2> genome_mapping_results_v1.txt || exit 1
fi

if [ -e $DIR/genome_non_mappers_v1.txt ]; then
echo "Mapping with 2 mismatches to "$GENOME"..."
# Map with bowtie, allow two mismatches, keep the non-mappers
# NOTE: we don't need the SAM header for the second file, so use "--sam-
    nohead"
bowtie -f -v 2 -a -m 1 -p $THREADS --sam --sam-nohead --un $DIR/
    genome_non_mappers_v2.txt \
  "$INDEX_GENOME" $DIR/genome_non_mappers_v1.txt > $DIR/genome_output_v2.
     sam 2> genome_mapping_results_v2.txt || exit 1
fi

# Combine the three output SAM files into one.
# Only the first file has a SAM HEADER, so put it first.
# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/genome_output_v0.sam $DIR/genome_output_v1.sam $DIR/
    genome_output_v2.sam > genome_output.sam

# set index to transposon
INDEX_GENOME=~/genomes/dmTE942

echo "Mapping with zero mismatches to TEs..."
# Map with bowtie, zero mismatches, keep the non-mappers
bowtie -f -v 0 -a -m 1 -p $THREADS --sam --un $DIR/TEs_non_mappers_v0.txt \
  "$INDEX_GENOME" $DIR/dmel-miscRNA_non_mappers_v3.txt > $DIR/TEs_output_v0
     .sam 2> TEs_mapping_results_v0.txt || exit 1

if [ -e $DIR/TEs_non_mappers_v0.txt ]; then
echo "Mapping with 1 mismatch to TEs..."
```

```
# Map with bowtie, allow one mismatch, keep the non-mappers
# NOTE: we don't need the SAM header for the second file, so use "--sam-
    nohead"
bowtie -f -v 1 -a -m 1 -p $THREADS --sam --sam-nohead --un $DIR/
    TEs_non_mappers_v1.txt\
  "$INDEX_GENOME" $DIR/TEs_non_mappers_v0.txt > $DIR/TEs_output_v1.sam 2>
     TEs_mapping_results_v1.txt || exit 1
fi

if [ -e $DIR/TEs_non_mappers_v1.txt ]; then
echo "Mapping with 2 mismatches to TEs..."
# Map with bowtie, allow two mismatches, keep the non-mappers
# NOTE: we don't need the SAM header for the second file, so use "--sam-
    nohead"
bowtie -f -v 2 -a -m 1 -p $THREADS --sam --sam-nohead --un $DIR/
    TEs_non_mappers_v2.txt \
  "$INDEX_GENOME" $DIR/TEs_non_mappers_v1.txt > $DIR/TEs_output_v2.sam 2>
     TEs_mapping_results_v2.txt || exit 1
fi

# Combine the three output SAM files into one.
# Only the first file has a SAM HEADER, so put it first.
# use AWK to filter out non-mappers (where "chrom" == "*" in the SAM file)
awk '$3 != "*"' $DIR/TEs_output_v0.sam $DIR/TEs_output_v1.sam $DIR/
    TEs_output_v2.sam > TEs_output.sam


# clean up temporary directory
rm -r $DIR || exit 1
```

In order to allow for comparisons between independent sRNA libraries, all mapping read counts were then normalized to reads per million genomic mappers (rpm).

Listing A.6: Call read normalization subscript

```
## read count normalization
# normalize sam file to total number of genomic mappers
~/bin/piRNA_pipeline_sub/genome_normalization.sh || exit 1
echo "...done"
```

The normalization was done using the output of the genome mapping, a SAM file. Since all subsequent steps were done using this SAM file as input, only the normalized read counts were propagated through the pipeline. The tool used to express the normalized read counts as fractions of the total library size was Column Normalizer (Copyright ©2009 by A. Gordon, gordon@cshl.edu, available from http://cancan.cshl.edu/labmembers/gordon/column_normalizer/). In a second step, those fractions were then simply multiplied by 1,000,000 in order to obtain rpm values.

piRNA_pipeline_sub/genome_normalization.sh

```
#!/bin/sh

set -o pipefail

# set temporary directory for genome normalization
DIR=$(mktemp -d --tmpdir=. -t gen-norm.XXXXXXX) || exit 1
```

```
7
8  echo "Normalizing genomic read counts..."
9  # remove header
10 sed -e '/^@/d;s/-/\t/' genome_output.sam > $DIR/tmp1 || exit 1
11
12 # normalize read counts
13 column_normalizer -o $DIR/tmp2 -v -s $DIR/tmp1 2 \
14   1> genome_normalized_reads.sum || exit 1
15
16 # normalize to reads per million
17 tawk '($2=$2*1000000) {print}' $DIR/tmp2 |
18   sed 's/\t/-/' > genome_output_normalized.sam || exit 1
19
20 ## reconstruct normalized sam file with original header
21 # save header
22 grep '^@' genome_output.sam > $DIR/sam_header || exit 1
23 # concatenate header and sam file
24 cat $DIR/sam_header genome_output_normalized.sam > $DIR/tmp3 || exit 1
25 # remove header-less sam file
26 rm genome_output_normalized.sam || exit 1
27 # put new sam file in place
28 mv $DIR/tmp3 genome_output_normalized.sam || exit 1
29
30 # create normalized fasta file for mappers
31 sed '/^@/d' genome_output_normalized.sam |
32   awk '{print ">"$1"\n"$10}' > genome_mappers.fa || exit 1
33
34 # clean up temporary directory
35 rm -r $DIR || exit 1
```

After read count normalization, the script then automatically creates several files for visualization of the mapping results. The file formats (BAM, bigWig and BED) were chosen for their compatibility with the UCSC genome browser (http://genome.ucsc.edu/). A combination of the samtools package (available from http://samtools.sourceforge.net/ (Li et al., 2009b)) and Aaron Quinlan's bedtools (available from https://code.google.com/p/bedtools/ (Quinlan and Hall, 2010)) can handle all these file formats.

The BAM format is the compressed binary version of the SAM mapping results format. The BAI file is the accompanying sorted index for faster access.

piRNA_pipeline_sub/bam_bai.sh

```
1  #!/bin/sh
2
3  set -o pipefail
4
5  # set temporary directory for genome mapping
6  DIR=$(mktemp -d --tmpdir=. -t bam.XXXXXXX) || exit 1
7
8  echo "Creating BAM/BAI files..."
9  # Create a bam file
10 samtools view -S -b "$ID"_output_normalized.sam > "$ID"_output.bam || exit
     1
```

```
11
12 # Sort the bam file
13 samtools sort "$ID"_output.bam "$ID"_output.sorted || exit 1
14
15 # Delete the unsorted BAM file, don't need it any more
16 rm "$ID"_output.bam || exit 1
17 # Rename the sorted BAM file
18 mv "$ID"_output.sorted.bam "$ID"_output.bam || exit 1
19 # Create an index (BAI) file
20 samtools index "$ID"_output.bam || exit 1
21
22 # clean up temporary directory
23 rm -r $DIR || exit 1
```

The bigWig format displays continuous data rather than single mappers. This format is especially helpful in displaying read 'density' over a given feature or chromosome. The programs necessary for the conversion of bedgraphs to bigWig files are available from http://hgdownload.cse.ucsc.edu/admin/exe/.

piRNA_pipeline_sub/bigwig.sh

```
1 #!/bin/sh
2
3 set -o pipefail
4
5 echo "Creating Coverage (BigWig) file..."
6 CHROMSIZE=~/genomes/chrom_sizes/"$GENOME"_chromInfo.txt
7
8 bedtools genomecov -ibam "$ID"_output.bam -g "$CHROMSIZE" -bg > "$ID"
    _coverage.bedgraph || exit 1
9 bedtools genomecov -ibam "$ID"_output.bam -g "$CHROMSIZE" -bg -strand + > "
    $ID"_coverage_sense.bedgraph || exit 1
10 bedtools genomecov -ibam "$ID"_output.bam -g "$CHROMSIZE" -bg -strand - > "
    $ID"_coverage_antisense.bedgraph || exit 1
11
12 bedGraphToBigWig "$ID"_coverage.bedgraph "$CHROMSIZE" "$ID"_coverage.bw ||
    exit 1
13 bedGraphToBigWig "$ID"_coverage_sense.bedgraph "$CHROMSIZE" "$ID"
    _coverage_sense.bw || exit 1
14 bedGraphToBigWig "$ID"_coverage_antisense.bedgraph "$CHROMSIZE" "$ID"
    _coverage_antisense.bw || exit 1
```

The BED format is a simple list of mapping coordinates, which can be manipulated with unix text tools or bedtools.

piRNA_pipeline_sub/bed.sh

```
1 #!/bin/sh
2
3 set -o pipefail
4
5 echo "Creating BED file..."
6 # Create a bed file for downstream applications
7 bedtools bamtobed -i "$ID"_output.bam > "$ID"_output.bed || exit 1
```

The annotation of sequenced reads is key to understand both the overall complexity of the sRNA library (i.e. identify putative contaminations or irregularities), as well as to identify subpopulations for subsequent analysis (i.e. transposon derived sequences etc.). This can be done taking several independent routes, which are explained in the following paragraphs.

First, the appropriate subscript has to be called.

Listing A.7: Call read annotation subscript

```
## annotate reads
#  1. intersect genomic coordinates with transposon locations
#  2. intersect all un-annotated reads with all other features.
#     only keep annotation with highest priority
~/bin/piRNA_pipeline_sub/annotation.sh || exit 1
echo "...done"
```

Reads can be annotated by either mapping to specific features, as previously described for the identification of rRNA or viral RNA derived sequences, or by intersecting genomic mapping coordinates with previously annotated features. The BED or GFF files needed for the latter approach can be downloaded from FlyBase (ftp://ftp.flybase.net/genomes/).

One obvious complication in this step is that any given sRNA can map to more than one feature (or its mapping coordinates can intersect with more than one annotation). Therefore, multiple annotations have to be categorized by an underlying priority list. Each read can then have multiple annotations and additional information reflecting their priorities, or only carry along one annotation, which is naturally the one with the highest priority.

This concept was realized in two ways. Mapping to sequences such as rRNAs or virally derived RNAs, and consequently only mapping non-mappers to the genome, gives these contaminations the highest priority. This approach could be considered conservative but allows for a very robust normalization downstream. The priority list decided upon for the annotation by genomic coordinates is entirely arbitrary and can be adjusted to the experiment's underlying biological question.

Table A.1.1: Annotation priority list

Priority	Annotation
-1	no_annotation
0	transposable_element
1	miRNA
2	pre_miRNA
3	ncRNA
4	rRNA
5	transposable_element_insertion_site
6	snRNA

Table A.1.1: Annotation priority list (continued)

Priority	Annotation
7	snoRNA
8	tRNA
9	five_prime_UTR
10	three_prime_UTR
11	exon
12	intron
13	CDS
14	mRNA
15	gene
16	pseudogene
17	polyA_site
18	exon_junction
19	regulatory_region
20	protein_binding_site
21	DNA_motif
22	enhancer
23	orthologous_to
24	insertion_site
25	complex_substitution
26	deletion
27	point_mutation
28	rescue_fragment
29	uncharacterized_change_in_nucleotide_sequence
30	sequence_variant
31	protein
32	mature_peptide
33	BAC_cloned_genomic_insert
34	breakpoint
35	chromosome_band

piRNA_pipeline_sub/annotation.sh

```
1 #!/bin/sh
```

```
2
3  set -o pipefail
4
5  # set temporary directory for genome mapping
6  DIR=$(mktemp -d --tmpdir=. -t annotation.XXXXXXX) || exit 1
7
8  echo "Annotating mapped reads..."
9
10 # identify transposon derived sequences with highest priority
11 bedtools intersect -bed -wao -f 1.00 -a genome_output.bed \
12   -b ~/genomes/dm3_genome/gff-bed/dmel-r5.43_FM_TEs.bed \
13   > $DIR/tmp1 || exit 1
14 # save transposons sequences
15 awk '$13>0' $DIR/tmp1 > $DIR/tmp2 || exit 1
16 # get all sequences not annotated as transposons
17 tawk '($13==0){print $1,$2,$3,$4,$5,$6}' $DIR/tmp1 \
18   > $DIR/tmp3 || exit 1
19 # annotate all non transposons sequences
20 bedtools intersect -bed -wao -f 1.00 -s -a $DIR/tmp3 \
21   -b ~/genomes/dm3_genome/gff-bed/dmel-r5.43_FM_genes_clean.bed \
22     > $DIR/tmp4 || exit 1
23 # combine transposons and non-transposons derived sequences
24 cat $DIR/tmp2 $DIR/tmp4 > $DIR/tmp5 || exit 1
25 # only keep annotation with highest priority
26 tawk '{print $1,$2,$3,$4,$11,$6,$10}' $DIR/tmp5 |
27   sort -k4,4 | groupBy -g 4 -c 5 -o min | sort -k1,1 > $DIR/tmp6 || exit 1
28 # make joinable sam file
29 sed '/^@/d' genome_output_normalized.sam |
30   sort -k1,1 > $DIR/tmp7 || exit 1
31 # keep header for later
32 grep '^@' genome_output_normalized.sam > $DIR/header || exit 1
33 # add annotation to original sam file
34 join -t " " $DIR/tmp7 $DIR/tmp6 | sort -k15,15 > $DIR/tmp8 || exit 1
35 # build final annotation file
36 join -t " " -1 15 -2 1 $DIR/tmp8 ~/genomes/dm3_genome/gff-bed/priority.txt
     |
37   cut -f2,3,4,5,6,7,8,9,10,11,12,13,14,15,16 > $DIR/tmp9 || exit 1
38 cat $DIR/header $DIR/tmp9 > final_annotations.txt || exit 1
39
40 # print summary information for reads and sequences
41 sed '/^@/d;s/^\([0-9]*\)-\([0-9]*\)/\2/' final_annotations.txt |
42   sort -k15,15 | groupBy -g 15 -c 1 -o sum |
43   awk '{printf "%s\t%.1f\n",$1,$2}' > annotations.summary.reads || exit 1
44 sed '/^@/d;s/^\([0-9]*\)-\([0-9]*\)/\2/' final_annotations.txt |
45   sort -k15,15 | groupBy -g 15 -c 1 -o count |
46   awk '{printf "%s\t%d\n",$1,$2}' > annotations.summary.seqs || exit 1
47
48 # calculate cluster counts
49 bedtools intersect -bed -wo -f 1.00 -a genome_output.bed \
50   -b ~/genomes/dm3_genome/gff-bed/piRNA_clusters.bed |
51   tawk '{print $4,$10}' | uniq | sort -u | sed 's/^\([0-9]*\)-\([0-9]*\)
       /\2/' |
52   sort -k2,2 | groupBy -g 2 -c 1 -o sum |
53   sort -k2,2nr  > $DIR/tmp10 || exit 1
54 column_normalizer -f -p 3 -a $DIR/tmp10 2 |
```

```
  tawk '{printf "%s\t%d\t%.1f\n",$1,$2,$3*100}' > cluster_counts.txt

# clean up temporary directory
rm -r $DIR || exit 1
```

After annotation, a number of small tasks was executed in order to gain knowledge about specific properties of the library. For example, the length distribution for all genomic mappers was calculated.

piRNA_pipeline_sub/length_distributions.sh

```
#!/bin/sh

set -o pipefail

# set temporary directory for length distributions
DIR=$(mktemp -d --tmpdir=. -t lengthdist.XXXXXXX) || exit 1

echo "Creating length distribution for genomic mappers...:"
fasta_formatter -i "$ID"_mappers.fa -t | sed 's/^[0-9]*-//' |
  tawk '(l=length($2)){print $1,l}' |
  sort -k2,2 | groupBy -g 2 -c 1 -o sum \
  > "$ID"_length_distribution.txt || exit 1

# separate sense and antisense
echo -e "length\tsense\tantisense" > "$ID"_length_distribution_sas.csv
sed '/^@/d;s/^\([0-9]*\)-\([0-9]*\)/\2/' "$ID"_output_normalized.sam |
  tawk '{if ($2==0) {print length($10),$1,0} else {print length($10),0,"-"
     $1}}' |
  sort -k1,1 | groupBy -g 1 -c 2,3 -o sum,sum \
  >> "$ID"_length_distribution_sas.csv || exit 1

# clean up temporary directory
rm -r $DIR || exit 1
```

Next, the fraction of reads that start with a U at their 5' end was calculated. An unusually high percentage of 1U reads would be a typical signature of primary piRNAs.

piRNA_pipeline_sub/1U.sh

```
#!/bin/sh

set -o pipefail

# set temporary directory for genome normalization
DIR=$(mktemp -d --tmpdir=. -t 1U.XXXXXXX) || exit 1

echo "Calculating percentage of Us at 5' end of small RNA..."

# only look at piRNA reads
sed '/^@/d' "$ID"_output_normalized_23nt.sam |
  awk '{
  if ($2==0) {
    NT = substr($10,0,1)
    } else
  if ($2==16) {
    if ( "C" == substr($10,length($10),1) ) NT = "G";else
```

```
    if ( "G" == substr($10,length($10),1) ) NT = "C";else
    if ( "A" == substr($10,length($10),1) ) NT = "T";else
    if ( "T" == substr($10,length($10),1) ) NT = "A"
    };
    {print NT}}' |
  sort | groupBy -g 1 -c 1 -o count > $DIR/"$ID".1U

# extract useful information (only 1U percentage)
column_normalizer $DIR/"$ID".1U 2 |
  awk '($1=="T"){printf "Percent 1U:\t%.1f%\n",$2*100}' \
  > "$ID"_1U.percent
```

Lastly, the number of small RNA 'neighbors' with a given overlap was calculated. This information can be used to identify biogenesis partners such as siRNA duplexes (with a 19nt overlap) or piRNA ping-pong partners (10nt overlap). Given that the underlying methodology was rather complex, both from a biological and bioinformatic standpoint, the scripts for these overlap calculations are discussed in a dedicated chapter (Appendix A.1.2).

The last step of the sRNA pipeline was designed to clean up all output files and produce a easy to read report file. This report file was then emailed to the user for easier access. The report file contained basic quality control measurements, mapping statistics and a summary of all annotations present in the analyzed sRNA library.

piRNA_pipeline_sub/report.sh

```
#!/bin/sh

echo "Creating report..."
##
## Create a friendly report file
##

echo "Mapping Results for $INPUT" > report.txt
echo "==========================" >> report.txt
echo "Results are stored in: " >> report.txt
pwd >> report.txt
echo >> report.txt
if [ -e genome_fastq_to_fasta.txt ]; then
  echo "Fastq to Fasta conversion:" >> report.txt
  cat genome_fastq_to_fasta.txt >> report.txt
fi
echo >> report.txt
if [ -e genome_fasta_clipper.txt ]; then
  echo "Fasta clipper:" >> report.txt
  cat genome_fasta_clipper.txt >> report.txt
fi
echo >> report.txt
if [ -e genome_fasta_collapser_1.txt ]; then
  echo "Fasta collapser:" >> report.txt
  cat genome_fasta_collapser_1.txt >> report.txt
fi
echo >> report.txt
if [ -e genome_fasta_collapser_2.txt ]; then
```

```
  echo "Fasta collapser of clipped fasta file:" >> report.txt
  cat genome_fasta_collapser_2.txt >> report.txt
fi
echo >> report.txt
if [ -e genome_fasta_collapser_unclipped.txt ]; then
  echo "Fasta collapser of unclipped sequences:" >> report.txt
  cat genome_fasta_collapser_unclipped.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e synthetic_mapping_results.txt ]; then
  echo "== Alignment with mismatches to synthetic:" >> report.txt
  cat  synthetic_mapping_results.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e dm3_viruses_mapping_results.txt ]; then
  echo "== Alignment with mismatches to dm3_viruses:" >> report.txt
  cat  dm3_viruses_mapping_results.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e dmel-miscRNA_mapping_results.txt ]; then
  echo "== Alignment with mismatches to dmel-miscRNA:" >> report.txt
  cat  dmel-miscRNA_mapping_results.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e dmel-tRNA_mapping_results.txt ]; then
  echo "== Alignment with mismatches to dmel-tRNA:" >> report.txt
  cat  dmel-tRNA_mapping_results.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e genome_mapping_results_v0.txt ]; then
  echo "== Alignment with 0 mismatches to dm3 (unique mappers):" \
  >> report.txt
  cat  genome_mapping_results_v0.txt >> report.txt
fi
if [ -e genome_mapping_results_v1.txt ]; then
  echo "== Alignment with 1 mismatch to dm3 (unique mappers):" \
  >> report.txt
  cat  genome_mapping_results_v1.txt >> report.txt
fi
if [ -e genome_mapping_results_v2.txt ]; then
  echo "== Alignment with 2 mismatches to dm3 (unique mappers):" \
  >> report.txt
  cat  genome_mapping_results_v2.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e TEs_mapping_results_v0.txt ]; then
  echo "== Alignment with 0 mismatches to TEs (multiple mappers):" \
  >> report.txt
  cat  TEs_mapping_results_v0.txt >> report.txt
```

```
fi
if [ -e TEs_mapping_results_v1.txt ]; then
  echo "== Alignment with 1 mismatch to TEs (multiple mappers):" \
  >> report.txt
  cat  TEs_mapping_results_v1.txt >> report.txt
fi
if [ -e TEs_mapping_results_v2.txt ]; then
  echo "== Alignment with 2 mismatches to TEs (multiple mappers):" \
  >> report.txt
  cat  TEs_mapping_results_v2.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e luc_mapping_results_v2.txt ]; then
  echo "== Alignment with 2 mismatches to luciferase (multiple mappers):" \
  >> report.txt
  cat  luc_mapping_results_v2.txt >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e genome_output.sam ]; then
  echo "Total number of genomic mappers:" >> report.txt
  cat genome_output.sam | sed '/^@/d' | cut -f1 | sed 's/^[0-9]*-//' |
  awk '(sum+=$1){}END{print sum}' >> report.txt
fi
echo >> report.txt
if [ -e dm3_viruses_output.sam ]; then
  echo "Total number of mappers to dm3 viruses:" >> report.txt
  cat dm3_viruses_output.sam | sed '/^@/d' | cut -f1 | sed 's/^[0-9]*-//' |
  awk '(sum+=$1){}END{print sum}' >> report.txt
fi
echo >> report.txt
if [ -e dmel-tRNA_output.sam ]; then
  echo "Total number of mappers to dm3 tRNAs:" >> report.txt
  cat dmel-tRNA_output.sam | sed '/^@/d' | cut -f1 | sed 's/^[0-9]*-//' |
  awk '(sum+=$1){}END{print sum}' >> report.txt
fi
echo >> report.txt
if [ -e dmel-miscRNA_output.sam ]; then
  echo "Total number of mappers to dm3 miscRNAs:" >> report.txt
  cat dmel-miscRNA_output.sam | sed '/^@/d' | cut -f1 | sed 's/^[0-9]*-//'
      |
  awk '(sum+=$1){}END{print sum}' >> report.txt
fi
echo >> report.txt
if [ -e synthetic_output.sam ]; then
  echo "Total number of mappers to synthetic sequences (k-mers etc):" \
  >> report.txt
  cat synthetic_output.sam | sed '/^@/d' | cut -f1 | sed 's/^[0-9]*-//' |
  awk '(sum+=$1){}END{print sum}' >> report.txt
fi
echo >> report.txt
if [ -e TEs_output_normalized.sam ]; then
  echo "Total number of reads mapping to TEs:" >> report.txt
  cat TEs_output_normalized.sam | sed '/^@/d' | awk '$2==16' |
```

```
  sed 's/^\([0-9]*\)-\([0-9]*\)/\2/' | sort -k3,3 | groupBy -g 3 -c 1 -o
      sum |
  awk '{printf "%s\t%d\n",$1,$2}' > TE_counts.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e cluster_counts.txt ]; then
  echo "Total number of reads mapping to 42AB and flamenco:" \
  >> report.txt
  echo "ID  reads percent" >> report.txt
  cat cluster_counts.txt | grep -E 'flam|42AB' >> report.txt
fi
echo >> report.txt
echo >> report.txt
if [ -e annotations.summary.reads ]; then
  echo "Normalized read counts per annotation:" >> report.txt
  cat annotations.summary.reads >> report.txt
fi
echo >> report.txt
if [ -e annotations.summary.seqs ]; then
  echo "Normalized sequence counts per annotation:" >> report.txt
  cat annotations.summary.seqs >> report.txt
fi
echo >> report.txt
# report ping-pong information
#cat *_offset_sam_1U10A >> report.txt
#echo >> report.txt

# create miRNA read counts for normalization purposes
bedtools intersect -bed -wo -f 1.00 -a genome_output.bed \
  -b /data/fmuerdte/genomes/dm3_genome/gff-bed/dmel-miRNAs.gff3 |
tawk '{print $4,$15}' | sed 's/^\([0-9]*\)-\([0-9]*\)/\2/' |
  sort -k2,2 | groupBy -g 2 -c 1 -o sum |
awk '{printf "%s\t%d\n",$1,$2}' > miRNA_counts.txt

# remove report files
if [ -e genome_fastq_to_fasta.txt ]; then
  rm genome_fastq_to_fasta.txt
fi
if [ -e genome_fasta_collapser_1.txt ]; then
  rm genome_fasta_collapser_1.txt
fi
if [ -e genome_fasta_collapser_2.txt ]; then
  rm genome_fasta_collapser_2.txt
fi
if [ -e genome_fasta_collapser_unclipped.txt ]; then
  rm genome_fasta_collapser_unclipped.txt
fi
if [ -e genome_fasta_clipper.txt ]; then
  rm genome_fasta_clipper.txt
fi
if [ -e genome_fasta_clipper_unclipped.txt ]; then
  rm genome_fasta_clipper_unclipped.txt
fi
if [ -e genome_normalized_reads.sum ]; then
```

```
  rm genome_normalized_reads.sum
fi
if [ -e annotations.summary.reads ]; then
  rm annotations.summary.reads
fi
if [ -e annotations.summary.seqs ]; then
  rm annotations.summary.seqs
fi
rm *_mapping_results*.txt
```

A.1.2 Molecular signatures of sRNA biogenesis pathways

The interplay between Aub and Ago3 in the biogenesis of secondary piRNAs creates an unusually high number of so called ping-pong partners that overlap by 10nt. Plotting all possible neighbors with different offsets within a certain window can therefore identify the presence or absence of piRNAs originating from this biogenesis pathway. If secondary piRNAs were present within the analyzed sRNA population, one would expect to see a peak at the 10nt overlap mark. Similarly, if the analyzed sRNA population were mostly composed of Dicer2 products, and both guide and passenger strands would be present, one would expect to see such siRNA duplexes as a high number of neighbors overlapping by 19nt (Figure A.1.1).

Before plotting, the neighbors for any given offset have to be identified, which can be done in two ways: one way is to map each sRNA in a library onto all other sRNAs within that library. If any given sRNA maps onto another sRNA in antisense direction and their 5' ends are 10nt apart, this pair can be flagged as a putative ping-pong pair and be counted. Another method is to map all sRNAs of the population against a target reference (i.e. transposon sequences or chromosomes). In a subsequent step, the overlaps between neighbors can be deduced from their mapping locations within the reference sequence. The latter is computationally preferable, since mapping to reference sequences is usually done prior to downstream analysis already, saving the user one mapping step. Once flagged as pairs with a given offset, each sequence of that pair contributes to the count for its offset either by adding 1 or its read count. Both options are available to the user and can be set at run-time. The final output of the script is a list of all identified pairs with their respective offset and sequence or read count (for both the same and the opposite strand). Since the previously mentioned signatures of biogenesis pathways such as the ping-pong signature only arise from sRNA pairs on opposite strands, the same strand information can be discarded.

input/bed_calculate_offsets.pl

```
#!/usr/bin/env perl
=pod
  small RNA offset counter

  Copyright (C) 2011 A. Gordon (gordon at cshl dot edu)
                and F. Muerdter (fmuerdte at cshl dot edu)

  License: Free for personal / academic / non-commercial use.

  See related publication:
     [ Production of artificial piRNAs in flies and mice ]
     [ Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B
        ., Gordon, A., Hannon, G.J., Aravin, A.A. ]
```

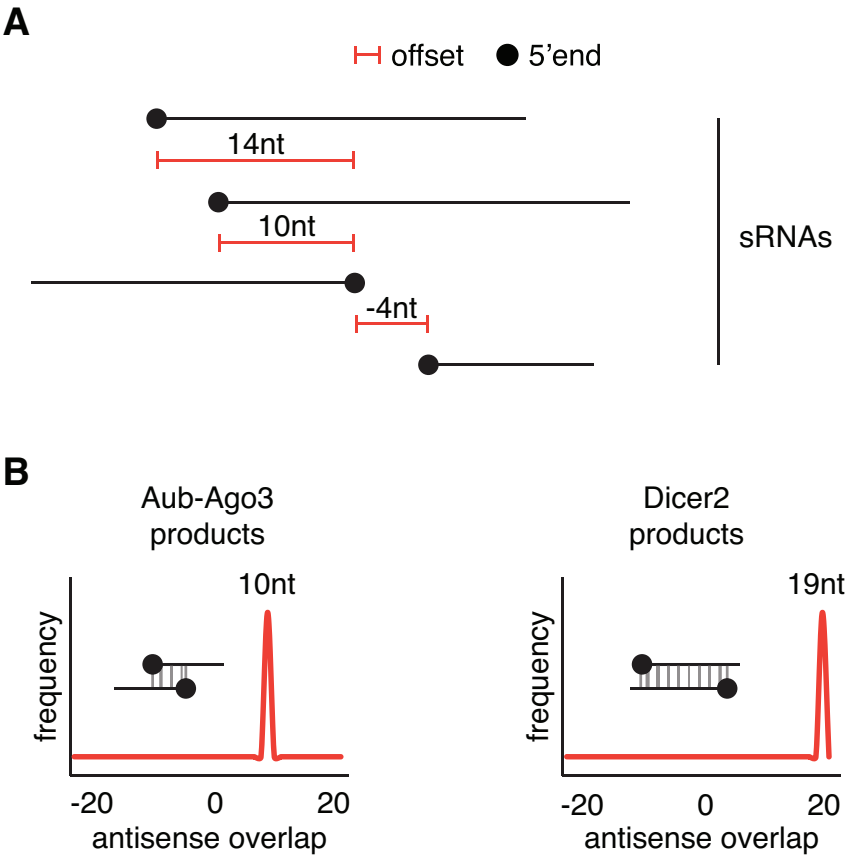


Figure A.1.1: sRNA biogenesis signatures can be detected in total RNA libraries. (A) The distance of 5' ends of small RNA neighbors within a given window can be expressed as offsets. (B) Two different sRNA biogenesis pathways create distinct signatures in sRNA populations.

```
    [ RNA (2012) ]
    [ doi: 10.1261/rna.029769.111 ]

  For questions, please contact the corresponding authors:
    A. A. Aravin, E—mail aaa at caltech dot edu.
  G. J. Hannon, E—mail hannon at cshl dot edu.

  When using this script (or derivatives there—of),
  please cite: [ Muerdter et al., 2012 ]
=cut

##
## Version 0.2
##
use strict;
use warnings;
use Data::Dumper;
use Carp;
use Getopt::Long;


```

```
34
35 ##
36 ## Forward function declarations
37 ##
38 sub parse_command_line();
39 sub show_help();
40 sub set_filter_region($);
41 sub load_external_intervals();
42 sub get_next_interval();
43 sub interval_in_region($$$);
44 sub initialize_empty_windows();
45 sub calculate_region_coverage();
46 sub scan_region_windows();
47 sub sort_intervals_to_process();
48
49 ##
50 ## Global Parameters
51 ##   Set by user on command line
52 my $region_chrom; # Process reads from this region only, ignore all other
       reads
53 my $region_start;
54 my $region_end;
55 my $use_multiplity_count; # Use multiplicity count from read name, column 4
       (N-M) where M is the multiplicity count.
56 my $extend_upstream=10;
57 my $extend_downstream=10;
58 my $header; # print header line
59 my $debug; # print a lot of debugging information to STDERR
60 my $external_intervals; #if set, we calculate the coverage from one BED
       file, but scan the intervals from another BED file
61
62 ## Global variables
63 my $region_size;
64 my $stats_lines_count=0;
65 my $stats_sequences_count=0;
66 my $stats_reads_count=0;
67
68 # Sequences = multiplitiy of 1 (one per line/interval)
69 # Reads = with multiplicity values
70 my @window_positive_sequences;
71 my @window_negative_sequences;
72 my @window_positive_reads;
73 my @window_negative_reads;
74 my @intervals_to_process;
75
76 ##
77 ## Program start
78 ##
79
80 parse_command_line();
81 initialize_empty_windows();
82 print STDERR "Loading intervals..." if $debug;
83 calculate_region_coverage();
84 if ($external_intervals) {
85   print STDERR "Loading external intervals..." if $debug;
```

```
  load_external_intervals();
}
print STDERR "Sorting intervals..." if $debug;
sort_intervals_to_process();
if ($header) {
  print join("\t", qw/chrom genomic_start strand window_position offset
      seq_count_same_strand read_count_same_strand seq_count_opposite_strand
       read_count_opposite_strand/), "\n";
}
scan_region_windows();

if ($debug) {
  print STDERR "Read $stats_lines_count lines\n";
  print STDERR "Processed $stats_sequences_count intervals (
      $stats_reads_count reads) in region $region_chrom:$region_start-
      $region_end\n";
}

##
## Program End
##

sub show_help()
{
  print<<EOF;
small RNA offset counter
Copyright (C) 2011 A. Gordon   (gordon\@cshl.edu)
               and F. Muerdter (fmuerdte\@cshl.edu)

  License: Free for personal / academic / non-commercial use.

  See related publication:
     [ Production of artificial piRNAs in flies and mice ]
     [ Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B
        ., Gordon, A., Hannon, G.J., Aravin, A.A. ]
     [ RNA (2012) ]
     [ doi: 10.1261/rna.029769.111 ]

  For questions, please contact the corresponding authors:
    A. A. Aravin, E-mail aaa\@caltech.edu.
  G. J. Hannon, E-mail hannon\@cshl.edu.

  When using this script (or derivatives there-of),
  please cite: [ Muerdter et al., 2012 ]

Usage: $0 [OPTIONS] INPUT.BED

Options:
 --help    =    This help screen.

 -r chrX:start-end
 --region chrX:start-end  =  Specify region of interest. Reads outside this
      region will be ignored.

 --up N         = Start window N upstream of the intervals 5" position.
```

```
 --down N       = End window N downstream of the intervals 5" position.

 -m
 --mult         = Take multiplicity count from name column (4th column).
                  Multiplicity should be in "N-M" format, where M is the
                     multiplicity value.

 --intervals FILE.BED = Compare the intervals in FILE.BED against the
    coverage of INPUT.BED.
                        By default, the intervals in INPUT.BED are compared
                             against their own coverage.

 -h
 --header       = Print header line.

 --debug        = Print debug information

Example:

\$ cat test.bed
chr1  95  125 1-7 0 +
chr1  100 150 1-13  0 +
chr1  100 140 1-4 0 +
chr1  80  111 1-20  0 -

\$ $0 --region chr1:0-2000 --mult --up 10 --down 10 --header test.bed
chrom genomic_start strand  window_position offset  seq_count_same_strand
    read_count_same_strand  seq_count_opposite_strand
    read_count_opposite_strand
chr1  95  + 95  5 2 17  0 0
chr1  100 + 100 -5  1 7 0 0
chr1  100 + 100 0 1 13  0 0
chr1  100 + 100 10  0 0 1 20
chr1  100 + 100 -5  1 7 0 0
chr1  100 + 100 0 1 4 0 0
chr1  100 + 100 10  0 0 1 20
chr1  110 + 110 -10 0 0 2 17

EOF
  exit(1);
}

sub parse_command_line()
{
  my $region;

  my $rc = GetOptions(
       "help" => \&show_help,
       "region|r=s" => \$region,
       "mult|m" => \$use_multiplity_count,
       "debug" => \$debug,
       "up|upstream=i" => \$extend_upstream,
       "down|downstream=i" => \$extend_downstream,
       "intervals=s" => \$external_intervals,
       "header|h" => \$header,
```

```
185     );
186
187   die "Error: invalid command line options.\n" unless $rc;
188   die "Error: missing region of interest (example: --region chrX:1000-2000)
          .\n" unless $region;
189   die "Error: external intervals file ($external_intervals) not found\n" if
          defined $external_intervals && ! ( -e $external_intervals ) ;
190
191   set_filter_region($region);
192 }
193
194 sub set_filter_region($)
195 {
196   my ($region) = shift or carp;
197
198   ($region_chrom, $region_start, $region_end) = $region =~ /^([\w-]+):(\d+)
          -(\d+)$/
199     or die "Error: invalid region ($region), expecting 'chrX:1000-2000'.\n"
          ;
200
201   $region_end--; #we assume the BED END position is 1-based
202
203   die "Error: invalid region ($region), start is bigger than end position.\
          n" if ($region_start > $region_end);
204
205   die "Error: invalid region ($region), size (end-start) is zero.\n" if (
          $region_start == $region_end);
206
207   $region_size = $region_end - $region_start  + 1;
208
209   print STDERR "Filter region: $region_chrom:$region_start-$region_end\n"
          if $debug;
210 }
211
212 =pod
213   small RNA offset counter
214
215   Copyright (C) 2011 A. Gordon (gordon at cshl dot edu)
216                  and F. Muerdter (fmuerdte at cshl dot edu)
217
218   License: Free for personal / academic / non-commercial use.
219
220   See related publication:
221      [ Production of artificial piRNAs in flies and mice ]
222      [ Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B
          ., Gordon, A., Hannon, G.J., Aravin, A.A. ]
223      [ RNA (2012) ]
224      [ doi: 10.1261/rna.029769.111 ]
225
226   For questions, please contact the corresponding authors:
227     A. A. Aravin, E-mail aaa at caltech dot edu.
228   G. J. Hannon, E-mail hannon at cshl dot edu.
229
230   When using this script (or derivatives there-of),
231   please cite: [ Muerdter et al., 2012 ]
```

```
232 =cut
233
234 sub get_next_interval()
235 {
236   my $line = <>;
237   return () unless defined $line; #End of file ?
238   chomp $line;
239
240   my ($chrom, $start, $end, $name, $quality, $strand, @other) = split /\t/,
          $line;
241
242   # Do some input validation
243   die "Input error: not enough columns in line $. (expecting at least 6 BED
          columns)\n" unless defined $strand;
244   die "Input error: bad start value ($start) on line $. (expecting a number
       )\n" unless $start =~ /^\d+$/;
245   die "Input error: bad end value ($end) on line $. (expecting a number)\n"
          unless $end =~ /^\d+$/;
246   die "Input error: bad strand value ($strand) on line $. (expecting + or
       -)\n" unless $strand eq "+" || $strand eq "-";
247
248   my $multiplicity = 1 ;
249
250   #extract multiplicity value from read name
251   if ($use_multiplity_count) {
252     (undef, $multiplicity) = $name =~ /^(\d+)-([-+]?[0-9]*\.?[0-9]+([eE
          ][-+]?[0-9]+)?)$/
253       or die "Input error: invalid multiplicity value on column 4 ($name)
             on line 4. (expecting N-M)\n";
254   }
255
256   return $chrom, $start, $end-1, $name, $strand, $multiplicity;
257 }
258
259 sub interval_in_region($$$)
260 {
261   my ($chrom, $start, $end) = @_ or carp;
262
263   return undef unless $chrom eq $region_chrom;
264   return undef unless $start >= $region_start;
265   return undef unless $end <= $region_end;
266
267   return 1;
268 }
269
270 sub initialize_empty_windows()
271 {
272   $#window_positive_sequences = $region_size;
273   $#window_negative_sequences = $region_size;
274   $#window_positive_reads = $region_size;
275   $#window_negative_reads = $region_size;
276
277   foreach my $index ( 0 .. $region_size ) {
278     $window_positive_sequences[$index] = 0 ;
279     $window_negative_sequences[$index] = 0 ;
```

```
280     $window_positive_reads[$index] = 0 ;
281     $window_negative_reads[$index] = 0 ;
282   }
283 }
284
285 sub load_external_intervals()
286 {
287   open FILE,"<",$external_intervals
288     or die "Error: failed to open external intervals file (
          $external_intervals): $!\n";
289   while ( my $line = <FILE> ) {
290     chomp $line;
291     my ($chrom, $start, $end, $name, $quality, $strand, @other) = split /\t
          /, $line;
292
293     # Do some input validation
294     die "Input error: not enough columns in file $external_intervals line $
          . (expecting at least 6 BED columns)\n" unless defined $strand;
295     die "Input error: bad start value ($start) in file $external_intervals
          line $. (expecting a number)\n" unless $start =~ /^\d+$/;
296     die "Input error: bad end value ($end) in file $external_intervals line
           $. (expecting a number)\n" unless $end =~ /^\d+$/;
297     die "Input error: bad strand value ($strand) in file
          $external_intervals line $. (expecting + or -)\n" unless $strand eq
          "+" || $strand eq "-";
298
299     my $multiplicity = 1 ;
300
301     push @intervals_to_process, [ $chrom, $start, $end, $name, $strand,
          $multiplicity ] ;
302   }
303   close FILE;
304 }
305
306 sub calculate_region_coverage()
307 {
308   while (my ($chrom, $start, $end, $name, $strand, $multiplicity) =
        get_next_interval()) {
309     $stats_lines_count++;
310     next unless interval_in_region($chrom, $start, $end);
311     $stats_sequences_count++;
312     $stats_reads_count+=$multiplicity;
313
314     if ( $strand eq "+" ) {
315       $window_positive_sequences[$start - $region_start] += 1;
316       $window_positive_reads[$start - $region_start] += $multiplicity;
317     } else {
318       $window_negative_sequences[$end - $region_start] += 1;
319       $window_negative_reads[$end - $region_start] += $multiplicity;
320     }
321
322     # load the intervals from the same input file used for coverage
          calculations
323     if (! defined $external_intervals ) {
```

```
324       push @intervals_to_process, [ $chrom, $start, $end, $name, $strand,
              $multiplicity ] ;
325     }
326   }
327 }
328
329 sub sort_intervals_to_process()
330 {
331   ## this copmlicated sort function sorts by: Strand [4], then Start[1],
          then end [2]
332   @intervals_to_process = sort {
333         ( $a->[4] ne $b->[4] ) ? ( $a->[4] cmp $b->[4] ) :
334           (
335             ( $a->[1] != $b->[1] ) ? ( $a->[1] <=> $b->[1] ) :
336             (
337              $a->[2] <=> $b->[2]
338             )
339           );
340         } @intervals_to_process ;
341 }
342
343 sub scan_region_windows()
344 {
345   ## Iterave over every nucleotide in the region
346   foreach my $interval_ref ( @intervals_to_process ) {
347
348     my ($interval_chrom, $interval_start, $interval_end,
349       $interval_name, $interval_strand, $interval_multiplicity ) = @{
              $interval_ref};
350
351     print STDERR "processing interval $interval_chrom:$interval_start-
          $interval_end strand $interval_strand name '$interval_name'
          multiplicity $interval_multiplicity\n" if $debug;
352
353     my $nuc_position = (( $interval_strand eq "+") ? $interval_start :
          $interval_end ) -  $region_start;
354     my $genomic_nuc_position = $region_start + $nuc_position;
355
356
357     ##
358     ## Calculate the window around this read (up/down stream is relative to
           the interval's strand)
359     ##
360     my ($extended_start, $extended_end);
361     if ( $interval_strand eq "+" ) {
362       $extended_start = $nuc_position - $extend_upstream;
363       $extended_end   = $nuc_position + $extend_downstream;
364     } else {
365       $extended_start = $nuc_position - $extend_downstream;
366       $extended_end   = $nuc_position + $extend_upstream;
367     }
368     $extended_start = 0 if $extended_start < 0 ;
369     $extended_end = $region_size-1 if $extended_end > $region_size-1;
370
371     ## Scan the window around (before/after) the nucleotide position
```

```
372     foreach my $window_index ( $extended_start .. $extended_end ) {
373       my ($sequences_count_same, $sequences_count_opposite,
374         $reads_count_same, $reads_count_opposite, $offset);
375
376       ##
377       ## Find the counts of reads/intervals in this window position
378       ##
379       if ( $interval_strand eq "+" ) {
380         $sequences_count_same = $window_positive_sequences[$window_index];
381         $reads_count_same = $window_positive_reads[$window_index];
382         $sequences_count_opposite = $window_negative_sequences[
              $window_index];
383         $reads_count_opposite = $window_negative_reads[$window_index];
384
385         $offset = $window_index - $nuc_position ;
386       } else {
387         $sequences_count_same = $window_negative_sequences[$window_index];
388         $reads_count_same = $window_negative_reads[$window_index];
389         $sequences_count_opposite = $window_positive_sequences[
              $window_index];
390         $reads_count_opposite = $window_positive_reads[$window_index];
391
392         $offset = $nuc_position - $window_index;
393       }
394
395       # At offset 0, subtract the current interval from the counts
396       # (only if NOT using external intervals )
397       if ( $offset == 0 && (!defined $external_intervals) ) {
398         $sequences_count_same -= 1 ;
399         $reads_count_same -= $interval_multiplicity;
400       }
401
402       if ( $sequences_count_same != 0 || $reads_count_same != 0 ||
403         $sequences_count_opposite != 0 || $reads_count_opposite != 0) {
404         print $region_chrom, "\t",
405           $genomic_nuc_position, "\t",
406           "+", "\t",
407           $nuc_position, "\t",
408           $offset, "\t",  ## This is the window offset
409           $sequences_count_same, "\t",
410           $reads_count_same, "\t",
411           $sequences_count_opposite, "\t",
412           $reads_count_opposite,
413           "\n";
414       }
415     }
416   }
417 }
```

One easy way to process the output into more meaningful numbers is grouping by the offset (column 5) and summing up the read count (column 9).

Listing A.8: Process script output for plotting

```
1 bed_calculate_offsets.pl --region chr1:0-2000 --mult --up 20 \
```

```
  --down 20 --header INPUT.bed > OUTPUT.txt
cat OUTPUT.txt | sed 1d | sort -k5,5 | groupBy -g 5 -c 9 -o sum
```

The resulting offsets and read counts can then be plotted in a bargraph. The y-axis should either display absolute read counts, or z-scores of offset frequency over the given window. Using z-scores allows the user to compare offset frequency plots from libraries with different read counts.

A.1.3 Analysis pipeline for the discovery of apiRNAs

The main principles of small RNA library analysis presented in appendix A.1.1 apply equally to the discovery of artificial piRNAs generated from exogenous sequences such as lacZ or GFP (see also chapter 2.1).

The major difference between the analysis of standard small RNA libraries and libraries generated from flies or mice harboring tagged piRNA clusters, was to take all reads of the library that did not map to any wildtype chromosome, and map those to the artificially introduced sequences. Each cassette (lacZ, GFP, etc.) had its own dedicated mapping index (generated with bowtie-build). The standard mapping subscript was designed in a way that the mapping index was variable, rather than being hard-coded into the script. Instead, the index was set within the main script and the subscript was invoked repeatedly after setting and exporting the variable defining the appropriate mapping index. This created output files specific to each reference sequence, which were then further analyzed with the same subscripts presented in appendix A.1.1.

Listing A.9: Repeated invocation of the same subscript with different mapping indexes

```
# set id to genome
ID="dm3_genome"
export ID

# invoke mapping subscript
~/bin/apiRNA_pipeline_sub/mapping.sh

# set id to lacZ
ID="lacZ"
export ID

# invoke mapping subscript
~/bin/apiRNA_pipeline_sub/mapping.sh

# set id to GFP
ID="GFP"
export ID

# invoke mapping subscript
~/bin/apiRNA_pipeline_sub/mapping.sh

# etc.
```

After setting the appropriate index, the same mapping subscript can be called repeatedly.

Listing A.10: An example mapping subscript using variable mapping indexes

```
#!/bin/sh

# set temporary directory for genome mapping
DIR=$(mktemp -d --tmpdir=. -t mapping.XXXXXXX) || exit 1

set -o pipefail

# set the proper mapping index based on the variable ID, set within the
   main script
INDEX=~/genomes/"$ID"

# Map the entire library to the genome, if the ID is set to genome,
# map only non-mapping reads otherwise
if [ $ID == "dm3_genome" ];
    then
      DATA="$BASENAME"_collapsed.fa
  else
    DATA=genome_non_mappers.txt
fi



echo "Mapping with zero mismatches to "$ID"..."
# Map with bowtie, zero mismatches, keep the non-mappers
bowtie -f -v 0 -a -m 1 -p 16 --sam --max "$ID"_max_mappers_v0.txt \
  --un "$ID"_non_mappers_v0.txt \
  "$INDEX" "$DATA" > "$ID"_output_v0.sam \
  2> "$ID"_mapping_results_v0.txt

# clean up temporary directory
rm -r $DIR || exit 1
```

A.1.4 Analysis pipeline for RNA-seq data

In general, the subscripts presented in appendix A.1.1 can be used to analyze RNA-seq data. However, RNA-seq data from total RNA input differs in some ways from small RNA sequencing data. RNA-seq libraries usually have a higher complexity (more unique sequences per read count), which made collapsing the library before mapping unnecessary. Also, cDNA from random primed total RNA consists of longer fragments, lowering the chance of sequencing into the 3' ligated adapter. This makes clipping of adapter sequences obsolete. Depending on the underlying biological question, pair-end sequencing and accordingly mapping of read mates may be preferable to the single-end read sequencing and mapping demonstrated in appendix A.1.1. As mentioned previously, the sequenced read may be polymorphic to the reference sequence, which is why up to two mismatches are commonly allowed in a typical small RNA (~20nt) read. Given that RNA-seq reads are usually longer when sequenced on Illumina platforms (up to 100nt), the possibility of more polymorphisms has to be taken into account. A seed and extend strategy (bowtie

option –n) may be preferable to the mapping strategy used for small RNA reads (bowtie option –v). Also, different mapping algorithms more sensitive for longer read may be considered (i.e. bowtie2).

After mapping, normalization of read counts is usually not necessary, since downstream applications such as DESeq (an R package to test for differential expression at the gene level (Anders and Huber, 2010)) use raw read counts for their statistical method. Instead, calculating total read counts mapping to any genomic feature (provided in a features file in bed or gff format) can be done using available tools such as htseq-count (available from http://www-huber.embl.de/users/anders/HTSeq/doc/count.html) or the bedtools package (available from https://code.google.com/p/bedtools/).

A.2 Manuscripts

Muerdter, F.*, Guzzardo, P.M.*, Gillis, J., Luo, Y., Yu, Y., Chen, C., Fekete, R., Hannon, G.J. (2013) A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila. *Molecular Cell*, 50 (5), pp. 736–748.
DOI: 10.1016/j.molcel.2013.04.006.

Guzzardo, P.M., Muerdter, F., Hannon, G.J. (2013) The piRNA pathway in flies: highlights and future directions. *Current Opinion in Genetics and Development*, 23 (1), pp. 44-52.
DOI: 10.1016/j.gde.2012.12.003.

Muerdter, F.*, Olovnikov, I.*, Molaro, A.*, Rozhkov, N.V.*, Czech, B., Gordon, A., Hannon, G.J., Aravin, A.A. (2012) Production of artificial piRNAs in flies and mice. *RNA*, 18 (1), pp. 42-52.
DOI: 10.1261/rna.029769.111.

*These authors contributed equally to this work

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

Molecular Cell
Resource

A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*

Felix Muerdter,[1,3,6] Paloma M. Guzzardo,[1,6] Jesse Gillis,[1,2] Yicheng Luo,[1] Yang Yu,[1] Caifu Chen,[4] Richard Fekete,[5] and Gregory J. Hannon[1,*]
[1]Watson School of Biological Sciences, Howard Hughes Medical Institute
[2]The Stanley Institute for Cognitive Genomics
Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[3]Zentrum für Molekularbiologie der Pflanzen, Entwicklungsgenetik, University of Tübingen, 72076 Tübingen, Germany
[4]Genetic Analysis R&D, Life Technologies Corporation, Foster City, CA 94404, USA
[5]Molecular Cell Biology R&D, Life Technologies Corporation, Austin, TX 78744, USA
[6]These authors contributed equally to this work
*Correspondence: hannon@cshl.edu
http://dx.doi.org/10.1016/j.molcel.2013.04.006

SUMMARY

A large fraction of our genome consists of mobile genetic elements. Governing transposons in germ cells is critically important, and failure to do so compromises genome integrity, leading to sterility. In animals, the piRNA pathway is the key to transposon constraint, yet the precise molecular details of how piRNAs are formed and how the pathway represses mobile elements remain poorly understood. In an effort to identify general requirements for transposon control and components of the piRNA pathway, we carried out a genome-wide RNAi screen in *Drosophila* ovarian somatic sheet cells. We identified and validated 87 genes necessary for transposon silencing. Among these were several piRNA biogenesis factors. We also found *CG3893* (*asterix*) to be essential for transposon silencing, most likely by contributing to the effector step of transcriptional repression. Asterix loss leads to decreases in H3K9me3 marks on certain transposons but has no effect on piRNA levels.

INTRODUCTION

Transposable elements populate virtually every eukaryotic genome. Although their presence can impart some benefits, when not properly controlled, transposons can compromise the genomic integrity of their host and its offspring. Hence, in all higher animals there are control mechanisms in place that prevent wholesale transposon mobilization (Malone and Hannon, 2009).

In *Drosophila*, the principal pathway that protects the inheritable genome is comprised of a catalog of small RNAs that interact with an animal-specific clade of Argonaute family proteins: the Piwi proteins (Ishizu et al., 2012). This pool of Piwi-interacting RNAs (piRNAs), which contains millions of distinct sequences bearing homology to transposable elements, functions as a molecular memory of transposon identity (Brennecke et al., 2007). Using their bound piRNAs as a guide, Piwi proteins (Piwi, Aubergine, and Argonaute 3) recognize and silence their targets. Failure of this pathway leads to defects in germline development and to sterility (Khurana and Theurkauf, 2010).

Genetic studies have uncovered a number of loci that are essential for the proper function of the piRNA pathway. Besides the core proteins of the Piwi clade, *flamenco* has long been implicated in transposon control. This discrete locus on the X chromosome of *Drosophila* was found to be a major determinant for silencing of the retroelement *gypsy* almost two decades ago, though its underlying molecular nature remained mysterious (Bucheton, 1995). Sequencing small RNAs bound to Piwi proteins and mapping these sequenced reads back to the genome revealed the true nature of the *flamenco* locus. Rather than being a protein coding gene, the *flamenco* transcript is the precursor to the majority of piRNAs expressed in the follicle cells of the ovary (Brennecke et al., 2007). This and other sites of abundant piRNA generation were termed piRNA clusters. What mechanisms mark *flamenco* and other cluster transcripts for processing into piRNAs remains largely unclear. Several studies have shed some light on this topic by identifying some of the protein factors that play a role in piRNA cluster transcription and transport. Rhino and Cutoff, as well as histone methylation marks deposited by Eggless (EGG), are necessary for cluster transcription (Klattenhoff et al., 2009; Pane et al., 2011; Rangan et al., 2011). In addition, UAP56, a helicase implicated in splicing and RNA export, was found to bind a germline piRNA cluster RNA and may escort the transcript from the nucleus to the nuage for processing (Zhang et al., 2012). Intriguingly, Rhino, Cutoff, and UAP56 all were reported to be specific to germline piRNA clusters, leaving factors involved in somatic cluster determination a mystery.

Mutagenesis screens for sterility phenotypes in *Drosophila* also highlighted factors that later were found to be elements of the piRNA pathway (Schüpbach and Wieschaus, 1989, 1991). Molecular analyses have begun to place these factors at specific

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

steps of the pathway, such as being required during the initiation and biogenesis phase or during the effector phase and silencing of transposons. However, very little is known about specific biochemical functions or enzymatic activities of the proteins that act at each step. One major insight came from the discovery of a trimming activity in insect cell lysates that shortens the 3′ end of putative piRNA precursors to their mature length (Kawaoka et al., 2011). However, the protein responsible for this activity remains unknown. Biochemical and structural studies of Zucchini (ZUC), which was previously implicated in the piRNA pathway, suggest it as a promising candidate for the nuclease that creates the 5′ ends of primary piRNAs (Ipsaro et al., 2012; Nishimasu et al., 2012). Whether Zucchini and a trimming enzyme together comprise the complete primary biogenesis machinery or other endo- or exonucleolytic cleavage events create intermediates that are further matured remains unknown.

Another enigmatic aspect of the pathway is precisely how Piwi-piRNA complexes silence their targets. In the case of somatic cells of the ovary, it has become clear that control of transposons occurs at the transcriptional level through Piwi-directed deposition of epigenetic marks. Recently, three conclusive studies showed that, upon Piwi depletion, transposons engage in active transcription and show a depletion of H3K9 trimethyl (H3K9me3) (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). In one of these studies, the authors place Maelstrom (MAEL) at the effector step of transcriptional repression (Sienski et al., 2012). Interestingly, loss of *mael* derepresses transposons without preventing H3K9me3 deposition, indicating that this modification may not be the definitive silencing mark. What the final silencing mark may be, and which proteins are responsible for establishing repressive chromatin marks over transposons, has yet to be determined.

Much of what has been learned of the piRNA pathway that operates in follicle cells relied on the use of a cultured ovarian somatic sheet (OSS) cell line (Niki et al., 2006). This cell line expresses microRNAs (miRNAs), endogenous small interfering RNAs (endo-siRNAs), and piRNAs (Lau et al., 2009; Saito et al., 2009). With an active siRNA and primary piRNA pathway in place, genetic requirements for primary piRNA biogenesis and transposon silencing can be investigated.

Here, we describe a genome-wide screen that builds a foundation for addressing many open questions relevant to piRNA biogenesis and effector functions. By individually assaying more than 41,000 long, double-stranded RNAs (dsRNAs), targeting every annotated gene in the *Drosophila* genome, and examining their effect on transposon expression levels, we describe a comprehensive genetic framework for transposon control. We reveal piRNA biogenesis factors and place proteins, which to our knowledge have not been implicated in the piRNA pathway, at the effector step.

RESULTS

An RNAi Screen for Elements of the Somatic piRNA Pathway

In order to assay transposon derepression upon knockdown (KD) of any given target gene, we established a sensitive assay for *gypsy* messenger RNA (mRNA) levels. Based on quantitative PCR (qPCR) with hydrolysis probes (TaqMan), this assay specifically detects the spliced subgenomic transcript of the retrotransposon (Figure 1A). The expression of this transcript is known to be highly sensitive to disruption of the piRNA pathway, even more so than its unspliced counterpart (Pélisson et al., 1994).

Knockdown of target genes was accomplished by transfecting dsRNAs from two independent genome-wide libraries with a total of 41,342 dsRNAs. The average count of dsRNAs per gene was 2.28, targeting 13,914 genes with valid IDs in FlyBase (McQuilton et al., 2012). Additionally, the two libraries contained 1,045 negative controls, 2,097 dsRNAs without an annotated target, and 2,301 dsRNAs targeting the Heidelberg collection of predicted genes.

We transfected OSS cells with individual dsRNAs in 48-well plates, lysed the cells 5 days later, and used the lysate for reverse transcription (Figure 1A). The qPCR results were normalized to their respective plate using *Z* scores (Ramadan et al., 2007). As a secondary metric, we calculated the fold change in relation to the median. The knockdown of *piwi* in this setting led to a *gypsy* mRNA signal that was detectable by qPCR much earlier than the average of the plate. In four biological replicates of *piwi* knockdown, the average normalized signal for *gypsy* was almost 5 SDs away from the median of its plate (Figure S1A). Hence, our assay for transposon derepression is both sensitive and robust, at least given that the RNA interference (RNAi) trigger is of good quality. When comparing several independent dsRNAs, we saw that there was considerable variance in this respect. dsRNAs against known components of the pathway, such as *armitage* (*armi*), led to consistent derepression of *gypsy*; however, levels varied from 25- to 70-fold (Figure S1B). Since we assayed several dsRNAs against each annotated gene, we felt confident that the majority of pathway components could be identified.

Out of 41,342 tested dsRNAs, 33,780 met our criteria for inclusion in our analysis; of these, 320 dsRNAs met the criteria for primary hit selection (Figure 1B and Table S1). Included in this list were 18 dsRNAs without annotated targets, which were disregarded. All genes that were previously implicated in *gypsy* control were outliers in the primary screen (Figure S1B). Knockdown of pathway components such as *capsuleen*, *hen1*, *egg*, or *squash* was not expected to cause strong *gypsy* derepression based on existing literature and indeed did not cause derepression of *gypsy* in our assay (Olivieri et al., 2010; Rangan et al., 2011). After choosing *Z* score and fold change cutoffs for hit selection based on 217 green fluorescent protein (GFP) negative controls, only 3 out of 645 (0.5%) additional negative controls scored as weak hits (Figure S1C).

To ask whether genes that affect transposon control show preference toward specific annotation groups, we performed functional enrichment analysis on our primary data set. After multiple testing correction, 215 functions were associated with transposon silencing (corrected $p < 0.05$), many with strong potential relevance and very significant enrichment (the top 20 functions have a corrected $p < 1 \times 10^{-6}$; Table S2). Among the most significant, we find expected cellular components like the Yb body, but also more surprising functions like regulation of growth of symbiont in host. While genes implicated in the piRNA pathway drive several of these enrichments, all scoring highly in the primary screen, our candidate hits intersect with these in some of the

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

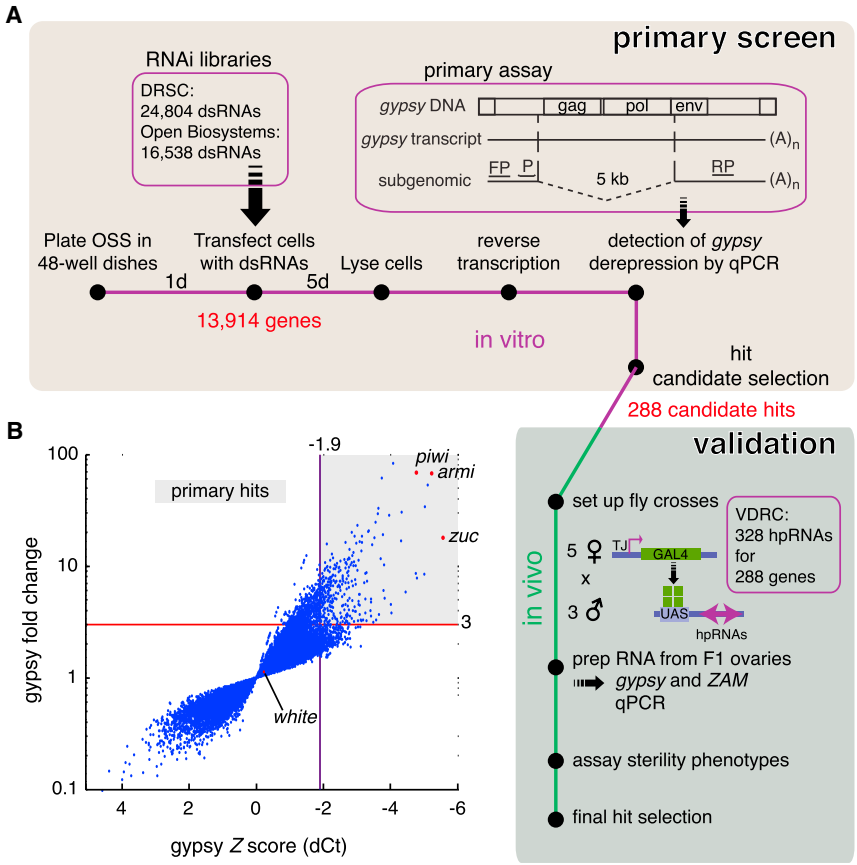


Figure 1. A Genome-wide RNAi Screen for piRNA Pathway Components Acting in the Somatic Compartment of *Drosophila* Ovaries

(A) A workflow of the primary RNAi screen in ovarian somatic sheet (OSS) cells and validation of primary hit candidates in vivo is shown. Each gene in the *Drosophila* genome was knocked down with one or more dsRNAs. At 5 days after transfection, cells were tested for increased levels of the *gypsy* retrotransposon. The primers and the hydrolysis probe used for the qPCR are shown (FP, forward primer; P, hydrolysis probe; RP, reverse primer). The dashed line indicates the ~5 kb segment not present in the subgenomic transcript. We further tested 288 genes in vivo using the Gal4/UAS system to drive hairpin RNAs (hpRNAs) within the *traffic jam* (TJ) expression domain.

(B) All transfected wells were assayed for levels of *gypsy* and one reference gene for normalization. Levels of *gypsy* expression are displayed as *Z* scores and fold change. The cutoffs for both *Z* score (<−1.9) and fold change (>3) are indicated as red lines. The shaded area shows the selection of primary hit candidates. Three positive controls (*piwi*, *armi*, *zuc*) and one negative control (*white*) are marked as red dots. Only wells that passed the filter for primary data point selection are shown. For all primary data points see Table S1. See also Figure S1.

enriched functions. For example, under dorsal appendage formation, *smt3* (SUMO) joins *armi* and *zuc* (Nie et al., 2009). Further work will be necessary to determine to what extent these unexpected intersections relate to biologically relevant connections.

Next, we compared protein interaction data to the full-ranked fold change list. For every gene in the genome, the degree to which that gene's interaction partners scored highly in the fold change list was measured (as receiver operating characteristics [ROC]). Of the top 20 genes with interaction partners significantly elevated in the fold change list, 18 are annotated as belonging to the proteasomal complex (which had 65 genes in total). This observation was remarkably significant ($p < 1 \times 10^{-40}$ after multiple testing correction), which may be due partially to the correlated interaction profiles of the proteasome complex. The two remaining genes not belonging to the proteasomal complex were *bx42*, a homolog of mammalian Skip (SKI-interacting protein), a protein implicated in splicing, and *calypso*, a histone 2A deubiquitinase (Makarov et al., 2002; Scheuermann et al., 2010). Both genes are highly expressed in ovaries, according to the modENCODE tissue expression data. Whether their interaction with particularly high scoring genes is indicative of any regulatory function remains to be tested.

In Vivo Validation of Primary Hits

We obtained 328 fly lines from the Vienna *Drosophila* RNAi Center (VDRC) containing inducible hairpin RNAs (hpRNAs) against 288 of our primary hit candidates (Dietzl et al., 2007). When crossed to females expressing a follicle cell-specific Gal4 driver (*traffic jam*), these hpRNAs can effectively knock down any given target gene within the same expression domain (Tanentzapf et al., 2007). Using hpRNAs against *aub* as a negative control, we observed highly significant changes in *gypsy* expression by qPCR when knocking down *armi* (Figure S2A). We confirmed the effect on the piRNA pathway by measuring two additional transposons, *ZAM* and *gypsy3*. All known pathway components that were identified as primary hits were validated using this approach except for Piwi, which showed developmental defects upon knockdown (Figure S2B). Harnessing this in vivo system, we validated 87 out of the 288 primary hit candidates (Figure 2A and Table S3). In order to be considered as validated, knockdown of the target gene had to result in upregulation of *gypsy* or *ZAM* by at least 2-fold. For crosses with male flies originating from the GD library (first generation) from VDRC, we used *ZAM* as a metric; for the KK library (second generation) we measured *gypsy* levels. This decision was based on the finding that negative controls from the GD library already showed higher basal levels of *gypsy* subgenomic transcript when compared to the KK library negative controls (Figures S3A–S3C).

Out of the 288 candidate hits, knockdown of 52 genes, including Piwi, led to such severe developmental defects that dissection of ovaries and confirmation of the initial screen result was not possible. However, several arguments can be made that this category harbors a substantial number of true hits. First, the genes in this category had an average *gypsy* fold change of 5.8 in

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

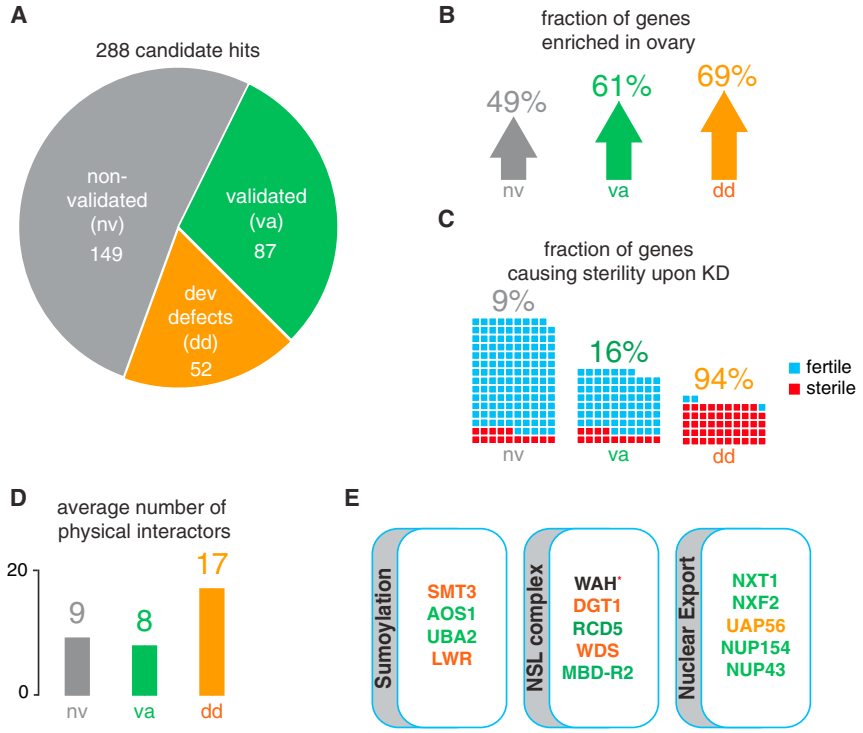


Figure 2. Primary Candidates Were Validated In Vivo

(A) The number of hit candidates that validated (va) or did not validate (nv) in vivo is shown. Genes that caused severe developmental defects upon knockdown and therefore could not be assayed are also indicated (dd). A full list of validated fly lines and corresponding transposon derepression information is available in Table S3.

(B) Validated hits are preferentially expressed in ovaries. The percentage of genes that are enriched in ovaries compared to whole fly is shown for the three categories. These data are based on mRNA signals on Affymetrix expression arrays available from FlyAtlas (Chintapalli et al., 2007).

(C) The fraction of genes causing sterility upon knockdown is shown. Each small box represents one gene, with blue and red indicating if flies were fertile or sterile, respectively, upon knockdown.

(D) The degree to which a gene may represent a node in a network in each of the classifications was measured by the number of physical interactors. Interaction data from BioGRID was used for this analysis (Breitkreutz et al. 2008).

(E) Components of the *Drosophila* sumoylation pathway, the nonspecific lethal complex, and proteins involved in nuclear export are primary hits that validate in vivo. WAH could not be validated in vivo because no RNAi fly was available at the time of submission (red asterisk). The text coloring of each gene indicates the result of the validation screen and is consistent with the categories in (A). See also Figure S2.

the primary data set, as compared to 3.3 for the primary hits that failed secondary validation. This average fold change was even higher than the validated subset (5.1-fold). Since primary fold changes as well as *Z* scores are a function of precision (the likelihood of a primary hit to be validated), this is indicative of the biological relevance of these hits (Figure S2C). Second, while the nonvalidated genes had a representation count in the dsRNA libraries of 2.84, the developmental defect set had a count of 2.42, which is significantly different for the two sets ($p \approx 0.0035$, Wilcoxon test). Thus, the genes of the developmental defect category were disadvantaged according to their annotation class with respect to their possibility to be a primary hit by chance in comparison to the candidates that failed validation, yet they had a much higher average fold change.

Both the validated and developmental defect sets were significantly enriched for genes with higher expression levels in ovaries as compared to whole fly (Figure 2B). While knockdown of genes within the nonvalidated category only led to sterility of 9% of the crosses, the fraction was 16% for the validated and 94% for the developmental defect set (Figure 2C). The extreme phenotypes we observe in the developmental set could imply more generic functions for these genes than simply action in the piRNA pathway. Indeed, around 90% have an average of 17 physical interactors, which is significantly higher than the other categories (Figure 2D).

When ranked by their fold changes in the validation screen, the majority of the somatic piRNA pathway components were among the strongest hits (Table 1). However, genes that, to our knowledge, have not been implicated in the pathway, scored highly as well. *nxt1*, a nuclear export factor, ranks first with *gypsy* levels almost 2,500-fold higher than the negative control (Figure 2E) (Herold et al., 2001). In addition, depletion of NXT1 also led to sterility. Knockdown of the RNA helicase *uap56*, which acts in the same export pathway (Herold et al., 2003), showed derepression of *gypsy* of up to 8-fold in the primary screen. First implicated in splicing, *uap56* was recently shown to be involved in transport of the primary piRNA transcript of dual-stranded clusters to the nuclear pore (Zhang et al., 2012). The knockdown of *uap56* in follicle cells affected germline development to such an extent that in vivo verification of the primary screen results was not possible. We were able to validate another mRNA export factor, *nxf2*, which interacts with *nxt1* (Table S3) (Herold et al., 2001).

NXT1 was previously reported to affect interactions with the nuclear pore complex as well (Lévesque et al., 2001). Hence, the presence of several nuclear pore components within the top 20 hits is not surprising: both *nup154* and *nup43* knockdowns caused similar derepression of *gypsy* (Table 1). Additionally, NUP154-deficient flies were sterile in our assay.

Another two genes ranking among the top 10 are uncharacterized as of yet: *CG3893* and *CG2183*. CG3893 shows homology to mammalian GTSF1. Even though no direct link to the piRNA pathway has been shown, this germline-specific factor also seems to be indispensible for transposon control in mice (Yoshimura et al., 2009). CG2183 is predicted to be a homolog of GASZ. This protein was previously implicated in the piRNA

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

Table 1. Top 20 Validated Hits

FlyBase ID	Symbol	Primary Screen Average Fold Change	Validation Screen Fold Change			Fertility	Comments
			Gypsy	*Zam*	*Gypsy3*		
FBgn0028411	*nxt1*	2	2452	3566	41	–	Involved in mRNA export from nucleus
FBgn0000928	*fs(1)Yb*	11	96	700	335	+	piRNA pathway component
FBgn0041164	*armi*	48	197	846	112	+	piRNA pathway component
FBgn0261266	*zuc*	19	809	549	9	+	piRNA pathway component
FBgn0263143	*vret*	4	74	315	22	+	piRNA pathway component
FBgn0036826	*CG3893* (*asterix*)	42	80	207	10	+	Contains two CHHC zinc fingers
FBgn0016034	*mael*	3	159	452	16	+	piRNA pathway component
FBgn0033273	*CG2183*	4	173	158	11	+	Fly homolog of GASZ
FBgn0029800	*lin-52*	5	153	153	8	+	dREAM complex subunit
FBgn0038016	*MBD-R2*	16	85	48	4	+	NSL complex subunit
FBgn0029113	*uba2*	3	84	12	8	+	Sumoylation E1 ligase
FBgn0034617	*CG9754*	2	26	61	14	+	No conserved domains
FBgn0027499	*wde*	3	40	120	12	+	Cofactor of Eggless
FBgn0021761	*nup154*	6	30	186	3	–	Structural constituent of nuclear pore
FBgn0003612	*Su(var)2-10*	2	9	20	4	–	dPIAS, putative SUMO E3 ligase
FBgn0001624	*dlg1*	2	16	2	1	+	Guanylate kinase
FBgn0003401	*shu*	7	14	416	1	+	piRNA pathway component
FBgn0038739	*CG4686*	4	13	1	1	+	Part of ribokinase/pfkB and DUF423 superfamilies
FBgn0038609	*nup43*	3	12	3	1	+	WD40-repeat-containing domain
FBgn0038925	*cchl*	0	12	1	2	–	Cytochrome *c* heme lyase

pathway in mice (Ma et al., 2009) and has now been validated as a piRNA pathway component in flies (Czech et al., 2013).

smt3 (SUMO), which was one of the highest scoring genes in the primary screen, could not be validated in vivo because of developmental defects that occured upon knockdown (Talamillo et al., 2008). However, the depletion of its E1 activating enzymes AOS1 and UBA2 caused consistent transposon derepression in the validation screen (Table 1 and Figure 2E). Knockdown of the E2 conjugating enzyme *lesswright* also caused developmental defects and could not be validated.

Another notable validated hit is *windei* (*wde*), which was previously reported as a cofactor of EGG in H3K9 trimethylation (Koch et al., 2009). While EGG depletion had no effect on *gypsy* expression in our assay, knockdown of *wde* caused derepression of *gypsy*, although to a lower extend than *ZAM* (Table S3).

Additionally, two genes involved in transcriptional regulation were identified. MBD-R2 is part of the nonspecific lethal complex (Figure 2E) (Raja et al., 2010). All members of this complex except for *rcd1* scored in the primary screen. *lin-52*, which is part of the dREAM transcriptional regulator complex, also scored highly with *gypsy* and *ZAM* (Lewis et al., 2004).

Characterization of Newly Described piRNA Pathway Components

In order to place some of the validated hits at particular steps within the piRNA pathway, we constructed RNA sequencing (RNA-seq) libraries from ovaries of biological replicates of *tj*-Gal4-driven hpRNA crosses. Mapping RNA-seq reads to transposon consensus sequences revealed the same levels of *gypsy* and *ZAM* derepression observed by qPCR (Figure 3A). When analyzed on a global scale, mainly transposons dominant within the somatic compartment of the ovary reacted significantly to the respective knockdown of the target gene in a *tj*-Gal4-dependent manner (false discovery rate [FDR] < 0.05) (Malone et al., 2009). Transposons like *HeT-A*, *roo*, or *Rt1b*, which were previously shown to be germline dominant, did not change expression levels (Figure 3A) (Malone et al., 2009). The patterns of derepression that we observed in knockdowns of known components of the pathway (*armi* and *mael*) remarkably resembled those observed in *CG3893* and *wde* knockdown.

We observed a similar behavior of CG3893 and known piRNA pathway components in their impacts on the expression of protein coding genes. While proteins like NXT1 or UBA2, with potentially more general functions beyond the piRNA pathway, impacted the expression of hundreds to thousands of genes, this was not the case for ARMI, MAEL, or CG3893, which show effects that are much more restricted to transposon transcripts (Figure 3B). We interpreted these results as an indication that CG3893 might act specifically within the piRNA pathway.

We further annotated validated hits based upon their impacts on piRNA levels. Interestingly, we saw a substantive drop in the abundance of piRNAs uniquely mapping to the soma-dominant *flamenco* piRNA cluster in the *nxt1*, *uba2*, and *wde* knockdowns (Figure 4A). To avoid skewing this result based on normalization to a piRNA-producing locus, which theoretically should not change in soma-specific knockdowns (i.e., *cluster 1/42AB*), we examined the rankings within each library of piRNA clusters based on the overall abundance of corresponding piRNAs. Given

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

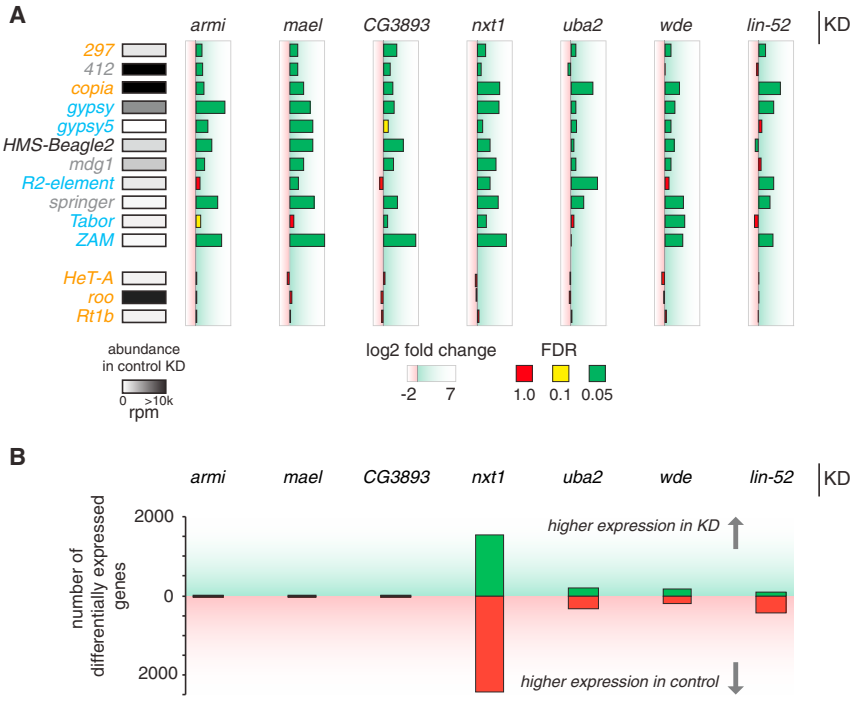


Figure 3. RNA-Seq Shows Changes in Gene and Transposon Expression upon Knockdown of Top Candidates In Vivo

(A) A subset of somatically expressed transposons is derepressed in the indicated KD. The classification of transposons according to Malone et al. (2009) is indicated in orange (germline dominant), gray (intermediate), blue (soma dominant), and black (unclassified). The absolute abundance of reads in control knockdown mapping to each transposon is shown in shades of gray. The log2 fold change of each target gene versus a negative control (*aub*) is shown. Color of the bars represents the significance of these fold changes and is indicated as an adjusted p value (FDR). Green indicates highly significant differences ($p \leq 0.05$), yellow indicates moderately significant changes ($0.05 < p \leq 0.1$), and red indicates nonsignificant changes ($0.1 < p \leq 1$) based on two biological replicates. Each knockdown is normalized to *aub* knockdown controls from their corresponding library (GD or KK). For differences in transposon abundance levels between both *aub* controls see Figure S3.

(B) The number of genes differentially expressed (adjusted $p < 0.05$) in each knockdown with respect to the control is shown. Green bars indicate the number of genes that have higher expression levels in the knockdown fly line, while red bars designate the number of genes with higher levels in the *aub* negative control.

that we only knock down each gene in the somatic compartment, only clusters within that expression domain (i.e., *flamenco*) should change their ranking. Validating this approach, we observed that in the *armi* knockdown, *flamenco* was the tenth most abundant cluster while it was the third most abundant in total RNA libraries from negative-control ovaries. Conversely, *42AB* and *X-upstream* (cluster 20A) remained at the top of the list in all tested knockdowns (Figure 4B). The two genes that mirrored *armi* are *nxt1* and *uba2*. *wde* and *lin-52* showed changes in *flamenco* piRNAs that altered its ranking, but not to the same extent. In none of the knockdowns did the length profiles of the remaining piRNAs from *flamenco* change (Figure 4C). When compared to their negative controls, piRNA levels did not change in the *mael* and *CG3893* knockdown. The same conclusions could be drawn when mapping to transposons consensus sequences: antisense populations of piRNAs with homology to soma-dominant transposons showed severe reductions in the *nxt1*, *uba2*, and *wde* knockdowns, which resembled patterns seen for the biogenesis factor *armi* (Figure 4D). Depletion of *mael*, *lin-52*, or *CG3893* did not show the same effect. Intriguingly, in the case of *lin-52* this did not coincide with the effects seen for *flamenco* mappers. None of the assayed knockdowns showed any changes in mature miRNA levels, indicating that the observed effects were specific to the piRNA pathway and not a general trend of all small RNA populations (Figure S4).

CG3893 Is Indispensible for Transposon Silencing in the Germline

piwi and *mael* have recently been shown to silence transposons in the somatic compartment of the ovary through effects on transposon transcription (transcriptional gene silencing or TGS) (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). *CG3893*, displaying patterns in global transposon derepression similar to *mael* and unaffected mature piRNA populations, appeared to be a promising candidate for a pathway component acting at the TGS effector step.

CG3893, a ~20 kDa protein, is a member of a family of proteins with unknown function (UPF0224), characterized by the presence of highly conserved zinc finger domains (Figure 5A). All five proteins of this family are weakly expressed in OSS cells and show germline-specific moderate to high expression in the ovary (Figure 5B, modENCODE tissue expression data [Graveley et al., 2011]). Out of all five members of the family, only *CG3893* showed strong effects on transposon mRNA abundance when knocked down in OSS cells (Figure 5C).

In order to obtain an additional model for a loss of function of CG3893, we searched for available transposon insertion lines. We investigated two available lines (204406, Kyoto *Drosophila* Genetic Resource Center [DGRC]; 22464, Bloomington *Drosophila* Stock Center at Indiana University). Line 204406 has a P element insertion into the first exon of CG3893, disrupting its N-terminal CHHC zinc finger domain (Figure S5A). Consistent with the insertion site, we identified truncated *CG3893* mRNAs by RNA-seq in libraries from homozygous animals. The levels of *CG3893* mRNA expression were also reduced in animals homozygous for this mutation when compared to heterozygous siblings. The results obtained by RNA-seq were confirmed by qPCR (Figure S5B). The second line (22462) has a P element insertion upstream of the first exon in either the promoter or the 5′ untranslated region

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

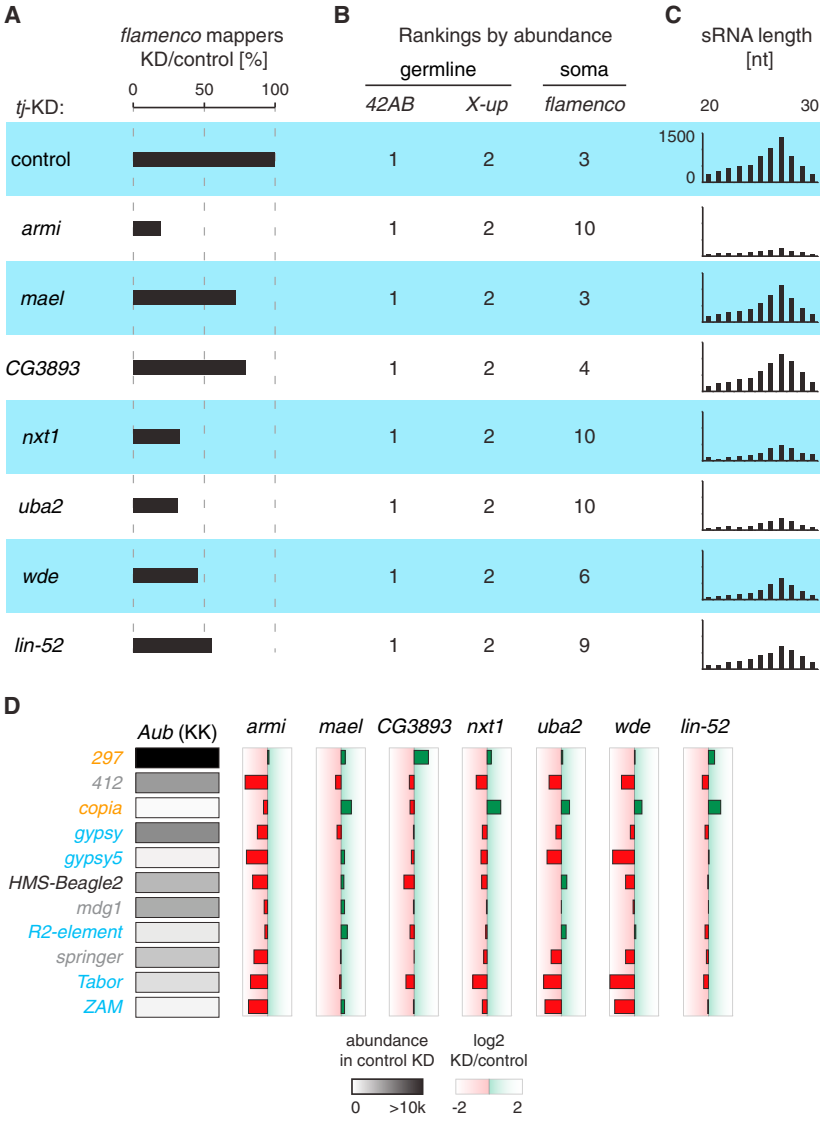


Figure 4. Biogenesis of Small RNAs from Somatic Clusters and Transposons Is Affected in Knockdowns of a Subset of Top Candidate Genes

(A) Percentages of total unique mappers (sense species, >23 nt) to *flamenco* in each knockdown (as indicated) in relation to the control knockdown are shown.

(B) The internal rankings for three representative piRNA clusters based on their representation in piRNA populations are displayed. Expression bias toward either domain (soma or germline) is indicated. Cluster definitions are in concordance with Brennecke et al. (2007).

(C) The size profiles of piRNAs mapping in sense orientation to *flamenco* in each knockdown (as in [A]) are plotted as total read count per million genomic mappers. As a control, we show that levels of microRNAs do not change in knockdowns versus negative control (Figure S4).

(D) piRNAs mapping to a subset of somatically expressed transposons are reduced when gene expression of a subset of top hits is disrupted. The classification of transposons according to Malone et al. (2009) is indicated in orange (germline dominant), gray (intermediate), blue (soma dominant), and black (unclassified). The absolute abundance of antisense piRNAs in an *aub* control mapping to each transposon is shown in shades of gray. The log2 fold change of each target gene versus a negative control is shown. See also Figure S4.

(UTR) of *CG3893*. Our RNA-seq data in control flies indicated a slightly extended *CG3893* transcript when compared to the gene model presented in FlyBase. Even though this insertion is not in any coding sequence, the observed phenotypes were severe: homozygous females were completely sterile and characterized by a complete absence of ovarian structures. This correlated with undetectable levels of *CG3893* transcript when assayed by qPCR, indicating a complete loss of function (Figure S5B). The phenotypes observed in females homozygous for the first insertion (204406) were milder, with ovaries developing to a rudimentary stage (Figure 5D). Nevertheless, this potentially hypomorphic mutation caused females to be sterile, demonstrating the negative impact of the insertion on *CG3893* function.

According to our current model of transposon silencing as a nuclear phenomenon, effectors at this step are expected to be nuclear as well. The mouse homolog of CG3893, Gtsf1, is reported to be mainly cytoplasmic in adult testes (Yoshimura et al., 2007). However, when overexpressed in OSS cells, GFP fusion proteins of CG3893 colocalized with Piwi in the nucleus (Figure 5E).

Our results to this point demonstrated the involvement of CG3893 in the somatic compartment of the ovary. In order to investigate its role in all tissues of the female germline, we generated RNA-seq and small RNA libraries from females heterozygous and homozygous for the exonic P element insertion. RNA-seq revealed a remarkable change in global transposon expression. Almost all classes of annotated transposons populating the *Drosophila* genome, except for the P element itself, showed upregulation in homozygous females when compared to their heterozygous sisters (Figure 5F). This derepression effect was equally strong for germline- and soma-dominant transposon classes. Yet, when mapping antisense piRNA reads to transposon consensus sequences, we see no change in the homozygous animals (Figure 5G).

Three recent publications demonstrate not only that piRNA-mediated transposon silencing is a nuclear phenomenon occurring through TGS, but also that it acts through deposition of silencing H3K9 trimethyl marks on active copies of transposons (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). Given its potential involvement in this

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

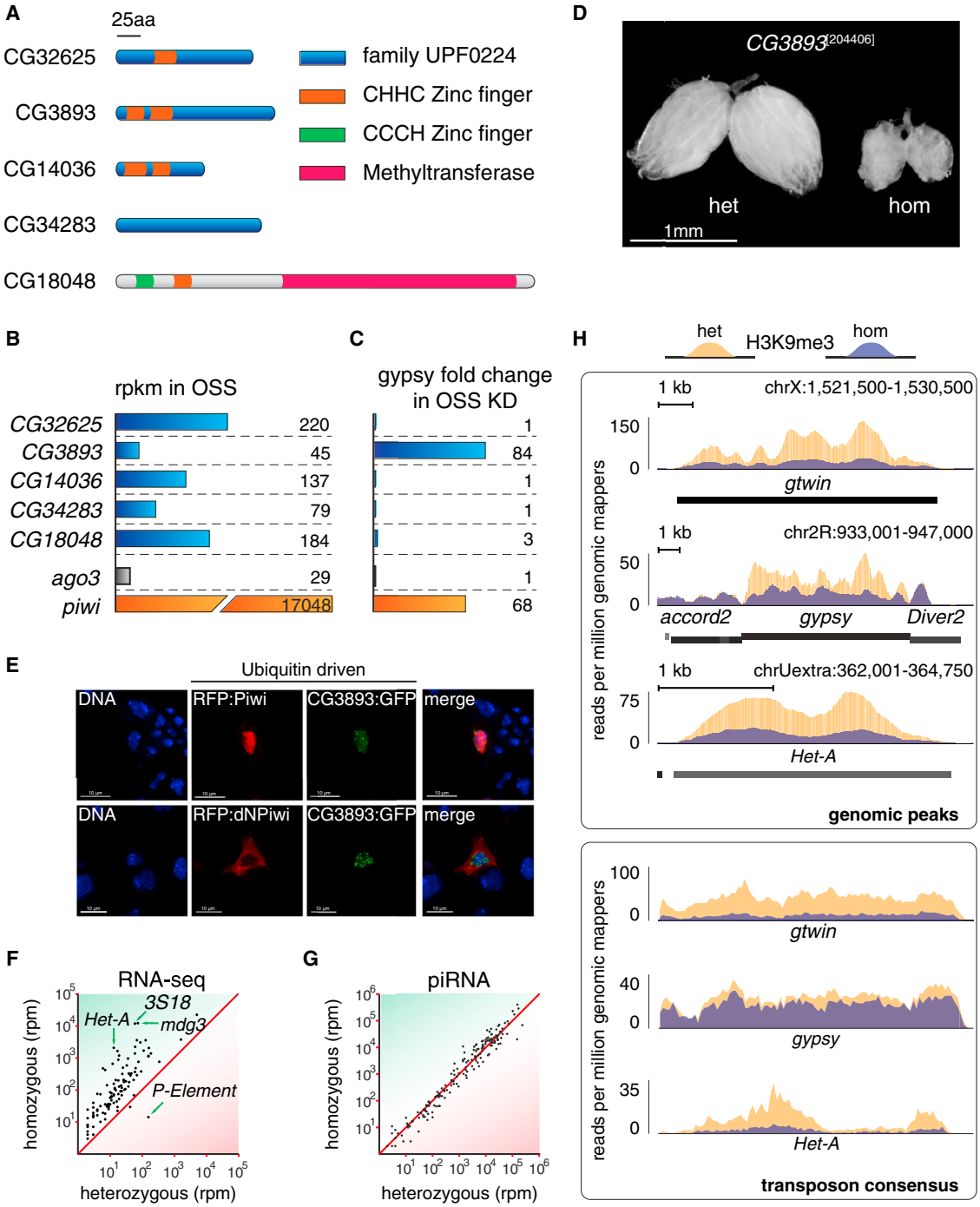


Figure 5. Disruption of CG3893 Function Has a Severe Impact on Transposon Silencing

(A) The five members of the *Drosophila* uncharacterized protein family UPF0224 and their domain structures are diagrammed. The conserved domains are highlighted as colored boxes.

(B) All five family members are weakly expressed in OSS cells. *piwi* and *ago3* expression levels are shown for comparison. Expression levels are based on the modENCODE cell line expression data and are displayed as reads per kilobase per million mapped reads (rpkm).

(C) CG3893, but no other members of its protein family, has a strong impact on transposon silencing upon knockdown in OSS cells. Effects of knockdown of *ago3* and *piwi* are shown for comparison. Numbers represent fold changes of *gypsy* levels with respect to the median fold change of the corresponding plate in the primary screen.

(D) The ovarian morphology of flies heterozygous or homozygous for a P element insertion in *CG3893* is shown (204406, Kyoto DGRC). For a more detailed view of the insertion and expression levels see Figure S5.

(legend continued on next page)

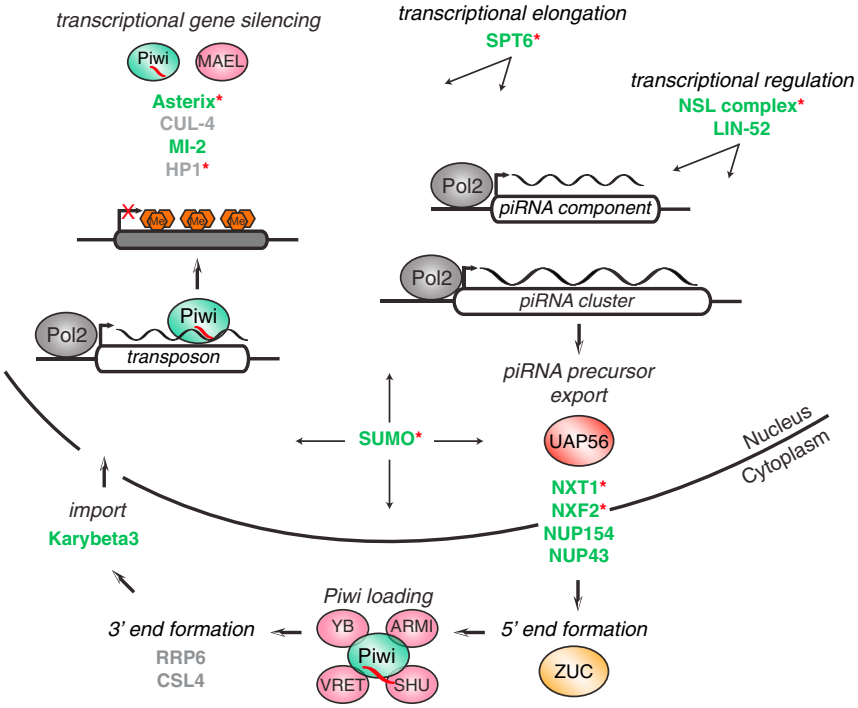


Figure 6. Potential Roles for Newly Identified piRNA Pathway Components

Known piRNA components are shown as bubbles. The newly identified genes are shown in bold text colored according to their validation status (color code as in Figure 2; Pol2, RNA polymerase 2; red hexagons represent H3K9me3). Red asterisks denote genes that validated in the germline screen done in parallel (Czech et al., 2013).

step, we investigated the effects of CG3893 disruption on this histone mark by performing chromatin immunoprecipitation sequencing (ChIP-seq) analysis for H3K9me3 in ovaries from heterozygous and homozygous females. Strikingly, homozygous females showed a marked reduction of H3K9me3 levels over a subset of peaks identified in heterozygous animals. These differential peaks overlapped with full-length insertions of both germline- and soma-dominant transposons while neighboring peaks over transposon fragments did not change (Figure 5H; fragments correspond to *accord2* and *diver*, center panel). We also observed the same changes when we mapped to consensus sequences of the corresponding transposon families, thereby aggregating signal from all genomic insertions (Figure 5H). We found that 75% of all peaks showed lower levels of H3K9me3 deposition in homozygous flies, with 28 peaks exhibiting a reduction to less than 50% of the read density compared to heterozygous siblings (Figure S5C). These peaks corresponded mostly to retrotransposons (LTR, LINE), while none of the peaks over DNA elements were affected. However, the loss of H3K9me3 in homozygous mutant animals was not exclusive to transposable elements, as a limited number of protein coding genes, such as *CG8964*, showed similar patterns (Figure S5D). Our RNA-seq data were in accordance with this finding: the *CG8964* transcript was increased ~26-fold in homozygote versus heterozygote flies. Whether this effect is due to effects on a transposon insertion within its genomic locus or within another gene that controls *CG8964* expression, or whether this represents a piRNA-independent function of CG3893, remains to be seen. Because of its small size, yet powerful role in transposon silencing, we name *CG3893 asterix* (*arx*).

DISCUSSION

We are just beginning to understand the precise molecular steps necessary for piRNA biogenesis and successful silencing of transposons. In an effort to shed light on all aspects of the pathway, from piRNA biogenesis to the effector mechanisms of transposon control, we performed an unbiased, genome-wide RNAi screen in cultured ovarian somatic cells. The primary in vitro screen proved to be a robust and specific assay for transposon derepression, with all expected piRNA components scoring strongly. To assess the validity of the primary data set, we tested our top candidate hits for effects in vivo. In order to make this resource more accessible to the scientific community, we created a web resource with all primary and validation data points (http://somatic-pirnascreen.cancan.cshl.edu/).

Within our list of 87 validated genes, we have promising candidates for filling almost every gap in our current understanding of the piRNA pathway (Figure 6). For example, the identified RNA

(E) Tagged CG3893 colocalizes with Piwi in the nucleus of OSS cells when overexpressed in transient transfections. Nuclear Hoechst staining is blue, GFP-tagged CG3893 is green, and red fluorescent protein (RFP)-tagged Piwi or ΔNT-Piwi is shown in red (Saito et al., 2009).

(F) Transposons are highly upregulated upon disruption of *CG3893* in the P element insertion line. A scatter plot of reads per million (rpm) is shown for RNA-seq of heterozygous versus homozygous flies. Each dot represents one transposon consensus sequence. Only sequences mapping in the sense orientation are taken into account.

(G) piRNA levels are not affected by *CG3893* disruption. The number of piRNA reads mapped to the same transposon consensus sequences as in (F) is expressed in reads per million.

(H) Levels of H3K9me3 decrease dramatically on a subset of transposons upon depletion of CG3893. Density plots for normalized H3K9me3 ChIP-seq reads over three transposons, *gtwin*, *gypsy*, and *Het-A*, are shown. Yellow distributions correspond to levels in heterozygous flies, and blue distributions correspond to the homozygous state. The upper box shows three distinct genomic peaks over transposon insertions; the lower box shows the corresponding consensus sequences. For all identified H3K9me3 peaks and their read densities see Figure S5C.

Please cite this article in press as: Muerdter et al., A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*, Molecular Cell (2013), http://dx.doi.org/10.1016/j.molcel.2013.04.006

export factors and nucleoporins could act in the export of primary cluster transcripts to the cytoplasm. We also identify genes that are likely to affect transcription of these piRNA precursors. WDE was shown to be a cofactor of *eggless*, a gene required for transcription of clusters (Koch et al., 2009; Rangan et al., 2011). While depletion of EGG did not result in mobilization of *gypsy* (as previously shown), *wde* knockdown led to high levels of *gypsy* expression, hinting toward a role for WDE independent of EGG. *lin-52* was previously described as a transcriptional activator of Piwi (Georlette et al., 2007). However, Li et al. (2013) recently showed that A-MYB, which provides activity orthologous to the dREAM complex in mice, controls the expression of key pathway components as well as piRNA precursor transcripts.

Following nuclear export, piRNA precursors have to be further processed by an endonuclease to create the 5′ end of the mature piRNA. ZUC, which was recently shown to be a cytoplasmic, single-stranded RNA-specific endonuclease, is the most likely candidate for this function (Ipsaro et al., 2012; Nishimasu et al., 2012). The fact that we do not identify any other annotated endonuclease with comparable derepression phenotypes in our screen supports the role for ZUC in this step. Ribonuclease (RNase) P and RNase Z, both endonucleases implicated in transfer RNA (tRNA) processing, did score in the primary screen but could not be validated in vivo because of their severe developmental defects (Dubrovsky et al., 2004; Frank and Pace, 1998). After 5′ end formation and loading into Piwi, each piRNA is proposed to be further trimmed to its mature length. The only genes exhibiting exonuclease activity and scoring highly in our screen were *csl4* and *rrp6*, both components of the exosome (Andrulis et al., 2002). However, neither of the two genes could be validated in vivo due to arrested gonadal development in knockdowns.

Transcriptional silencing of transposons through Piwi is a nuclear process, and previous data have demonstrated that unloaded Piwi remains in the cytoplasm (Saito et al., 2010). One protein possibly involved in reimportation of loaded Piwi is Karybeta3, a homolog of Importin 5 (Mosammaparast and Pemberton, 2004), which emerged as a hit from our screen.

Upon reentry into the nucleus, Piwi is able to recognize transcription of active transposons through its bound piRNA and consequently silence them (Le Thomas et al., 2013; Rozhkov et al., 2013; Sienski et al., 2012). So far, only Piwi itself and MAEL have been implicated in this step. With Asterix, we present a component of the nuclear piRNA silencing machinery that is indispensible for transcriptional repression. However, even though we see lower levels of H3K9me3 in mutant animals compared to heterozygous siblings, Asterix most likely is not directly responsible for depositing these marks. The only conserved domains within the protein are predicted to be RNA binding (Andreeva and Tidow, 2008). This still leaves a need for identifying chromatin remodelers and histone methyltransferases (HMTs) that act as piRNA effectors in TGS. No annotated HMTs were hit in the screen; specifically, disruption of Su(var)3-9, which was recently implicated in epigenetic programming directed by piRNAs, did not lead to any significant transposon derepression, suggesting a possible redundancy in proteins that act in histone methylation (Huang et al., 2013).

Elongation factors also have been shown to play a role in chromatin modification. The elongation factor SPT6, which interacts with the nuclear exosome, emerged as another validated hit of the screen (Andrulis et al., 2002). Indeed, data from fission yeast implicated Spt6 in the silencing of heterochromatic repeats (Kiely et al., 2011). In the *spt6* mutants, decreased recruitment of the CLRC complex led to loss of H3K9me3 and, as a consequence, lower occupancy of the HP1 homolog Swi6. Cul-4, a homolog of a CLRC member in yeast that is involved in histone methylation in *Drosophila*, was a primary hit in the screen but led to developmental defects during validation (Higa et al., 2006; Hong et al., 2005).

Another validated hit involved in chromatin remodeling was *mi-2*, yet its knockdown only led to modest effects on transposon derepression (Brehm et al., 2000). Heterochromatin protein 1 (HP1), which interacts with Piwi, could not be validated due to developmental defects (Brower-Toland et al., 2007). Intriguingly, MI-2 and HP1 are both sumoylation substrates, implying a function for SUMO beyond the one in piRNA biogenesis demonstrated here (Nie et al., 2009).

In summary, our unbiased, genome-wide approach was successful in identifying likely candidates to fill in many of the gaps in our understanding of the molecular mechanisms of transposon control in *Drosophila*. Together with the findings of a transcriptome-wide screen for germline piRNA pathway components done in parallel (Czech et al., 2013), which showed substantial overlap with the top hits of our screen, we are confident to have identified a comprehensive set of pathway components, which to our knowledge have not been implicated in the piRNA pathway. Our data support the current view that the piRNA pathway is the major pathway exerting transposon control, given that both the primary and validation screens were dominated by known components of the piRNA machinery. Our meta-analysis on the primary data set as a whole, as well as the list of validated genes, will provide a resource for the field in efforts toward a greater depth of understanding of piRNA production and the mechanisms by which piRNAs silence transposons.

EXPERIMENTAL PROCEDURES

Cell Culture

OSS cells were cultured as previously described and transfected using Xfect Transfection Reagent according to manufacturer's guidelines (Clontech 631317) (Niki et al., 2006).

DNA Plasmids

Expression vectors of CG3893:GFP, RFP:Piwi, and RFP:ΔNTPiwi were made using the *Drosophila* Gateway Collection.

Imaging of Fluorescent Fusion Proteins in OSS

OSS cells were cotransfected with plasmids expressing the indicated fusion proteins using Cell Line Nucleofector Kit V (Amaxa Biosystems; program T-029). Fixed cells were stained with Hoechst 33342 (Invitrogen, R37601).

RNAi Libraries

Two *Drosophila* dsRNA libraries were used in this study: the Open Biosystems *Drosophila* RNAi Collection and the *Drosophila* RNAi Screening Center Genome-wide RNAi library (DRSC 2.0).

RNAi Screening

A detailed description of the primary screen can be found in the Supplemental Experimental Procedures. A basic workflow is shown in Figure 1A. All primers used are listed in Table S4.

Drosophila Stocks and Husbandry

Fly stocks are listed in the Supplemental Experimental Procedures. A description of husbandry and validation screen procedures can be found in the Supplemental Experimental Procedures.

RNA Isolation and qPCR Assays

Total RNA from ten ovaries was extracted with Trizol and purified by organic extraction followed by isopropanol precipitation. After DNase treatment, complementary DNA (cDNA) was synthesized from 800 ng RNA using oligo dT primers and SuperScript III Reverse Transcriptase (Life Technologies). qPCR was performed to assay levels of *gypsy*, *ZAM*, *gypsy3*, and rp49. Fold changes for transposons were calculated using the delta Ct method (Livak and Schmittgen, 2001). All primers used are listed in Table S4.

RNA-Seq and Analysis

For RNA-seq libraries, 2.5–5 μg of total RNA was depleted of ribosomal RNA using the Epicentre Ribo-Zero rRNA Removal Kits (Human/Mouse/Rat) following the manufacturer's directions. Libraries were prepared using the Illumina ScriptSeq v2 RNA-Seq Library Preparation Kit and were sequenced on an Illumina HiSeq platform. Details on analysis can be found in the Supplemental Experimental Procedures.

Small RNA Cloning and Analysis

For small RNA libraries, total RNA was depleted of 2S rRNA, and libraries were constructed using the Illumina TruSeq small RNA Sample Preparation Kit following the manufacturer's protocol. Details on analysis can be found in the Supplemental Experimental Procedures.

ChIP-Seq

ChIP from 50 ovaries was done as described in Ram et al. (2011) and Garber et al. (2012), with some modifications. Details on the methodology and analysis can be found in the Supplemental Experimental Procedures.

Statistical Procedures

Details on enrichment analysis and statistical procedures can be found in the Supplemental Experimental Procedures.

ACCESSION NUMBERS

RNA-seq, ChIP-seq, and small RNA data have been deposited in the Gene Expression Omnibus database under accession number GSE46009.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures, four tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.molcel.2013.04.006.

ACKNOWLEDGMENTS

We are greatly indebted to Norbert Perrimon (DGRC) for giving us reagents. We are grateful to Sabrina Boettcher who managed all logistics and supplies. We would like to thank Alon Goren for invaluable help with ChIP-seq cloning. Molly Hammell aided in analyzing RNA-seq and ChIP-seq data. Leah Sabin provided helpful insight and advice. We wish to thank Stephanie Muller, Assaf Gordon, and Astrid Haase for help with robotics and library normalization. We would like to thank Ben Czech, Jon Preall, Sho Goh, and Julius Brennecke for sharing data prior to publication. Assistance with sequencing was provided by Emily Lee. We would also like to thank Jo Leonardo and all members of the Hannon lab for vital support. P.M.G. is a NIH trainee on the CSHL WSBS NIH Kirschstein-NRSA predoctoral T32 GM065094 grant, a William Randolph Hearst Scholar, and a Leslie Quick Junior Fellow. A grant from T. and V. Stanley supported J.G.'s work. Work in the Hannon laboratory is supported by a grant from the National Institutes of Health (5R01GM062534) and a kind gift from Kathryn W. Davis. G.J.H. is an investigator of the HHMI.

Received: February 6, 2013
Revised: April 4, 2013
Accepted: April 5, 2013
Published: May 9, 2013

REFERENCES

Andreeva, A., and Tidow, H. (2008). A novel CHHC Zn-finger domain found in spliceosomal proteins and tRNA modifying enzymes. Bioinformatics *24*, 2277–2280.

Andrulis, E.D., Werner, J., Nazarian, A., Erdjument-Bromage, H., Tempst, P., and Lis, J.T. (2002). The RNA processing exosome is linked to elongating RNA polymerase II in Drosophila. Nature *420*, 837–841.

Brehm, A., Längst, G., Kehle, J., Clapier, C.R., Imhof, A., Eberharter, A., Müller, J., and Becker, P.B. (2000). dMi-2 and ISWI chromatin remodelling factors have distinct nucleosome binding and mobilization properties. EMBO J. *19*, 4332–4341.

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bähler, J., Wood, V., et al. (2008). The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. *36* (Database issue), D637–D640.

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.

Brower-Toland, B., Findley, S.D., Jiang, L., Liu, L., Yin, H., Dus, M., Zhou, P., Elgin, S.C., and Lin, H. (2007). Drosophila PIWI associates with chromatin and interacts directly with HP1a. Genes Dev. *21*, 2300–2311.

Bucheton, A. (1995). The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. Trends Genet. *11*, 349–353.

Chintapalli, V.R., Wang, J., and Dow, J.A.T. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat. Genet. *39*, 715–720.

Czech, B., Preall, J.B., McGinn, J.T., and Hannon, G.J. (2013). A transcriptome-wide RNAi screen in the Drosophila ovary reveals novel factors of the germline piRNA pathway. Mol. Cell *50*. Published online May 9, 2013. http://dx.doi.org/10.1016/j.molcel.2013.04.007.

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

Dubrovsky, E.B., Dubrovskaya, V.A., Levinger, L., Schiffer, S., and Marchfelder, A. (2004). Drosophila RNase Z processes mitochondrial and nuclear pre-tRNA 3′ ends in vivo. Nucleic Acids Res. *32*, 255–262.

Frank, D.N., and Pace, N.R. (1998). Ribonuclease P: unity and diversity in a tRNA processing ribozyme. Annu. Rev. Biochem. *67*, 153–180.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z., et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Mol. Cell *47*, 810–822.

Georlette, D., Ahn, S., MacAlpine, D.M., Cheung, E., Lewis, P.W., Beall, E.L., Bell, S.P., Speed, T., Manak, J.R., and Botchan, M.R. (2007). Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells. Genes Dev. *21*, 2880–2896.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

Herold, A., Klymenko, T., and Izaurralde, E. (2001). NXF1/p15 heterodimers are essential for mRNA nuclear export in Drosophila. RNA *7*, 1768–1780.

Herold, A., Teixeira, L., and Izaurralde, E. (2003). Genome-wide analysis of nuclear mRNA export pathways in Drosophila. EMBO J. *22*, 2472–2483.

Higa, L.A., Wu, M., Ye, T., Kobayashi, R., Sun, H., and Zhang, H. (2006). CUL4-DDB1 ubiquitin ligase interacts with multiple WD40-repeat proteins and regulates histone methylation. Nat. Cell Biol. *8*, 1277–1283.

Hong, E.J., Villén, J., Gerace, E.L., Gygi, S.P., and Moazed, D. (2005). A cullin E3 ubiquitin ligase complex associates with Rik1 and the Clr4 histone H3-K9 methyltransferase and is required for RNAi-mediated heterochromatin formation. RNA Biol. *2*, 106–111.

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.

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.

Ishizu, H., Siomi, H., and Siomi, M.C. (2012). Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines. Genes Dev. *26*, 2361–2373.

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

Khurana, J.S., and Theurkauf, W. (2010). piRNAs, transposon silencing, and Drosophila germline development. J. Cell Biol. *191*, 905–913.

Kiely, C.M., Marguerat, S., Garcia, J.F., Madhani, H.D., Bähler, J., and Winston, F. (2011). Spt6 is required for heterochromatic silencing in the fission yeast Schizosaccharomyces pombe. Mol. Cell. Biol. *31*, 4193–4204.

Klattenhoff, C., Xi, H., Li, C., Lee, S., Xu, J., Khurana, J.S., Zhang, F., Schultz, N., Koppetsch, B.S., Nowosielska, A., et al. (2009). The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. Cell *138*, 1137–1149.

Koch, C.M., Honemann-Capito, M., Egger-Adam, D., and Wodarz, A. (2009). Windei, the Drosophila homolog of mAM/MCAF1, is an essential cofactor of the H3K9 methyl transferase dSETDB1/Eggless in germ line development. PLoS Genet. *5*, e1000644.

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.

Lévesque, L., Guzik, B., Guan, T., Coyle, J., Black, B.E., Rekosh, D., Hammarskjöld, M.L., and Paschal, B.M. (2001). RNA export mediated by tap involves NXT1-dependent interactions with the nuclear pore complex. J. Biol. Chem. *276*, 44953–44962.

Lewis, P.W., Beall, E.L., Fleischer, T.C., Georlette, D., Link, A.J., and Botchan, M.R. (2004). Identification of a Drosophila Myb-E2F2/RBF transcriptional repressor complex. Genes Dev. *18*, 2929–2940.

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.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402–408.

Ma, L., Buchold, G.M., Greenbaum, M.P., Roy, A., Burns, K.H., Zhu, H., Han, D.Y., Harris, R.A., Coarfa, C., Gunaratne, P.H., et al. (2009). GASZ is essential for male meiosis and suppression of retrotransposon expression in the male germline. PLoS Genet. *5*, e1000635.

Makarov, E.M., Makarova, O.V., Urlaub, H., Gentzel, M., Will, C.L., Wilm, M., and Lührmann, R. (2002). Small nuclear ribonucleoprotein remodeling during catalytic activation of the spliceosome. Science *298*, 2205–2208.

Malone, C.D., and Hannon, G.J. (2009). Small RNAs as guardians of the genome. Cell *136*, 656–668.

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.

McQuilton, P., St Pierre, S.E., and Thurmond, J.; FlyBase Consortium. (2012). FlyBase 101—the basics of navigating FlyBase. Nucleic Acids Res. *40* (Database issue), D706–D714.

Mosammaparast, N., and Pemberton, L.F. (2004). Karyopherins: from nuclear-transport mediators to nuclear-function regulators. Trends Cell Biol. *14*, 547–556.

Nie, M., Xie, Y., Loo, J.A., and Courey, A.J. (2009). Genetic and proteomic evidence for roles of Drosophila SUMO in cell cycle control, Ras signaling, and early pattern formation. PLoS ONE *4*, e5905.

Niki, Y., Yamaguchi, T., and Mahowald, A.P. (2006). Establishment of stable cell lines of Drosophila germ-line stem cells. Proc. Natl. Acad. Sci. USA *103*, 16325–16330.

Nishimasu, H., Ishizu, H., Saito, K., Fukuhara, S., Kamatani, M.K., Bonnefond, L., Matsumoto, N., Nishizawa, T., Nakanaga, K., Aoki, J., et al. (2012). Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature *491*, 284–287.

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.

Pane, A., Jiang, P., Zhao, D.Y., Singh, M., and Schüpbach, T. (2011). The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline. EMBO J. *30*, 4601–4615.

Pélisson, 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.

Raja, S.J., Charapitsa, I., Conrad, T., Vaquerizas, J.M., Gebhardt, P., Holz, H., Kadlec, J., Fraterman, S., Luscombe, N.M., and Akhtar, A. (2010). The nonspecific lethal complex is a transcriptional regulator in Drosophila. Mol. Cell *38*, 827–841.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M., et al. (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628–1639.

Ramadan, N., Flockhart, I., Booker, M., Perrimon, N., and Mathey-Prevot, B. (2007). Design and implementation of high-throughput RNAi screens in cultured Drosophila cells. Nat. Protoc. *2*, 2245–2264.

Rangan, P., Malone, C.D., Navarro, C., Newbold, S.P., Hayes, P.S., Sachidanandam, R., Hannon, G.J., and Lehmann, R. (2011). piRNA production requires heterochromatin formation in Drosophila. Curr. Biol. *21*, 1373–1379.

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

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., Ishizu, H., Komai, M., Kotani, H., Kawamura, Y., Nishida, K.M., Siomi, H., and Siomi, M.C. (2010). Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes Dev. *24*, 2493–2498.

Scheuermann, J.C., de Ayala Alonso, A.G., Oktaba, K., Ly-Hartig, N., McGinty, R.K., Fraterman, S., Wilm, M., Muir, T.W., and Müller, J. (2010). Histone H2A deubiquitinase activity of the Polycomb repressive complex PR-DUB. Nature *465*, 243–247.

Schüpbach, T., and Wieschaus, E. (1989). Female sterile mutations on the second chromosome of Drosophila melanogaster. I. Maternal effect mutations. Genetics *121*, 101–117.

Schüpbach, T., and Wieschaus, E. (1991). Female sterile mutations on the second chromosome of Drosophila melanogaster. II. Mutations blocking oogenesis or altering egg morphology. Genetics *129*, 1119–1136.

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.

Talamillo, A., Sánchez, J., and Barrio, R. (2008). Functional analysis of the SUMOylation pathway in Drosophila. Biochem. Soc. Trans. *36*, 868–873.

Tanentzapf, G., Devenport, D., Godt, D., and Brown, N.H. (2007). Integrin-dependent anchoring of a stem-cell niche. Nat. Cell Biol. *9*, 1413–1418.

Yoshimura, T., Miyazaki, T., Toyoda, S., Miyazaki, S., Tashiro, F., Yamato, E., and Miyazaki, J.-i. (2007). Gene expression pattern of Cue110: a member of the uncharacterized UPF0224 gene family preferentially expressed in germ cells. Gene Expr. Patterns *8*, 27–35.

Yoshimura, T., Toyoda, S., Kuramochi-Miyagawa, S., Miyazaki, T., Miyazaki, S., Tashiro, F., Yamato, E., Nakano, T., and Miyazaki, J.-i. (2009). Gtsf1/Cue110, a gene encoding a protein with two copies of a CHHC Zn-finger motif, is involved in spermatogenesis and retrotransposon suppression in murine testes. Dev. Biol. *335*, 216–227.

Zhang, F., Wang, J., Xu, J., Zhang, Z., Koppetsch, B.S., Schultz, N., Vreven, T., Meignin, C., Davis, I., Zamore, P.D., et al. (2012). UAP56 couples piRNA clusters to the perinuclear transposon silencing machinery. Cell *151*, 871–884.

Molecular Cell, Volume *50*

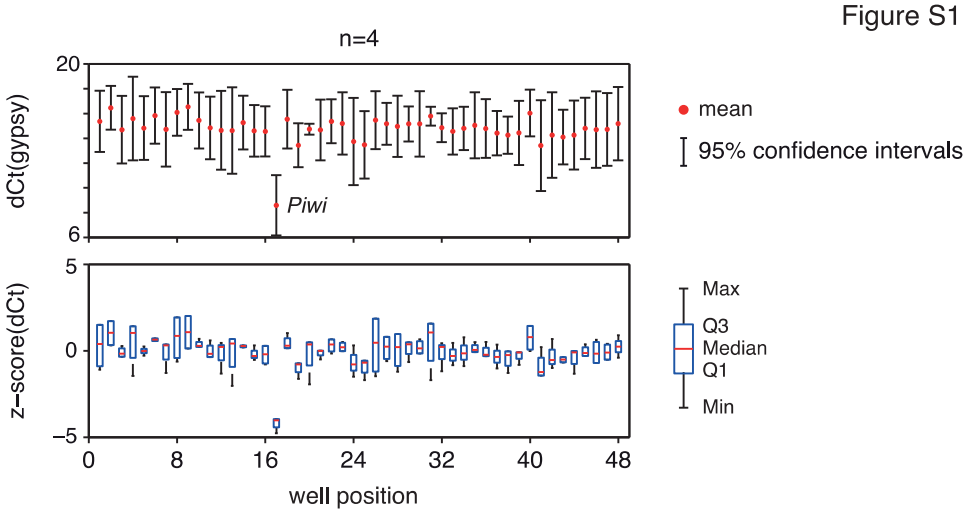
Supplemental Information

A Genome-wide RNAi Screen Draws a Genetic Framework for Transposon Control and Primary piRNA Biogenesis in *Drosophila*

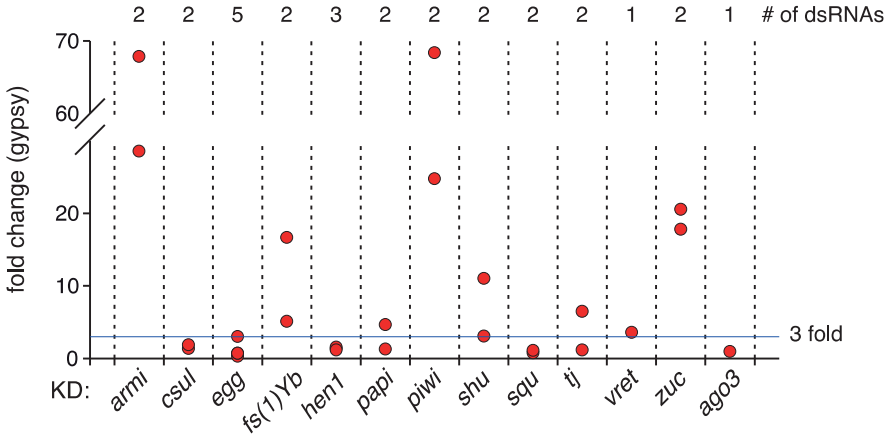
Felix Muerdter, Paloma M. Guzzardo, Jesse Gillis, Yicheng Luo, Yang Yu, Caifu Chen, Richard Fekete, and Gregory J. Hannon

Figure S1

A



B



C

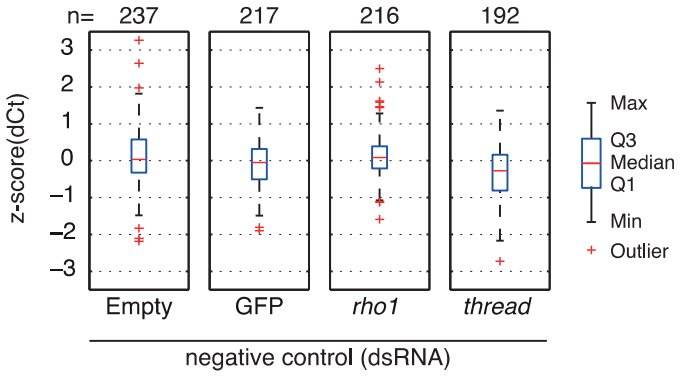


Figure S1. Performance and Controls for the Primary Screen, Related to Figure 1

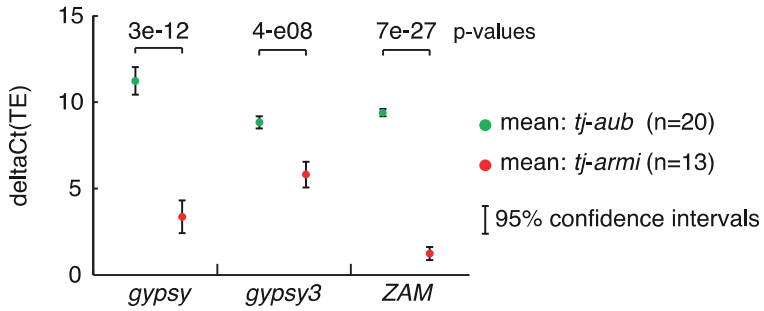
A) For 48 genes, including the positive control *piwi*, dsRNAs were transfected in four independent biological replicates. The upper graph shows the means and 95% confidence intervals of *gypsy* levels relative to a reference gene. The lower graph shows the individual z-scores as box plots for all 48 wells after normalization to the median of the plate. Piwi is a clear outlier in all four independent experiments.

B) The primary screen results in significant fold changes for all known somatic piRNA pathway components. Ago3 is shown as a negative control. The number of independent dsRNAs against each gene is indicated on top of the graph. The threshold for primary hit selection (3 fold up-regulation of *gypsy*) is marked in blue.

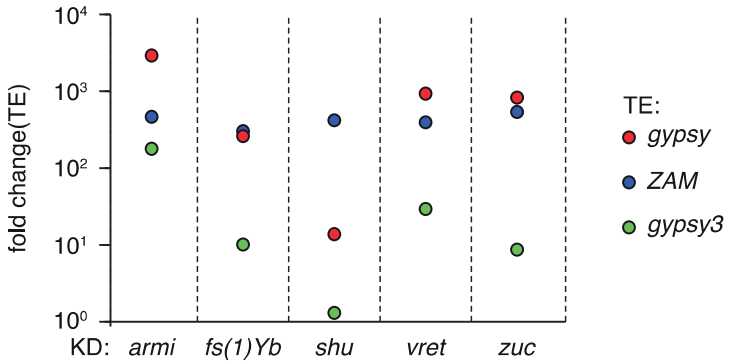
C) z-scores for 862 negative controls in the primary screen are shown as boxplots. Outliers are indicated as red crosses. The number of independent transfections of each dsRNA is indicated above.

Figure S2

A



B



C

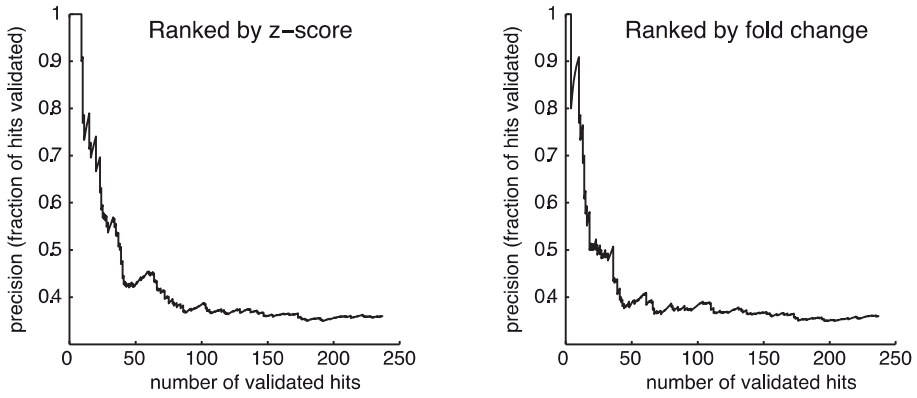


Figure S2. Performance and Controls for the Validation Screen, Related to Figure 2

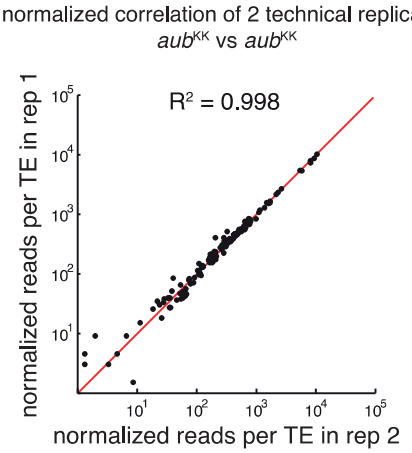
A) Knockdown of *armi* leads to highly significant differences in transposon expression when compared to a negative control (Aub). Shown are mean delta Ct values and 95% confidence intervals for three transposons assayed by qPCR. The number of biological replicates is indicates in brackets. The results of a t-test for significance are indicated as p-values for each transposon.

B) Knockdown of each component of the somatic piRNA pathway, which scored in the primary screen, has strong effects on the expression levels of three transposons *in vivo*. The fold change for each transposon upon knockdown is displayed on a log scale.

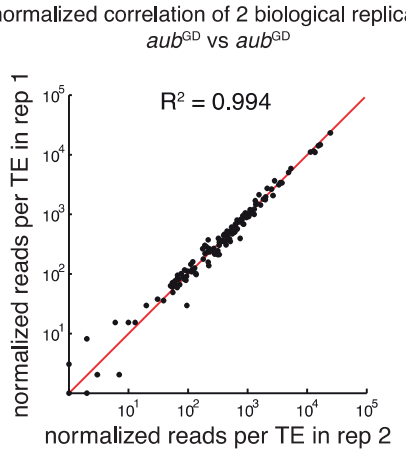
C) Z-scores and fold changes are a function of precision. Precision is the fraction of validated hits out of the total number of hits (validated and non-validated). The number of validated hits is shown on the x-axis. All dsRNAs for validated genes were used to cover the depicted range. Thus, if genes had dsRNAs producing z-scores or fold changes outside the range needed for primary hit selection, the genes' final annotation as validated or non-validated was assigned to those dsRNAs.

Figure S3

A



B



C

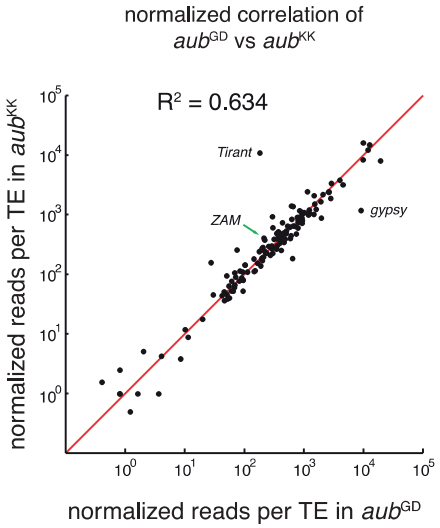


Figure S3. Two Negative Control Lines from Two Available VDRC Fly Libraries Show Different Transposon Expression Levels for Gypsy and Tirant, Related to Figure 3

A) Scatter plots of normalized reads mapping to TE consensus from RNA-seq data are shown. The squared correlation coefficient for two technical replicates of the KK line is indicated. The red line indicates where data points would show equal numbers in both samples.

B) The results for two biological replicates of *aub* flies from the GD library are shown.

C) Two Aub hpRNA lines from the KK and the GD library are compared. Data points for three transposons are highlighted: *gypsy* and *Tirant* as significantly differentially expressed and *ZAM* as the transposon used for hit calling in the validation screen.

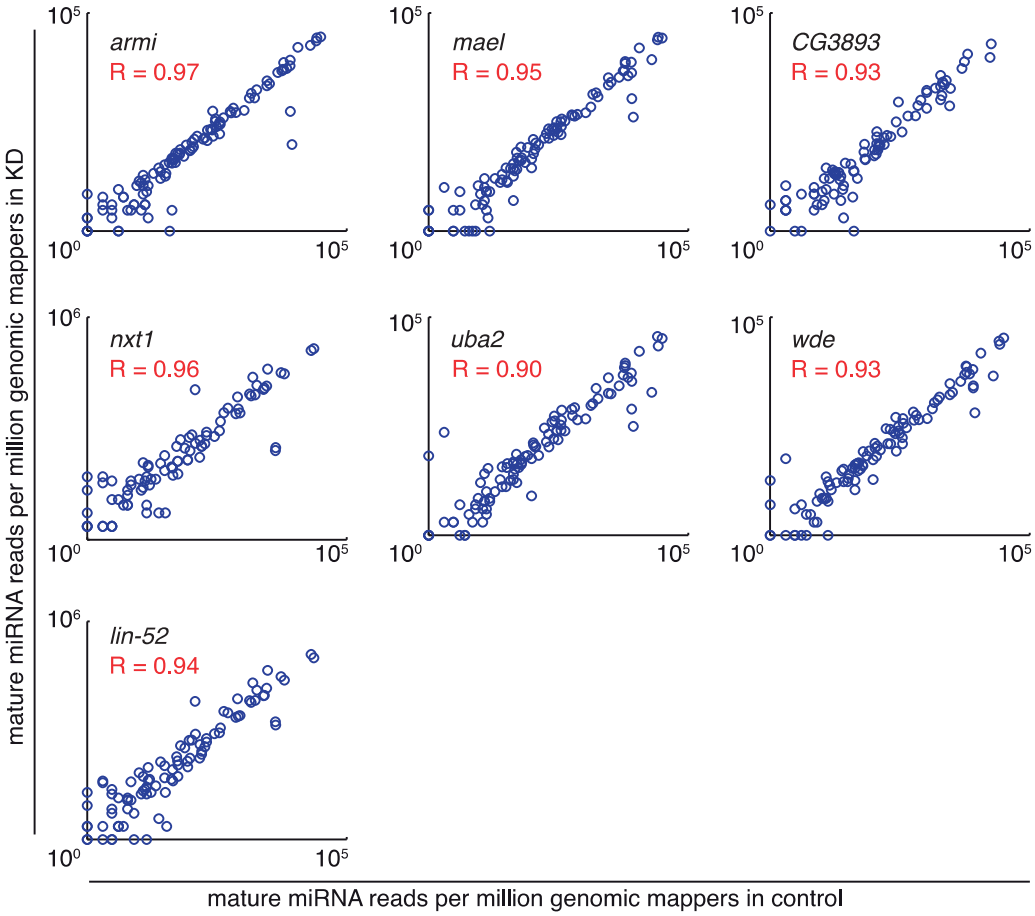


Figure S4. miRNA Populations in Knockdowns, Related to Figure 4

There are no significant changes in microRNA levels upon knock down of a number of validated hits. Scatter plots for mature microRNA in reads per million genomic mappers are shown for all follow-up genes compared to the Aub negative control. Pearson correlation coefficients are shown for each population.

Figure S5

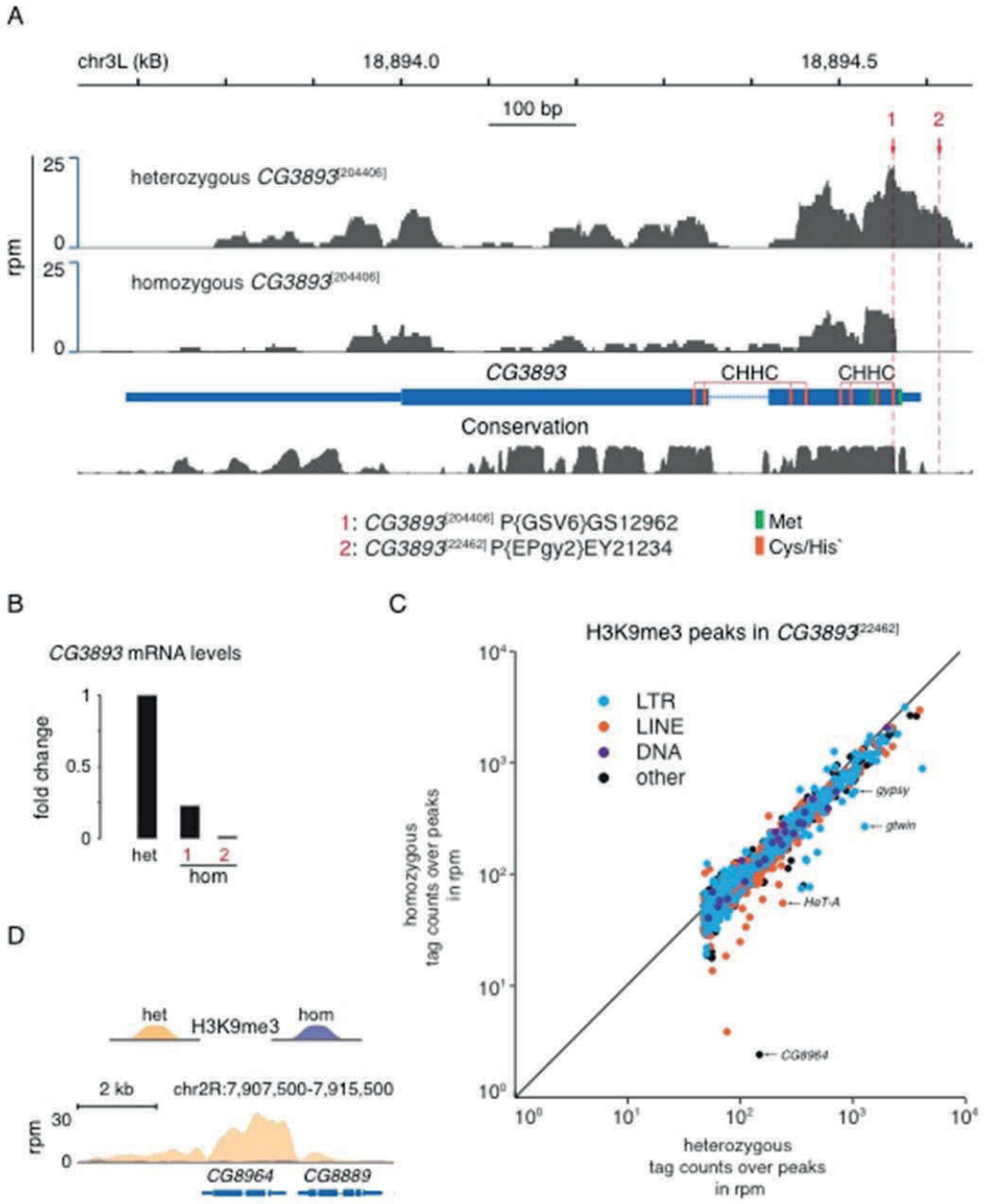

A
chr3L (kB)
18,894.0
18,894.5
100 bp
1
2
rpm
25
0
heterozygous CG3893[204406]
homozygous CG3893[204406]
CG3893
CHHC
CHHC
Conservation
1: CG3893[204406] P{GSV6}GS12962
2: CG3893[22462] P{EPgy2}EY21234
Met
Cys/His
B
CG3893 mRNA levels
fold change
1
0.5
0
het
1
2
hom
C
H3K9me3 peaks in CG3893[22462]
LTR
LINE
DNA
other
gypsy
gtwin
HeT-A
CG8964
homozygous tag counts over peaks in rpm
heterozygous tag counts over peaks in rpm
D
het
H3K9me3
hom
2 kb
chr2R:7,907,500-7,915,500
rpm
30
0
CG8964
CG8889

Figure S5. Two P Element Insertions Disrupt CG3893 Function, Related to Figure 5

A) Density plots of reads mapping to *CG3893* from RNA-seq libraries corresponding to the heterozygous and homozygous *CG3893*$^{[204406]}$ insertion line are shown. 1 and 2 shown in red designate the insertion points of P-elements in *CG3893*$^{[204406]}$ and *CG3893*$^{[22462]}$, respectively. Beneath, the FlyBase gene model for *CG3893* is shown with green boxes designating Met translation start sites and in red boxes are positions of Cys and His amino acids that make up the CHHC zinc fingers. Under the gene model, conservation is shown.

B) Both *CG3893* P-element insertion lines disrupt expression of its mRNA transcript but to different extents. qPCR for levels of *CG3893* transcript in heterozygous and homozygous flies are shown. Each homozygous fly is normalized to its corresponding heterozygous sibling. 1 corresponds to *CG3893*$^{[204406]}$ and 2 to *CG3893*$^{[22462]}$.

C) Read densities over a subset of H3K9me3 peaks identified in *CG3893*$^{[204406]}$ heterozygous flies are lower in homozygous siblings. The read densities are expressed as reads per million genomic mappers. All peaks are annotated and divided into four annotation categories: transposable elements (LTR, LINE and DNA elements) and 'other' (intra- and intergenic annotations). The three highlighted transposon peaks correspond to the plots shown in Figure 5H. A detailed view of the genic peak is shown in Figure S5D).

D) H3K9me3 peaks over genic regions are affected in *CG3893*$^{[204406]}$ homozygous flies. Read densities are expressed as reads per million genomic mappers (rpm).

Table S2. Top Enriched GO Terms in the Primary Screen, Related to Figure 1

GO term	p-value[1]	GO ID
Yb body	0	GO:0070725
negative regulation of growth of symbiont in host	0	GO:0044130
stem cell development	0	GO:0048864
negative regulation of multi-organism process	0	GO:0043901
positive regulation of Ras protein signal transduction	0	GO:0046579
dorsal appendage formation	0	GO:0046843
germ-line stem cell maintenance	8.E-10	GO:0030718
regulation of mRNA 3' end processing	6.E-09	GO:0031440
male germ-line stem cell division	6.E-09	GO:0048133
gene silencing by RNA	2.E-08	GO:0031047
negative regulation of transposition	2.E-07	GO:0010529
imaginal disc-derived wing expansion	3.E-07	GO:0048526

[1]after multiple test correction

Table S4. Primers and Probes Used in this Study, Related to Supplemental Experimental Procedures

Target	Sequence (5'-3')	Primer:Probe
Primary Screen		
gypsy fwd.	CCAACAATCTGAACCCACCAATCTA	
gypsy rev	AGTACCCGCCACAACCTTTAAG	
gypsy probe	CAAACAGGGTAGTTAAGTTAG	4.5:1
cib fwd.	GCCAGCATCCCAGCTTAGTAGT	
cib rev	GCTGGGGCGGCCATCTT	
cib probe	CGCTTCGCCAATCCA	1.5:1
Validation Screen		
gypsy fwd.	CAGGCGACAAACAGGGTAG	
gypsy rev	GTTCAAACACCAGCACATCC	
gypsy probe	AC ACAGGAATGTAGTTGGCATGCGA	4:1
gypsy3 fwd.	GACATACTGAAGGGCGAGAAC	
gypsy3 rev	TCAGGGTATCTAAGGGTGACG	
gypsy3 probe	CAAGGTAGAATTTTCCGAAGCGCAGC	4:1
ZAM fwd.	GGTATGGAAGATGTGGGTGTC	
ZAM rev	TCCTCTTCACCGTATCCCTAG	
ZAM probe	TCGCCGTAATACTCACCTGGACACT	4:1
rp49 fwd.	GTCGGATCGATATGCTAAGCTG	
rp49 rev	CAGATACTGTCCCTTGAAGCG	
rp49 probe	TTGTCGATACCCTTGGGCTTGCG	1:1
General qPCR primers and probes		
CG3893 fwd.	TCGTCATCCCAGTTCTCCT	
CG3893 rev	CATTTGATACCAGAGCCCCAG	
CG3893 probe	CGAAGACACCAGACACGCGAAGAT	1:1
2s-rRNA depletion antisense oligo		
2s-rRNA	AGTCTTACAACCCTCAACCA TATGTAGTCCAAGCAGCACT	

Supplemental Experimental Procedures

Cell Culture

OSS cell were cultured in Shields and Sang M3 Insect media (Sigma) supplemented with 10% FBS, 5% fly extract, 0.6mg/ml glutathione and 10mg/ml insulin as previously described (Niki et al., 2006). Cells were transfected using Xfect transfection reagent according to manufacturer's guidelines (Clontech, #631317).

DNA Plasmids

Expression vectors of CG3893:GFP, RFP:Piwi and RFP:ΔNTPiwi driven by an ubiquitin promoter were made using the Drosophila Gateway Collection (Terence Murphy, Carnegie Institute of Washington, Baltimore, MD). To construct expression clones, coding sequences of CG3893, Piwi and ΔNTPiwi (excluding the first 72 aa) were PCR-amplified from ovarian cDNA and cloned into pENTR/ D-TOPO, and then recombined with either destination vector pURW (DGRC1282), for Piwi and ΔNTPiwi or pUWG (DGRC 1284), for CG3893.

Imaging of Fluorescent Fusion Proteins in OSS

OSS cells were co-transfected with plasmids expressing the indicated fusion proteins using Cell Line Nucleofector kit V (Amaxa Biosystems; program T-029). 48 hours after transfection, cells were plated on glass coverslips. 24 hours later, cells were stained with Hoechst 33342 (NucBlue live cell stain; Invitrogen, R37601) and immediately fixed in 2% formaldehyde/PBS at room temperature for 5min. After three 10min PBS washes, coverslips were mounted in proLong antifade (Invitrogen, P7481) and examined under a fluorescent microscope (Nikon Eclipse Ti). Z-stack images were taken with 40X magnification and the final images were de-convoluted under the default manufacturer settings.

RNAi Libraries

Two Drosophila dsRNA libraries were used in this study, the Open Biosystems (now Thermo Scientific) Drosophila RNAi Collection version 1.0/2.0 and the Drosophila RNAi Screening Center Genome-wide RNAi library (DRSC 2.0).

RNAi Screening

OSS cells were plated in 48-well dishes (79,000 cells/well). The following day cells were transfected with 500ng of dsRNA, 0.3µl Xfect reagent and 9.7µl Xfect Buffer. To do this procedure in a robust way the Epmotion robot (Eppendorf) was used to prepare the transfection mixture in a 96-well plate and to pipette the mixture onto the cells. Approximately 12 hours post transfection, cells were washed with PBS and media was replaced. An additional media change was done on day 3 post-transfection to avoid drying of wells. On day 5 post-transfection cells were lysed with 150ul of Lysis Buffer (10mM KCl, 10mM Tris pH8, 1.5mM MgCl2, 0.5% NP-40, 60 units RNasin) per well and shaken for 5min at 300 rpm. For the DRSC library, instead of the Lysis Buffer, Ambion Cells-to-Ct Lysis Reagent (Life Technologies cat 4391848M) was used to lyse cells. Following the 5 minutes of shaking, 15µl of Stop Solution (Life Technologies cat 4402960) was added to stop the lysis reaction, mixed by pipetting, and left for 2 minutes at room temperature. Lysates were transferred to a 96-well PCR plate. 22.5µl of the lysate was used as input for a 50µl reverse transcription (RT) reaction and then incubated at 37ºC for 1 hour and 95ºC for 5 min. The RT master mix and enzyme used were those provided in the TaqMan Gene expression Cells-to-CT kit (4399002). Both the transfer of the lysate to 96-well plates, as well as the RT reaction set-up was done using the Epmotion. After cDNA synthesis, 2µl of the cDNA was used as input in a qPCR reaction to assay levels of *gypsy* and cib in a multiplexed reaction, using TaqMan Fast Advanced Master Mix (Life Technologies cat 4444965) on an Eppendorf MasterCycler EP realplex machine. Levels of *gypsy* subgenomic transcript and the reference gene *ciboulot* were assayed using hydrolysis probes spanning splice junctions and the resulting Ct values were expressed as delta-Ct z-scores (distance in standard deviations from the plate median) and fold change (in relation to the plate median). For further analysis, we ignored extreme outliers for the reference gene and wells in which *gypsy* could not be detected after 38 cycles of qPCR. Targets were called a primary hit if the *gypsy* delta-Ct z-score was lower than or equal to -1.9 and the *gypsy* fold change higher than or equal to 3. We called an additional 22

genes a primary hit based on b-score normalization (Ramadan et al., 2007). After calling primary hits Primers and probes are listed in Table S4.

Drosophila Stocks and Husbandry

For crosses in the validation round we used *tj*-GAL4 (DGRC stock 10455); GS12962 (DGRC stock 204406) and EY21234 (Bloomington stock 22462) are P-element insertions into the CG3893 locus. The 328 fly stocks corresponding to the candidate hits were ordered from Vienna Drosophila Resource Center (VDRC) and the Drosophila RNAi Resource center. The trans-IDs used by VDRC are listed in Supplementary Table S3. Lines from the DGRC are indicated with the prefix TRIP. For all crosses performed during the validation screen, five tj-GAL4 females and three VDRC hpRNA males were crossed and left in vials for five days, when parental flies were removed from the vial. Eight days after, ten female and three male F1 flies were put into new vials with yeast. After two days, ovaries from female flies were dissected. Eight days later, we checked the vials for the presence of larvae to test for fertility.

RNA Isolation and qPCR Assays

Ovaries from 10 F1 flies were dissected for each cross. Ovaries were washed once with cold PBS and homogenized in 1 ml of Trizol reagent. Total RNA was purified by phenol chloroform extraction followed by isopropanol precipitation according to the Trizol protocol. RNA was then subjected to DNase treatment using Ambion Turbo DNA-free kit at 37ºC for 30 minutes according to the manufacturer's protocol (Life Technologies). cDNA was synthesized with 800ng RNA as input using oligo dT primers (dT20) and Superscript III Reverse Transcriptase (Life Technologies) at 50ºC for 50 minutes, followed by 15 minutes at 70ºC. Next, qPCR was performed to assay levels of *gypsy*, *ZAM*, *gypsy3* and rp49. Using hydrolysis probes with FAM and HEX fluorescent reporters, we multiplexed the qPCR for the transposon and rp49. Primers and probes are listed in Table S4. Fold changes for transposons were calculated using the delta Ct method (Livak and Schmittgen, 2001). In the case of the GD library we compared each knockdown to an average of 5 biological replicates of White negative controls, for the KK library

we used 5 biological replicates of Aub negative controls. All primers were tested for efficiency in single and multiplexed reactions. Only primers for which efficiency was not impaired in the multiplexed reactions were used.

RNA-Seq and Analysis

For RNA-seq libraries, 2.5-5ug of total RNA was depleted of ribosomal RNA using the Epicenter Ribo-Zero rRNA Removal Kits (Human/Mouse/Rat), following the manufacturer's directions. Libraries were prepared using the Illumina Script Seq v2 RNA-Seq library preparation kit and were sequenced on an Illumina HiSeq platform for 36 cycles in a single end run. After collapsing all reads into a non-redundant list (cloning counts were preserved), they were mapped to Drosophila viral, tRNA and miscRNA (rRNA, snoRNA etc) sequences using the short read aligner Bowtie (Langmead et al., 2009). Only sequences in each library that did not map to either of these contaminants were then mapped to the Drosophila genome with up to two mismatches. Additionally, only uniquely mapping sequences were considered for further analysis. The same reads were mapped to a custom index of transposon consensus sequences with up to 2 mismatches (Kaminker et al., 2002). Reads mapping to up to 2 locations were considered for further analysis. For differential expression analysis of transposons we aggregated read counts mapping to these consensus sequences in sense orientation. For differential expression analysis of genes, we used htseq-counts (Part of the 'HTSeq' framework, version 0.5.3p3) to assess read counts per gene. In both cases we used the R package DESeq to call differential expression at a FDR cutoff of 0.05 based on two biological replicates (Anders and Huber, 2010).

Small RNA Cloning and Analysis

For small RNA libraries, 2.5 µg of total RNA was depleted of the 2S rRNA by annealing an antisense primer (Table S4, 95ºC to 25ºC in ~1h) followed by RNase H digestion at 37ºC for 30 minutes in 5X FS buffer (from Superscript III Reverse Transcriptase Kit, Life Technologies; RNase H was from NEB, M0297S). The remaining RNA was used as input. Libraries were

constructed using the Illumina TruSeq small RNA sample Prep kit following the manufacturer's protocol. For analysis of sRNA populations of CG3893 heterozygous and mutant animals, we used 50 ng of size selected RNA (19-28nt) as input. After sequencing on a Illumina HiSeq single-end 36 run, the TruSeq adapter (TGGAATTCTCGGGTGCCAAGGAACTCCAGTCAC) was clipped from the 3' end of the read and sequences shorter than 15 nt were discarded from further analysis. The remaining sequences were collapsed into a non-redundant list and mapped to Drosophila viral, tRNA and miscRNA (rRNA, snoRNA etc) sequences using the short read aligner Bowtie (Langmead et al., 2009). Only non-mapping reads were consequently mapped to the D. melanogaster genome (D. melanogaster Apr. 2006 [BDGP R5/dm3]). Up to two mismatches were allowed. Read counts of uniquely mapping reads were normalized to reads per million genomic mappers and compared to a negative control: in the case of knockdowns using long hpRNAs from the VDRC KK libraries we used Aub (106999^{KK}), in the case of GD libraries we used White (30033^{GD}). The same reads were mapped to a custom index of transposon consensus sequences with up to 2 mismatches (Kaminker et al., 2002). Reads mapping uniquely were considered for further analysis. The percentages of flamenco mappers displayed in Figure 4A are based on read counts normalized to 42AB. The rankings displayed in Figure 4B are calculated based on aggregated read counts of unique mappers to piRNA clusters defined in Brennecke et al. (2007). For size profiles, we used the same negative control libraries for comparison, which were normalized to the same scale in order to accurately compare across knockdowns. For analysis of transposons we aggregated read counts mapping to consensus sequences in sense orientation and normalized to the counts of three germline dominant transposons (*roo*, *Rt1b* and *Het-A*).

ChIP-Seq

ChIP was done as described in Ram et al. (2011) and Garber et al. (2012), with some modifications. Approximately fifty ovaries were dissected from heterozygous or homozygous flies into cold PBS and washed once with PBS. Ovaries were then fixed in 1.8% formaldehyde

for 10 minutes, then quenched by adding glycine to 0.125M and immediately placed on ice. Tissue was then homogenized by douncing five times with pestle A (Kontes). Washed once with PBS supplemented with protease inhibitors (Roche) and pellet was flash frozen in liquid nitrogen. Pellets were then thawed on ice and resuspended in 1mL Lysis Buffer (1% SDS, 10mM EDTA, 50mM Tris-HCl, pH 8.1) and lysed for 10 minutes in ice. Chromatin was sheared to 200-800bp using a Branson sonifier (model S-450D). After clearing lysate by centrifugation, 9mLs of Dilution Buffer (0.01% SDS, 1.1%Triton X-100, 1.2mM EDTA, 16.7mM Tris-HCl, pH 8.1, 167mM NaCl) were added to the lysate and 5mLs of the lysate were incubated with a 50ul of an equal mixture of conjuagted protein A and G Dynabeads (Invitrogen). To conjugate beads, they first had been washed once in Blocking Buffer (1X PBS, 0.5% TWEEN 20, 0.5% BSA), then coupled for 1 hour at 4ºC with 5ug of H3K9me3 antibody (Abcam 8898) and finally washed twice with Blocking Buffer to remove excess antibody. Lysate and conjugated magnetic beads were rotated at 4ºC overnight. Beads were then resuspended in 200ul cold RIPA buffer (10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 14 mM NaCl, 1% Triton X-100, 0.1% SDS, 0.1% DOC) and transfer to a 96-well plate. All further separation steps were performed in the 96-well plate magnet. Beads were washed five times with 200ul cold RIPA, two times with RIPA buffer supplemented with 500 mM NaCl, two times with LiCl buffer (10 mM TE, 250mM LiCl, 0.5% NP-40, 0.5% DOC), and once with TE (10Mm Tris-HCl pH 8.0, 1mM EDTA). Samples were eluted in 50 µl of 0.5% SDS, 300 mM NaCl, 5 mM EDTA, 10 mM Tris HCl pH 8.0. The eluate was reverse cross-linked at 65ºC for 4 hours and then treated with 2ul of RNaseA (Roche, 11119915001) for 30 min followed by 2.5 µl of Proteinase K (NEB, P8102S) for two hours. Library preparation was done as indicated in Garber et al. (2012), but without automation. In brief, to purify DNA 120ul of Ampure XP beads (Agencourt) were added to the reverse cross-linked samples, mixed by pipetting and incubated for 2 minutes. Samples were then placed on the magnetic stand for 4 minutes to separate beads, followed by 2 washes with 70% ethanol and air dried for 4 minutes and eluted in 10mM Tris-HCl pH 8.0. Library was constructed by

performing DNA end-repair, A-base addition, adaptor ligation and enrichment PCR. After each step DNA was purified by adding 20% PEG and 2.5 M NaCl to the reaction, to allow DNA to bind to Ampure XP beads already in the tube. Samples were not moved form their original well position, until after PCR enrichment. The libraries were sequenced on the Illumina HiSeq platform for 76 cycles in a pair-end run. The resulting reads were mapped with Bowtie 2 with the preset option --sensitive (Langmead and Salzberg, 2012). Only read pairs mapping concordantly were used for further analysis. For calling H3K9me3 peaks, annotation of called peaks and visualization, we used the HOMER software package (Heinz et al., 2010). Enrichments of H3K9me3 signal were calculated using input libraries as a control signal. All peaks within 8kb distance from each other were merged into regions. We used annotatePeaks from Homer to then calculate the tag counts in heterozygous and homozygous libraries over those regions, normalizing each tag to reads per million genomic mappers.

Statistical Procedures

Enrichment analysis was conducted using 2714 gene sets from the gene ontology (Ashburner et al., 2000; Barrell et al., 2009). This constituted the complete complement of gene sets in GO with between five and 100 Drosophila genes annotated to them in either the cellular component or biological process branch of GO. Molecular function substantially overlapped with biological process in many top functions and was excluded to diminish redundancy. Significance was calculated using an adaptation of the ROC-based approach described in (Gillis et al., 2010) and elsewhere. To obtain a ranking for the genes, dsRNA z-scores and fold changes were independently averaged for each gene. These scores were then converted into ranks and averaged (effectively weighting them equally). Based on the ROC_{50} approach first described in (Gribskov and Robinson, 1996), all scores outside of the top 50 were regarded as tied. Statistical enrichment of the GO functions was then calculated (Mann-Whitney U test) with multiple test correction (Benjamini, 1995).

Supplemental References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics *25*, 25-29.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome biology *11*, R106.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., and Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. Nucleic acids research *37*, D396-403.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B Met *57*, 289-300.

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.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z.*, et al.* (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. Molecular cell *47*, 810-822.

Gillis, J., Mistry, M., and Pavlidis, P. (2010). Gene function analysis in complex data sets using ErmineJ. Nature Protocols *5*, 1148-1159.

Gribskov, M., and Robinson, N.L. (1996). Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. Computers & chemistry *20*, 25-33.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription

factors prime cis-regulatory elements required for macrophage and B cell identities. Molecular cell *38*, 576-589.

Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M.*, et al.* (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome biology *3*, RESEARCH0084.

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 biology *10*, R25.

Livak, K.J., and Schmittgen, T.D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods *25*, 402-408.

Niki, Y., Yamaguchi, T., and Mahowald, A.P. (2006). Establishment of stable cell lines of Drosophila germ-line stem cells. Proceedings of the National Academy of Sciences of the United States of America *103*, 16325-16330.

Ram, O., Goren, A., Amit, I., Shoresh, N., Yosef, N., Ernst, J., Kellis, M., Gymrek, M., Issner, R., Coyne, M.*, et al.* (2011). Combinatorial patterning of chromatin regulators uncovered by genome-wide location analysis in human cells. Cell *147*, 1628-1639.

Ramadan, N., Flockhart, I., Booker, M., Perrimon, N., and Mathey-Prevot, B. (2007). Design and implementation of high-throughput RNAi screens in cultured Drosophila cells. Nature Protocols *2*, 2245-2264.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Genetics & Development

The piRNA pathway in flies: highlights and future directions

Paloma M Guzzardo, Felix Muerdter and Gregory J Hannon

Piwi proteins, together with their bound Piwi-interacting RNAs, constitute an evolutionarily conserved, germline-specific innate immune system. The piRNA pathway is one of the key mechanisms for silencing transposable elements in the germline, thereby preserving genome integrity between generations. Recent work from several groups has significantly advanced our understanding of how piRNAs arise from discrete genomic loci, termed piRNA clusters, and how these Piwi-piRNA complexes enforce transposon silencing. Here, we discuss these recent findings, as well as highlight some aspects of piRNA biology that continue to escape our understanding.

Address
Watson School of Biological Sciences, Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, United States

Corresponding author: Hannon, Gregory J (hannon@cshl.edu)

Current Opinion in Genetics & Development 2012, **23**:xx–yy

This review comes from a themed issue on **Cancer genomics**

Edited by **Nahum Sonenberg** and **Nissim Hay**

0959-437X/$ – see front matter, Published by Elsevier Ltd.

http://dx.doi.org/10.1016/j.gde.2012.12.003

The piRNA pathway

Germ cells are the only cell type of an organism that contribute genetic material to future progeny. It is therefore essential that the integrity of this genome is preserved to protect reproductive success. One threat placed on germ cells is the movement of mobile genetic elements, or transposons, which correspond to a large fraction of the eukaryotic genome. Although transposons provide some benefits in driving evolution, their uncontrolled expression can lead to loss of genome integrity [1]. One of the major ways in which transposable elements (TEs) are kept under control is via Piwi-interacting RNAs (piRNAs) [2,3]. piRNAs are a class of small RNAs bound by the Piwi clade of Argonaute (Ago) proteins. As with all members of the Ago family, Piwi clade proteins rely on sequence complementarity to identify their targets, which for piRNAs are most commonly transposable elements. The importance of this pathway is evident; Piwi proteins are highly conserved throughout evolution, and their loss of function leads to gross defects in gametogenesis and to sterility.

With many aspects of this pathway being studied in a range of organisms, it is impossible to summarize all recent insights. Therefore, we will focus specifically on the piRNA pathway in the ovary of *Drosophila melanogaster*, which has been one of the main model organisms in the study of this pathway and which has helped establish the framework for how it functions.

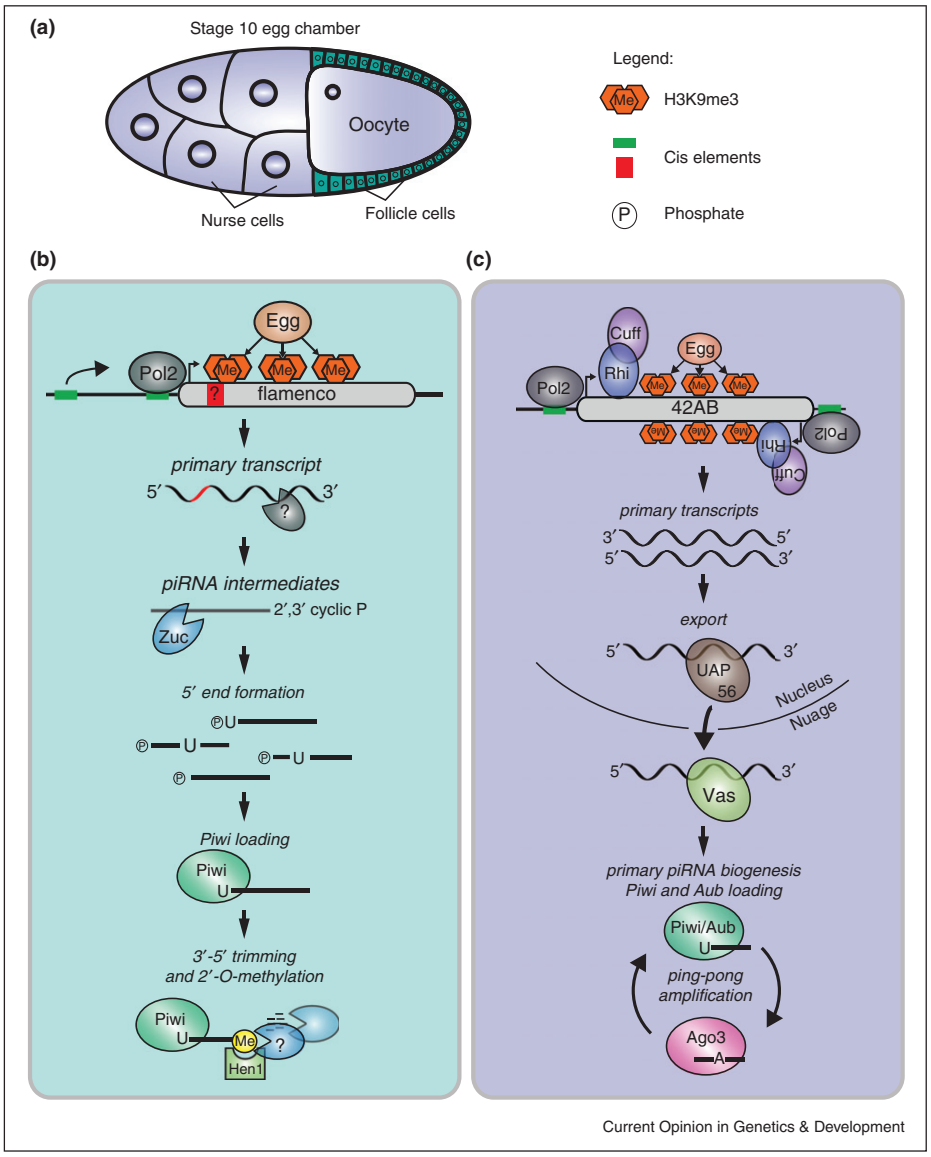
An intriguing aspect of piRNA biology in *Drosophila* ovaries is that there are two distinct iterations of the pathway active in this tissue: one in the germ cells and one in the follicle cells, cells of somatic origin that surround and support the developing germ cells [4,5] (Figure 1a). Controlling TEs in both of these cell types is important, since active transposons found within follicle cells, such as those from the gypsy family of retrovirus-like transposons, can form viral particles and infect the oocyte [6]. The somatic and germline piRNA pathways are distinct mainly because of the different expression patterns of the three fly Piwi proteins. While Aubergine (Aub) and Argonaute (Ago3) are exclusively found in the nuage of germ cells, Piwi is found in the nuclei of both germ cells and follicle cells [7–10]. Therefore, the somatic pathway acts only through piRNAs generated by primary biogenesis, while in germ cells, in addition to primary biogenesis, a more complex piRNA amplification loop exists that depends on the slicer activity of Aub and Ago3 [9,10]. Understanding the less complex primary piRNA pathway acting in somatic cells has provided a basic mechanistic framework of piRNA biogenesis that is likely shared between both somatic and germline piRNAs.

Taking advantage of the ease with which genetic manipulations can be done in *Drosophila*, studies of the small RNA populations in different piRNA mutants, together with other general molecular and cell biological analyses, such as localization studies and measurements of protein-protein interactions, have provided the main bulk of experimental data in the piRNA field [11]. The availability of cell lines derived from follicle cells (OSS/OSC) has also aided the study of the piRNA pathway [12–14]. To date, there are more than two-dozen proteins implicated in the piRNA pathway. However, many of the specific molecular steps that occur to generate a piRNA and that enable a piRNA to silence transposons remain unclear. In this review we will provide a brief summary of what is known about the piRNA pathway as well as discuss the open questions in the field.

How are piRNAs made?

The majority of piRNAs arise from specific genomic loci, known as piRNA clusters, which are found in pericentromeric heterochromatin [9]. Other sources of piRNAs

Figure 1



A model for piRNA biogenesis in the *Drosophila* ovary. **(a)** Two distinct piRNA pathways are active in a stage 10 egg chamber of the *Drosophila* ovary. The nurse cells that provide nutrients to the oocyte and the oocyte itself make up the germ cells of the ovary, shown in blue. The monolayer of somatic follicle cells surrounding the oocyte is shown in green. Nuclei are indicated as circles within each cell. **(b)** In follicle cells, primary piRNAs arise from *flamenco* and are processed through a cascade of enzymatic cuts. Transcription by RNA polymerase II (Pol2) depends on deposition of Histone 3 Lysine 9 trimethyl marks (H3K9me3) by Eggless (Egg). Regulatory *cis*-acting elements, indicated as green boxes, upstream of the transcriptional start site could affect Pol2 recruitment and transcription. Additionally, clusters could carry *cis* elements within themselves, shown in red, that affect downstream processing. After processing of the primary cluster transcript by unknown activities, piRNA intermediates are cleaved by the nuclease, Zucchini (Zuc). After 5′ end formation, transcripts with a U at the first position are preferentially loaded into Piwi. Trimming activity, which could be carried out by redundant nucleases, shortens the transcript to its mature length. This process is coupled to 2′-O-methylation by Hen1. **(c)** The transcription of clusters in germ cells can occur bidirectionally. In addition to Egg, the HP1 homolog Rhino (Rhi) and Cutoff (Cuff) are essential for transcription. Subsequently, the helicase UAP56 binds the primary transcript and escorts it to the nuclear periphery. There, it is handed over to another RNA helicase Vasa (Vas) and arrives at its site of biogenesis, the nuage. After primary processing by similar machinery as in (a), primary piRNAs are loaded into Piwi and Aub, and potentially Ago3. These primary piRNAs can be used to kick-start the ping-pong amplification cycle, which silences transposons post-transcriptionally.

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

do exist, such as the 3′UTRs of protein coding genes and dispersed euchromatic copies of TEs [9,14,15]. piRNA clusters contain remnants of transposons and serve as a catalog of sequences previously defined as targets for silencing. Exposure to a new transposon can lead to the expansion of this catalog and control of the TE, while omission from the catalog can mean that the element escapes repression [16•]. Brennecke *et al.* defined over 140 such clusters in *Drosophila* and saw that these clusters could be uni-directionally or bidirectionally transcribed [9]. Most of these clusters are active specifically in germ cells, while only a single major cluster (*flamenco*) drives transposon silencing in the soma. In general, germline clusters have two promoters, one on either side of the cluster (e.g. *cluster 42AB*), and are transcribed bidirectionally, while *flamenco* is uni-directionally transcribed.

Little is understood about what defines a piRNA cluster, how clusters are transcribed, and how this process is regulated. To date, we have no knowledge of transcription factors that regulate cluster expression. Clusters seem to be expressed in a cell-type specific manner, so there must be cell-type specific transcription factors enforcing this pattern. The promoters of clusters and their regulatory elements have not been defined, but in the case of *flamenco*, existing evidence suggests a single, discrete promoter, since a *P-element* insertion at the beginning of the cluster abolishes piRNA production, even ~200 kb away from the insertion point [9,17].

Some studies suggest a role for chromatin context in regulating cluster transcription. Deposition of Histone 3 Lysine 9 trimethyl marks (H3K9me3) was proposed to be necessary for cluster transcription, since mutations in Eggless (Egg, dSETDB1), a histone methyltransferase, lead to decreases in H3K9me3 deposition, and in the levels of cluster transcripts within both germ cells and somatic cells [18] (Figure 1b and c). As expected, these decreases in cluster transcription led to a reduction of mature piRNAs and upregulation of TEs. Rhino, a Heterochromatin Protein 1 homolog, and Cutoff (Cuff), a yeast Rai1-like nuclease, physically interact, and together bind specifically to bidirectionally transcribed clusters in the germline to promote their transcription [19,20]. Both proteins are found in nuclear foci in germ cells and depend on each other for their nuclear localization. How these factors promote cluster transcription remains unclear. Although Rhino and Cutoff are predominantly nuclear, their depletion is sufficient to disrupt Aub and Ago3 localization in nuage [19,20].

In another study addressing the role of chromatin context in cluster identity, Muerdter and colleagues found that when a cluster was taken out of its normal heterochromatic genomic context and placed in a euchromatic locus, it is still able to produce piRNAs [21]. This implies that clusters themselves contain sufficient information, possibly through *cis*-elements or secondary structure, to trigger piRNA production. However, it is also possible that information in the modified cluster is capable of recreating the chromatin context necessary for its expression, since the authors did not verify the euchromatic status of the cluster after insertion. In summary, more research is needed to understand the determinants of cluster identity; whether it be the chromatin context of the cluster, sequences in or surrounding the cluster that are important for transcription, or if it is sequences recognized within the transcript after transcription that then mark it as a piRNA producing transcript.

Following cluster transcription, the current model states that the primary transcript is exported to the cytoplasm, where it is processed into primary piRNAs that are loaded into Piwi or Aub. A recent study by Zhang *et al.* shed some light on how cluster transcripts are escorted from the transcription site to the nuage where processing is thought to occur [22]. The study shows that UAP56, a putative helicase, co-localizes with Rhino in nuclear foci. Mutation of UAP56 leads to germline transposon upregulation, decrease of piRNAs mapping to germline clusters, and disruption of Aub, Ago3, and Vasa from nuage. Based on how the Rhino-UAP56 foci are positioned next to the nuclear pore, and the finding that UAP56 and Vasa bind germline cluster transcripts, the authors proposed a model in which UAP56 escorts the primary transcript through the nuclear pore to nuage, where the transcript is handed over to Vasa and funneled into the biogenesis machinery. Since UAP56 is believed to be germ cell specific, factors that mediate export in the follicle cells remain a mystery. Whether the cluster transcript is exported as one long RNA or if some processing occurs in the nucleus to generate smaller piRNA intermediates to be exported, remains unknown.

After the cluster transcript is exported, it must be processed into piRNAs. Since Piwi-bound piRNAs have a strong preference for a uridine at the 5′ end (1 U) [9], this suggests a model of primary piRNA biogenesis wherein the 5′ end of the piRNA is generated first, followed by preferential loading of piRNA intermediates with a 5′ U into Piwi, followed by 3′ trimming. The variable lengths of primary piRNAs (23–29nt) could result from a footprint specific to the Piwi protein into which the intermediate is loaded, since the size of the RNA binding pocket probably varies slightly between each protein, and Aub, Ago3 and Piwi associated piRNAs are of slightly different lengths.

The factors responsible for 5′ and 3′ end formation have yet to be uncovered. However, recent advancements were made in our understanding of one piRNA protein that may be involved with end formation. Nishimasu *et al.* and Ipsaro *et al.* both revealed the crystal structure of the piRNA pathway protein Zucchini (Zuc) [23••,24••].

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

Based on its structure, Zuc shows a preference for binding specifically single stranded RNA. *In vitro* studies demonstrated that both the mouse and *Drosophila* Zuc protein had endoribonuclease activity [23••,24••], contradictory to previous reports implicating Zuc as a phospholipase [25,26]. The cleaved RNA product bore a 5′-monophosphate group, a characteristic of mature piRNAs. These data make Zuc the principal candidate for 5′ end formation. Both studies failed to show association of Zuc with piRNA precursors, which would have made the argument for its role as the 5′ nuclease much stronger, given that it shows no sequence preferences. Unlike most other piRNA factors, Zuc localizes to the mitochondrial membrane, and loss of this nuclease in either the germline, or the soma, results in a dramatic reduction of piRNAs [4,25–28]. The role that mitochondria could play in the piRNA pathway remains enigmatic, though its ancient connections to antiviral responses, for example it serving as the location at which the RIG-I pathway operates, is provocative [29]. In flies and mice, Piwi proteins are localized to discrete cytoplasmic structures associated with mitochondria [3], but whether this is purely to allow compartmentalization of the pathway, or whether it implies a further role of mitochondrial activity in the piRNA pathway is unclear.

The precise biochemical mechanism of piRNA 3′ end formation remains a mystery. Recent work in a cell line derived from silkworm ovaries, BmN4, has brought the field closer to identifying the 3′ generating enzyme [30••]. Kawaoka and colleagues established an *in vitro* 3′ trimming assay using BmN4 cell extracts. The authors found that Siwi (silkworm Piwi) binds transcripts with a bias toward 1 U, and that extended precursor transcripts could be trimmed in extracts, in a Mg 2+ dependent manner, to mature piRNA length. It had been determined previously that piRNAs are 2′-O-methylated at their 3′ termini by Hen1, and the addition of this modification was observed to be coupled with the trimming activity [31,32]. The importance of the 3′ terminal modification remains uncertain, because mutants of Hen1 have no detectable phenotype [31,32]. These findings are in accordance with the model that piRNA precursors bind to Piwi in the cytoplasm, and then are trimmed and methylated at the 3′ terminus. Unfortunately, the molecular nature of the trimming activity remains enigmatic; 'trimmer' could not be purified due to its insoluble nature. Moreover, no exonuclease has yet emerged as a candidate trimmer from genetic screen, which could indicate that multiple redundant trimmers exist or that trimmer has essential functions that mask an ability to isolate it as a piRNA pathway mutant.

Our current model follows the idea that Piwi must be loaded with a mature piRNA in order to be imported into the nucleus. Successful loading of Piwi-family proteins with primary piRNAs requires several other players. Although there are some distinguishing factors between the loading process in somatic and germ cells, many proteins are shared between the two pathways. The common proteins involved in biogenesis are Armitage (Armi), an RNA helicase, Shutdown (Shu), a cochaperone, and Vreteno (Vret) a TUDOR domain containing protein [27,28,33–37]. Although we understand little of the precise role of any of these proteins, mutation of any one disrupts localization of Piwi, and levels of associated piRNAs decrease dramatically [4,28,34–36]. It is important to note that mutations in Shu and Vret lead to delocalization of all three Piwi proteins in the germline, while Zuc and Armi mutants delocalize Piwi, but not Aub and Ago3. This could mean that Shu and Vret play a more general role in primary biogenesis involving Piwi and Aub, while Armi only aids Piwi in the piRNA loading process.

In the soma, Yb, a TUDOR-domain protein that also contains an RNA helicase motif, is an important additional factor for primary biogenesis. This protein localizes to foci in the cytoplasm, together with all other known loading components [27,33,38]. Zuc, the putative 5′ nuclease, localizes to mitochondria, many of which are adjacent to Yb bodies, supporting the role of these structures in Piwi RISC assembly. In Zuc mutants, Vret, Armi, Shu, and Yb all accumulate in enlarged Yb bodies with Piwi, suggesting that when the 5′ end of the piRNA cannot be generated, the loading machinery accumulates in the foci in response to a stall in biogenesis [27,28,33,35]. In the germline, there are no Yb bodies, and Yb is not expressed. Current evidence suggests that two Yb-related proteins, Brother of Yb and Sister of Yb, might serve the role played by Yb in the cytoplasm [28].

In germ cells, the loading process seems to occur in the nuage, where Aub and Ago3 localize. The function of the nuage is unknown, but many piRNA factors are found there, suggesting an important role in the piRNA pathway. One important difference between germ cells and the soma is that in germ cells, Aub and Ago3 engage in an adaptive, slicer-dependent loop termed the ping-pong cycle, which specifically amplifies the piRNA response against active elements [9,10]. In this model, Aub, bound to cluster-derived piRNAs, recognizes an active transposon transcript and cleaves it, generating the 5′ end of a new sense piRNA, which associates with Ago3. Subsequently, sense strand piRNA-loaded Ago3 can recognize complementary sequences in cluster transcripts, and through its slicer activity can generate a new antisense Aub bound piRNA, completing the cycle. According to the ping-pong model of piRNA amplification, Aub and Ago3 must be catalytically active in order to cleave new piRNAs from expressed transposons or piRNA cluster transcripts. However, the phenotypes of catalytically inactive mutants have never been described. While Aub and Ago3 seem to be responsible for generating

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

the 5′ end of each piRNA amplified through ping-pong, how the 3′ end is generated remains unknown, though it may proceed through the action of the same trimmer that is used for primary biogenesis.

In order to initiate the ping-pong cycle, piRNAs loaded into Aub are required. These come from two sources. One is primary biogenesis. The second is maternally deposited Aub, as the protein is loaded into developing oocytes along with associated piRNAs [8,39,40]. The importance of maternally deposited piRNAs is evident from analyses of hybrid dysgenesis models. In these cases, maternal deposition of piRNAs, produced by ping-pong and corresponding to the *I-element* or *P-element*, correlates with initiation of ping-pong in progeny and with effective element silencing [40]. For the *I-element*, as mothers age, their progeny have a reduced probability of being sterile even in the absence of the ability of the mother to use active *I-elements* as ping-pong substrates [16•,41]. For *P-elements*, even the dysgenic progeny can regain some fertility as the animals age. This suggests that perhaps primary piRNAs corresponding to those elements accumulate with age in the mother or offspring to a level sufficient to confer resistance.

How do piRNAs silence transposons?

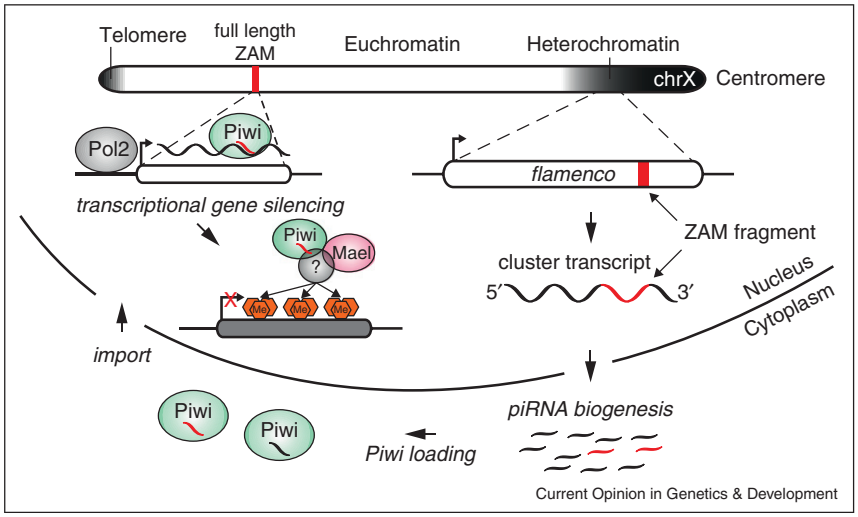
It seems evident that in germ cells Aub and Ago3 silence transposons through post-transcriptional gene silencing (PTGS). These two proteins posses slicer activity and cleave active TE transcripts during the ping-pong amplification cycle. By using the cleavage products to make more piRNAs, this cycle is able to amplify its response to actively transcribed elements [9,10].

The mechanism by which Piwi silences transposons proved much more difficult to dissect. It had long been suspected that Piwi mediates transcriptional gene silencing (TGS) of TEs through impacts on chromatin, due mainly to several provocative clues. First, Piwi is a nuclear protein, and this localization is essential to its silencing capability. A mutant Piwi lacking its nuclear localization signal is found in the cytoplasm and is incapable of silencing TEs but binds piRNAs to wildtype levels [14,27,42]. In addition, Piwi's slicer activity is not necessary for silencing, as a catalytically dead Piwi mutant rescues the null mutant phenotype [14,27].

Many studies have suggested that Piwi could silence transposons at a transcriptional level by inducing changes in histone marks, much like the mechanism by which small RNAs induce heterochromatin formation in yeast [43]. In fact, the murine piRNA pathway silences transposable elements by inducing chromatin changes, ultimately resulting in DNA methylation [44,45]. In *Drosophila*, several studies support a role for Piwi in acting through TGS in the ovary; multiple groups have reported changes in histone marks on a handful of transposons upon disruption of the piRNA pathway [42,46,47], and a study by Shpiz *et al.* detected an increase in several nascent TE transcripts upon Piwi knockdown (KD) [48]. However, it was the recent study by Sienski and colleagues that definitively demonstrated that Piwi silences transposons at the transcriptional level, triggering changes in chromatin state genome wide (Figure 2). The authors took advantage of the OSS/OSC cell line and did side-by-side comparisons of RNA Polymerase II (Pol2) occupancy, trimethylation of H3K9 (a common mark of heterochromatin), nascently transcribed RNA, and steady state RNAs at a global level in Piwi KD versus control cells [49••]. They observed that in the absence of Piwi, Pol2 occupancy on transposons increased, along with an increase of nascent TE transcripts and steady state RNA levels. Furthermore, levels of H3K9me3 marks on transposons dropped in the Piwi KD as compared to controls. Interestingly, the authors also observed that many TE sequences dispersed in euchromatin trigger the formation of an H3K9me3 island that is dependent on Piwi and on transcription of the locus. This strongly implicates an RNA-recognition mode for Piwi-dependent silencing. The study also identified Maelstrom (Mael), a protein previously implicated in the germline piRNA pathway, as playing a role in transcriptional silencing of transposons [50,51]. Upon Mael KD, there was no change in levels of mature piRNAs, but there were increases in Pol2 occupancy on TEs and nascent transcripts. Interestingly, levels of H3K9me3 did not decrease when Mael was depleted; rather, H3K9 methylation appeared to spread downstream of the TE insertion, in some cases for up to 30 kb. This places Mael downstream of Piwi in silencing of TEs. The precise mechanism by which Piwi influences chromatin state remains elusive. Other than Mael, no other effector protein has been identified. One likely candidate to play a role in this process is Heterochromatin Protein 1a (HP1a), which is believed to bind H3K9 methyl groups [52,53]. HP1a has been shown to interact with Piwi, and its depletion leads to TE derepression [47,54]. The current model of piRNA-mediated TGS proposes that Piwi RISC recognizes nascent transposon transcripts by sequence complementarity and then, with the help of Mael, recruits silencing machinery to trigger histone modifications at the site of transcription (Figure 2). The association of Piwi with chromatin seems to be unstable, as the authors were unable to map it to TE loci using chromatin immunoprecipitation. It is clear that other silencing effectors in addition to H3K9 are necessary because Mael mutants do not lose H3K9me3, but have upregulation of TE transcripts. Further experiments are needed to fully understand this process. Even though it seems likely that TGS is the main silencing mode for Piwi, there remains a possibility that it is also acting through PTGS at some level. This study did not address the role of Piwi in the germline nucleus but it

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

Figure 2



Transcriptional silencing of transposable elements by Piwi-piRNA complexes in the soma. The X chromosome of *Drosophila melanogaster* (chrX) is shown. A simplistic view of its chromatin state is indicated in shades of gray. The transcriptionally active euchromatin in white harbors a full-length copy of the retroelement, *ZAM* (indicated as a red box). An inactive remnant of the same element (in red) can be found within the *flamenco* piRNA cluster in pericentromeric heterochromatin. After transcription and processing of *flamenco*, this fragment gives rise to antisense piRNAs that are loaded into Piwi in the cytoplasm (indicated as red piRNA species). Upon reimport into the nucleus, these Piwi-piRNA complexes recognize active transcription of the full-length *ZAM* copy by RNA polymerase II (Pol2) based on sequence complementarity. This recognition leads to the recruitment of additional factors such as Maelstrom (Mael) and unknown chromatin remodelers. Ultimately, the deposition of H3K9me3 marks leads to loss of Pol2 occupancy and the transcriptional silencing of *ZAM*.

seems likely that it will also silence TEs by TGS in that setting.

Germ cells might prove to be more complex because of the presence of Aub and Ago3. Although these two proteins are engaged in the ping-pong cycle in the nuage, spatially separated from Piwi in the nucleus, there seems to be a more intimate connection between these proteins than has been generally appreciated. A strong indication of this connection is that in Aub and Ago3 mutants, levels of Piwi protein decrease [5]. Furthermore, in an Ago3 mutant, the levels of Piwi-bound piRNAs decrease and there is a shift in their sense versus antisense bias [5]. Considered together, these data indicate that there is significant crosstalk between Piwi and the ping-pong cycle. One point to remember is that, although ping-pong is thought to occur mainly between Aub and Ago3, there are a significant number of Piwi:Ago3 ping-pong pairs detected in ovaries [9]. Further studies will be critical in understanding the relationship between Piwi and ping-pong, and which mechanisms are employed to silence TEs in the germline.

What is the function of maternally deposited Piwi RISC complexes?

Piwi and Aubergine, together with their bound piRNAs, are maternally deposited in the embryo and accumulate in the pole plasm, which gives rise to the future germline [8,39,40]. These maternally contributed complexes are thought to be essential in priming the piRNA pathway to be able to successfully silence elements. Previous studies have revealed that hybrid dysgenesis is caused by the failure to maternally deposit piRNAs corresponding to a paternally contributed transposon [40]. These maternally contributed Piwi and Aub RISCs may serve to jump-start the silencing pathway to target elements even before zygotic transcription has begun. Therefore, maternally deposited complexes could be one of the triggers to initiate the ping-pong cycle, which will continue throughout the life of the organism.

A recent study offers another important role for these inherited complexes. de Vanssay *et al.* found that maternally deposited piRNAs could be involved in the specification of a piRNA cluster [55•]. In a previous study, the group characterized a phenomenon known as trans-silencing effect (TSE) in which *P-element* derived transgenes inserted in a heterochromatic region can silence a distinct *P-element* derived transgene inserted at a euchromatic locus. Using this system the authors found that a transgene cluster that induces strong silencing can convert a separate, homologous locus that is normally incapable of trans-silencing, into a strong silencer, in a heritable manner. This effect is dependent on maternally deposited piRNA complexes. This implies that the inherited piRNA complexes are needed to reestablish piRNA

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

cluster definitions in the progeny. Consequently, the piRNA pathway may completely reset and cluster identity be re-acquired between each generation. This concept is analogous to piRNA-driven transposon silencing in mammals; during primordial germ cell development, the germline is stripped of all DNA methylation, which is then reacquired on TEs through the action of piRNA-driven *de novo* methylation [45]. Since *Drosophila* lacks the ability to methylate DNA, maternally deposited piRNA complexes may serve a similar role in identifying TEs in the developing progeny. However, further work is necessary to evaluate this hypothesis. For instance, it would be interesting to specifically eliminate the maternally inherited pool of Piwi RISCs to observe if cluster definitions are lost.

In embryos, although maternally deposited Piwi and Aub are both localized to pole plasm in early embryogenesis, their localization patterns rapidly change during the cellularization of the embryo. While Aub continues to reside exclusively in pole cells, Piwi localizes to the nuclei of every cell of the embryo, and continues to do so until ~12 hours after egg laying [40,54]. What role might Piwi play in somatic nuclei during embryogenesis? One interesting possibility, especially considering the recent findings implicating Piwi in TGS, is that the protein is establishing silencing marks on transposons throughout the somatic compartment. In fact, many studies have implicated Piwi in positional effect variegation (PEV), a clearly somatic effect, and have observed Piwi binding on polytene chromosomes [54,56,57]. Perhaps the suppression of transgenes observed during PEV is mediated by Piwi-induced chromatin silencing in early embryogenesis, and is maintained throughout development. Extensive additional work will be necessary to fully understand the role of maternally deposited Piwi and Aub, but there is no doubt that there are many fascinating discoveries to be made in this area.

Conclusions

It has been almost a decade since the discovery of piRNAs, and many advances have been made toward understanding the general function of the pathway. However, surprisingly little is known about several key aspects of piRNA biology, such as the mechanistic details of piRNA biogenesis and how the downstream targets of the pathway are silenced. This is because many of the important players in the pathway still remain unknown. A genome-wide screen for piRNA pathway factors would aid in identifying all proteins involved, so that a full genetic framework could finally and conclusively be established. There is also an overwhelming need to develop biochemical assays that recapitulate several aspects of the piRNA pathway *in vitro*. These could bring much needed mechanistic insights into precisely how the pathway operates. Some progress has been made in this direction with the development of the silkworm trimming assay [30••]. Following the introduction of the Drosophila OSS/OSC cell lines by Niki *et al.*, both genome wide screens and *in vitro* assays have become feasible [12]. Given these tools and recent advances described here, it is easy to imagine that we will see many more exciting discoveries and insights into how small RNAs provide an immune defense against mobile elements.

Acknowledgements

We would like to thank Clare Rebbeck and Leah Sabin for critical comments on the manuscript and helpful discussion. We would also like to thank Julius Brennecke for sharing of data before publication. P.M.G. is a NIH trainee on a CSHL WSBS NIH Kirschstein-NRSA pre-doctoral award (T32 GM065094), a William Randolph Hearst Scholar and a Leslie Quick Junior Fellow. Work in the Hannon laboratory is supported by grants from the NIH and by a kind gift from Kathryn W. Davis. G.J.H. is an investigator of the HHMI.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Levin HL, Moran JV: **Dynamic interactions between transposable elements and their hosts**. *Nat Rev Genet* 2011, **12**:615-627.
2. Malone CD, Hannon GJ: **Small RNAs as guardians of the genome**. *Cell* 2009, **136**:656-668.
3. Siomi MC, Sato K, Pezic D, Aravin AA: **PIWI-interacting small RNAs: the vanguard of genome defence**. *Nat Rev Mol Cell Biol* 2011, **12**:246-258.
4. Malone CD, Brennecke J, Dus M, Stark A, Mccombie WR, Sachidanandam R, Hannon GJ: **Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary**. *Cell* 2009, **137**:522-535.
5. Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM *et al.*: **Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies**. *Cell* 2009, **137**:509-521.
6. Lécher P, Bucheton A, Pelisson A: **Expression of the Drosophila retrovirus gypsy as ultrastructurally detectable particles in the ovaries of flies carrying a permissive flamenco allele**. *J Gen Virol* 1997, **78(Pt 9)**:2379-2388.
7. Cox DN, Chao A, Lin H: **piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells**. *Development (Cambridge, England)* 2000, **127**:503-514.
8. Harris AN, Macdonald PM: **Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C**. *Development (Cambridge, England)* 2001, **128**:2823-2832.
9. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ: **Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila**. *Cell* 2007, **128**:1089-1103.
10. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC: **A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila**. *Science (New York, NY)* 2007, **315**:1587-1590.
11. Ishizu H, Siomi H, Siomi MC: **Biology of PIWI-interacting RNAs: new insights into biogenesis and function inside and outside of germlines**. *Genes Dev* 2012, **26**:2361-2373.
12. Niki Y, Yamaguchi T, Mahowald AP: **Establishment of stable cell lines of Drosophila germ-line stem cells**. *Proc Natl Acad Sci USA* 2006, **103**:16325-16330.

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

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

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

15. Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC: **A broadly conserved pathway generates 3′UTR-directed primary piRNAs**. *Curr Biol* 2009, **19**:2066-2076.

16. • Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE: **Adaptation to P element transposon invasion in *Drosophila melanogaster***. *Cell* 2011, **147**:1551-1563.
This paper demonstrates how the piRNA system can evolve its silencing repertoire in response to challenge by a new transposon.

17. Robert V, Prud, apos, homme N, Kim A, Bucheton A, Pelisson A: **Characterization of the flamenco region of the *Drosophila melanogaster* genome**. *Genetics* 2001, **158**:701-713.

18. Rangan P, Malone CD, Navarro C, Newbold SP, Hayes PS, Sachidanandam R, Hannon GJ, Lehmann R: **piRNA production requires heterochromatin formation in Drosophila**. *Curr Biol* 2011, **21**:1373-1379.

19. Pane A, Jiang P, Zhao DY, Singh M, Schüpbach T: **The Cutoff protein regulates piRNA cluster expression and piRNA production in the Drosophila germline**. *EMBO J* 2011, **30**:4601-4615.

20. Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS, Nowosielska A *et al.*: **The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters**. *Cell* 2009, **138**:1137-1149.

21. Muerdter F, Olovnikov I, Molaro A, Rozhkov NV, Czech B, Gordon A, Hannon GJ, Aravin AA: **Production of artificial piRNAs in flies and mice**. *RNA (New York, NY)* 2012, **18**:42-52.

22. Zhang F, Wang J, Xu J, Zhang Z, Koppetsch BS, Schultz N, Vreven T, Meignin C, Davis I, Zamore PD *et al.*: **UAP56 couples piRNA clusters to the perinuclear transposon silencing machinery**. *Cell* 2012, **151**:871-884.

23. •• Ipsaro JJ, Haase AD, Knott SR, Joshua-Tor L, Hannon GJ: **The structural biochemistry of Zucchini implicates it as a nuclease in piRNA biogenesis**. *Nature* 2012.
This paper reports the biochemistry and three dimensional structure of the mouse Zucchini protein and implicates it as the nuclease that forms piRNA 5′ ends.

24. •• Nishimasu H, Ishizu H, Saito K, Fukuhara S, Kamatani MK, Bonnefond L, Matsumoto N, Nishizawa T, Nakanaga K, Aoki J *et al.*: **Structure and function of Zucchini endoribonuclease in piRNA biogenesis**. *Nature* 2012.
Along with the Ipsaro paper, this report of the structure and biochemistry of Drosophila Zucchini may have solved the mystery of piRNA 5′ processing.

25. Huang H, Gao Q, Peng X, Choi S-Y, Sarma K, Ren H, Morris AJ, Frohman MA: **piRNA-associated germline nuage formation and spermatogenesis require MitoPLD profusogenic mitochondrial-surface lipid signaling**. *Dev Cell* 2011, **20**:376-387.

26. Watanabe T, Chuma S, Yamamoto Y, Kuramochi-Miyagawa S, Totoki Y, Toyoda A, Hoki Y, Fujiyama A, Shibata T, Sado T *et al.*: **MITOPLD is a mitochondrial protein essential for nuage formation and piRNA biogenesis in the mouse germline**. *Dev Cell* 2011, **20**:364-375.

27. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC: **Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in Drosophila**. *Genes Dev* 2010, **24**:2493-2498.

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

29. Arnoult D, Soares F, Tattoli I, Girardin SE: **Mitochondria in innate immunity**. *EMBO Rep* 2011, **12**:901-910.

30. •• Kawaoka S, Izumi N, Katsuma S, Tomari Y: **3′ end formation of PIWI-interacting RNAs in vitro**. *Mol Cell* 2011, **43**:1015-1022.
This paper reports one of the few successful attempts at addressing questions of piRNA biology using *in vitro*, biochemical approaches and provides a mechanism, if not the enzyme, for piRNA 3′ end formation

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

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

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

34. Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, Hannon GJ, Lehmann R: **Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila**. *Development (Cambridge, England)* 2011, **138**:4039-4050.

35. Olivieri D, Senti K-A, Subramanian S, Sachidanandam R, Brennecke J: **The cochaperone shutdown defines a group of biogenesis factors essential for all piRNA populations in Drosophila**. *Mol Cell* 2012, **47**:954-969.

36. Preall JB, Czech B, Guzzardo PM, Muerdter F, Hannon GJ: **shutdown is a component of the Drosophila piRNA biogenesis machinery**. *RNA (New York, NY)* 2012, **18**:1446-1457.

37. Xiol J, Cora E, Koglgruber R, Chuma S, Subramanian S, Hosakawa M, Reuter M, Yang Z, Berninger P, Palencia A *et al.*: **A role for Fkbp6 and the chaperone machinery in piRNA amplification and transposon silencing**. *Mol Cell* 2012, **47**:970-979.

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

39. Megosh HB, Cox DN, Campbell C, Lin H: **The role of PIWI and the miRNA machinery in Drosophila germline determination**. *Curr Biol* 2006, **16**:1884-1894.

40. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ: **An epigenetic role for maternally inherited piRNAs in transposon silencing**. *Science (New York, NY)* 2008, **322**:1387-1392.

41. Grentzinger T, Armenise C, Brun C, Mugat B, Serrano V, Pelisson A, Chambeyron S: **piRNA-mediated transgenerational inheritance of an acquired trait**. *Genome Res* 2012, **22**:1877-1888.

42. Klenov MS, Sokolova OA, Yakushev EY, Stolyarenko AD, Mikhaleva EA, Lavrov SA, Gvozdev VA: **Separation of stem cell maintenance and transposon silencing functions of Piwi protein**. *Proc Natl Acad Sci USA* 2011, **108**:18760-18765.

43. Huisinga K, Elgin S: **Small RNA-directed heterochromatin formation in the context of development: what flies might learn from fission yeast**. *Biochim Biophys Acta* 2008, **1789**:3-16.

44. Aravin AA, Sachidanandam R, Girard A, Fejes Toth K, Hannon GJ: **Developmentally regulated piRNA clusters implicate MILI in transposon control**. *Science (New York, NY)* 2007, **316**:744-747.

45. Aravin AA, Bourc'his D: **Small RNA guides for de novo DNA methylation in mammalian germ cells**. *Genes Dev* 2008, **22**:970-975.

46. Klenov MS, Lavrov SA, Stolyarenko AD, Ryazansky SS, Aravin AA, Tuschl T, Gvozdev VA: **Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the *Drosophila melanogaster* germline**. *Nucleic Acids Res* 2007, **35**:5430-5438.

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

47. Wang SH, Elgin SCR: **Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line**. *Proc Natl Acad Sci USA* 2011, **108**:21164-21169.

48. Shpiz S, Olovnikov I, Sergeeva A, Lavrov S, Abramov Y, Savitsky M, Kalmykova A: **Mechanism of the piRNA-mediated silencing of Drosophila telomeric retrotransposons**. *Nucleic Acids Res* 2011, **39**:8703-8711.

49. •• Sienski G, Dönertas D, Brennecke J: **Transcriptional silencing of transposons by Piwi and Maelstrom and its impact on chromatin state and gene expression**. *Cell* 2012, **151**:964-980.
This paper provides a genome-wide view of changes in chromatin structure upon Piwi silencing and provides concrete evidence that Piwi operates in the soma by regulating transposons at the level of their transcription.

50. Findley SD, Tamanaha M, Clegg NJ, Ruohola-Baker H: **Maelstrom, a Drosophila spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage**. *Development (Cambridge, England)* 2003, **130**:859-871.

51. Lim AK, Kai T: **Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster***. *Proc Natl Acad Sci USA* 2007, **104**:6714-6719.

52. Lachner M, apos O, Carroll D, Rea S, Mechtler K, Jenuwein T: **Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins**. *Nature* 2001, **410**:116-120.

53. Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T: **Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain**. *Nature* 2001, **410**:120-124.

54. Brower-Toland B, Findley SD, Jiang L, Liu L, Yin H, Dus M, Zhou P, Elgin SCR, Lin H: **Drosophila PIWI associates with chromatin and interacts directly with HP1a**. *Genes Dev* 2007, **21**:2300-2311.

55. • de Vanssay A, Bougé A-L, Boivin A, Hermant C, Teysset L, Delmarre V, Antoniewski C, Ronsseray S: **Paramutation in Drosophila linked to emergence of a piRNA-producing locus**. *Nature* 2012, **490**:112-115.
The authors implicate maternally deposited piRNAs as being important for specifying the identity of piRNA clusters.

56. Haynes K, Caudy A, Collins L, Elgin S: **Element 1360 and RNAi components contribute to HP1-dependent silencing of a pericentric reporter**. *Curr Biol* 2006, **16**:2222-2227.

57. Pal-Bhadra M: **Heterochromatic silencing and HP1 localization in Drosophila are dependent on the RNAi machinery**. *Science (New York, NY)* 2004, **303**:669-672.

Please cite this article in press as: Guzzardo PM, et al.: The piRNA pathway in flies: highlights and future directions, Curr Opin Genet Dev (2013), http://dx.doi.org/10.1016/j.gde.2012.12.003

Production of artificial piRNAs in flies and mice

FELIX MUERDTER,[1,2,5] IVAN OLOVNIKOV,[3,4,5] ANTOINE MOLARO,[1,5] NIKOLAY V. ROZHKOV,[1,5] BENJAMIN CZECH,[1,2] ASSAF GORDON,[1] GREGORY J. HANNON,[1,6] and ALEXEI A. ARAVIN[3,6]

[1]Howard Hughes Medical Institute, Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA
[2]Zentrum für Molekularbiologie der Pflanzen, Entwicklungsgenetik, University of Tübingen, 72076 Tübingen, Germany
[3]California Institute of Technology, Division of Biology, Pasadena, California 91125, USA
[4]Institute of Molecular Genetics, Russian Academy of Sciences, 123182 Moscow, Russia

ABSTRACT

In animals a discrete class of small RNAs, the piwi-interacting RNAs (piRNAs), guard germ cell genomes against the activity of mobile genetic elements. piRNAs are generated, via an unknown mechanism, from apparently single-stranded precursors that arise from discrete genomic loci, termed piRNA clusters. Presently, little is known about the signals that distinguish a locus as a source of piRNAs. It is also unknown how individual piRNAs are selected from long precursor transcripts. To address these questions, we inserted new artificial sequence information into piRNA clusters and introduced these marked clusters as transgenes into heterologous genomic positions in mice and flies. Profiling of piRNA from transgenic animals demonstrated that artificial sequences were incorporated into the piRNA repertoire. Transgenic piRNA clusters are functional in non-native genomic contexts in both mice and flies, indicating that the signals that define piRNA generative loci must lie within the clusters themselves rather than being implicit in their genomic position. Comparison of transgenic animals that carry insertions of the same artificial sequence into different ectopic piRNA-generating loci showed that both local and long-range sequence environments inform the generation of individual piRNAs from precursor transcripts.

Keywords: piwi; noncoding RNA; piRNA

INTRODUCTION

In several animals, including *Drosophila* and mammals, piRNAs have been shown to form the core of a small RNA-based innate immune system that recognizes and represses mobile elements (Saito et al. 2006; Vagin et al. 2006; Aravin et al. 2007a; Brennecke et al. 2007; Gunawardane et al. 2007; Malone and Hannon 2009; Siomi et al. 2011). This function is essential for proper germ-line development, and mutations in the piRNA pathway lead to male and/or female sterility (Cox et al. 2000; Harris and Macdonald 2001; Li et al. 2009; Malone and Hannon 2009). In essence, piRNAs play a major role in defining genomic content as being transposon related; piRNAs comprise a catalog of transposon sequences that an organism has defined as targets for repression (Brennecke et al. 2007). Omission from that catalog can mean that an element escapes repression. In the case of flies, the lack of an effective piRNA-based definition for the *I-* or *P-element* in some strains means that introduction of even this single transposon can lead to highly penetrant sterility (Pelisson 1981; Rubin et al. 1982; Brennecke et al. 2008).

Sequencing of piRNA populations has revealed their extreme diversity; literally, millions of distinct piRNA sequences can be identified in a single individual (Aravin et al. 2006, 2007b; Girard et al. 2006; Brennecke et al. 2007; Houwing et al. 2007; Lau et al. 2009). Genomic mapping indicates that piRNAs arise from three different types of loci. First, the dominant source of piRNAs can be found in so-called piRNA clusters (Aravin et al. 2006, 2007b; Brennecke et al. 2007). These loci range from a few kilobases to >200 kb in size. They are often strongly enriched in transposon sequences, in accord with a function of the piRNA pathway in transposon control (Vagin et al. 2006; Brennecke et al. 2007; Gunawardane et al. 2007). In the majority of cases, clusters generate a mixture of small RNAs, with some sense and some antisense to each targeted transposon. Second, piRNAs can be derived from protein-coding genes, with these almost invariably being sense species from 3′ UTRs (Aravin et al. 2008; Robine et al.

[5]These authors contributed equally to this work.
[6]**Corresponding authors.**
E-mail aaa@caltech.edu.
E-mail hannon@cshl.edu.
Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.029769.111.

2009; Saito et al. 2009). It is as yet unclear whether a single transcript isoform can be either translated into protein or processed into small RNAs or whether a specific transcript variant serves as a piRNA precursor. Only a few genes give rise to piRNAs, and these do not show uniformly high expression, suggesting that some specific determinant or motif, rather than a high-transcript abundance marks specific genes for processing. Third, piRNAs can arise from dispersed, euchromatic transposon copies (Brennecke et al. 2007, 2008; Aravin et al. 2008). These are often full length and close to consensus, representing the potentially active representatives of each transposon family.

The three types of piRNA generative loci produce small RNAs through two different mechanisms. piRNA clusters and genic loci generate "primary" piRNAs, which appear to be sampled from long, single-stranded transcripts through the action of an unknown nucleolytic machinery (Aravin et al. 2006, 2007b; Brennecke et al. 2007; Malone et al. 2009). Abundant primary piRNAs share no apparent sequence or structural motifs except for the presence of a 5′ terminal U residue (1U), which may reflect a binding preference of some Piwi family proteins. Secondary piRNAs are produced through a slicer-dependent mechanism, termed the ping-pong cycle and have a characteristic bias for an A at position 10 (paired with the 1U in the primary piRNA) (Brennecke et al. 2007; Gunawardane et al. 2007).

Combined analysis of piRNA sequences and animals bearing mutations in piRNA pathway components has led to a model for the role of these small RNAs (Malone and Hannon 2009; Saito and Siomi 2010; Senti and Brennecke 2010; Siomi et al. 2011). piRNA clusters produce a multitude of individual piRNAs, and the sequence content of piRNA cluster defines sequences of mature piRNAs generated from it. With the notable exception of pachytene piRNAs that are expressed during male meiosis in mouse, piRNA clusters in both flies and mice are highly enriched in transposable element sequences. The sequence content of the piRNA clusters determines the capacity of the system to respond to a given element, in essence comprising an organisms' evolving molecular definition of transposons. Inherent in this scenario is the ability of the system to adapt to colonization by new elements by incorporating their sequence into a piRNA cluster. A clear example can be found in the *P-element*, which swept through global *Drosophila melanogaster* populations after the sequestration of common laboratory strains (Rubin et al. 1982). Laboratory strains have no ability to repress the *P-element*. In retrospect, studies of strains with natural or acquired *P-element* resistance suggested that integration of the element into a piRNA cluster was key to its control (Ronsseray et al. 1991, 1996, 2003).

Here, we sought to test whether the ability to translate new genomic content into small RNAs was a general characteristic of piRNA loci in flies and mice. We find that clusters can be programmed to produce artificial piRNAs (apiRNAs). Furthermore, we were able to separate functional piRNA clusters from their native genomic locations, indicating that the clusters themselves contain sufficient information to funnel their RNA products into the piRNA biogenesis pathway. We made use of marked transgenic clusters that carry insertions of the same sequence into different contexts to evaluate the features that lead to the production of individual piRNA species. We find that critical determinants lie both in the local and long-range sequence environments of the piRNA cluster.

RESULTS AND DISCUSSION

The current model for acquiring piRNA-dependent resistance against new transposon invasion implies that insertion of active transposons into an existing piRNA cluster leads to the generation of new piRNA species and enables element repression. This model suggests that any sequence, if inserted into a piRNA cluster, will lead to the generation of new piRNAs. Though attractive, this model has not been rigorously tested. Acquisition of natural resistance against transposable elements by transposition into piRNA clusters is difficult to study in an experimental setting. However, this scenario can be modeled using transgenes carrying new sequence information within a piRNA-generating locus.

Over the years, large collections of *Drosophila* stocks have been produced that carry transgenes integrated randomly throughout the genome. We took advantage of these tools by searching for integration events in native piRNA clusters. The line P{lArB}A171.1F1 (also known as P-1152) has a 18.3-kb construct P{lArB} integrated into a telomeric piRNA cluster on the X-chromosome (chromosomal location 1A) (Wilson et al. 1989; Roche and Rio 1998). The P{lArB} transgene contains sequences derived from the *hsp70*, *Adh*, and *rosy* genes of *D. melanogaster* and a bacterial *lacZ* gene. Unlike P{lArB} insertions in other genomic sites, P-1152 has unusual properties. It is able to suppress the expression of other *lacZ* transgenes in germ cells, a phenomenon termed *trans*-silencing (Fig. 1A; Supplemental Fig. S1; Ronsseray et al. 1991). The P{lArB} insertion in P-1152 is mapped to the Telomere Associated Sequence (TAS) repeats that produce abundant piRNAs from both genomic strands. These piRNAs are loaded into Piwi, Aub, and Ago3 in the germ cells of *D. melanogaster* ovaries (Brennecke et al. 2007). Aub and Ago3-loaded piRNAs derived from TAS repeats display the characteristic features of the ping-pong amplification cycle, including a prevalent 10-nt 5′ overlap of sense and antisense species and an enrichment for an A at position 10 of secondary piRNAs. The *trans*-silencing properties of P-1152 transgene and the association of these properties with its localization in the piRNA cluster suggested that insertion of *lacZ* into an existing piRNA cluster led to the generation of new anti-*lacZ* piRNAs that are able to suppress cognate transcripts in germ cells. Indeed, the presence of small

RNAs complementary to *lacZ* was recently demonstrated using RNAse-protection assay in ovaries of the P-1152 line (Todeschini et al. 2010).

To analyze more deeply any artificial piRNAs derived from the P{lArB} transgene, we sequenced small RNAs from ovaries of the P-1152 line, examining a size range from 18 to 29 nt. This includes piRNAs, siRNAs, and miRNAs. P{lArB} generated abundant small RNA species that mapped to both genomic strands (Fig. 1B). Their size profile indicated that the majority were likely piRNAs, ranging from 23 to 27 nt, while a minor fraction corresponded to 21-nt endo-siRNAs that are also a product of bidirectionally transcribed piRNA loci (Fig. 1C; Czech et al. 2008; Lau et al. 2009). Further analysis confirmed that the 23- to 27-nt RNAs were genuine piRNAs that could be separated into primary (1U-biased) and secondary (10A-biased) populations (Fig. 1D; data not shown). Transgene piRNAs mapping to opposite genomic strands tended to have a 10-nt overlap between their 5′ ends that is a characteristic feature of the ping-pong cycle (Fig. 1E). Notably, P{lArB} contains the only *lacZ* sequence information in the P-1152 strain. Since signatures of the ping-pong cycle were evident for *lacZ*-derived piRNAs, this demonstrates unequivocally that cluster transcripts derived from the plus and minus genomic strands can participate in the piRNA amplification loop. Native *Adh* and *rosy* transcripts are not processed into piRNAs in wild-type flies (data not shown). Therefore, it is unlikely that any specific signals that trigger piRNA processing might be present in these genes. Moreover, bacterial sequences are unlikely to have evolved as a trigger for piRNA production. Thus, our results indicate that, when present in the context of a piRNA cluster, virtually any sequence can serve as a substrate for piRNA biogenesis. We confirmed previous observations that the P{lArB} transgene inserted in TAS is able to silence *lacZ* expression from separate, euchromatic locations (Supplemental Fig. S1), demonstrating that artificial anti-*lacZ* piRNAs are functional and able to silence transcripts that share sequence content in *trans*.

piRNAs are processed from the entire P{lArB} transgene independently of the origin of the inserted fragments; both *D. melanogaster* and bacterial sequences generate piRNAs with similar efficiency (Fig. 1B). Throughout the construct there are approximately twofold more piRNAs derived from the plus than from the minus genomic strand independently of the orientation of the genes within the construct, just as is observed for native components of the cluster. For example, *Adh* and *rosy* have different orientations, but for both fragments the majority of piRNAs are mapped to the plus genomic strand. RT–PCR shows that *rosy* transcripts are present in ovaries of P-1152 females, but absent in wild-type flies or flies that have a P{lArB} insertion outside of the piRNA cluster (Supplemental Fig. S2), indicating that *rosy* expression is dependent on insertion of P{lArB} into TAS. Overall, both the distribution of piRNAs along P{lArB} transgene and RT–PCR results suggest that transcript of both plus- and minus-strand RNAs, which are processed to piRNAs, initiates outside of the transgenic construct, likely within adjacent TAS sequences.

Mapping of piRNAs to P{lArB} revealed that intronic sequences present within *Adh* and *rosy* gave rise to piRNA from both genomic strands. Even when present in the sense orientation, where the intron could have been removed by the splicing apparatus, piRNA levels remained comparable in adjacent intronic and exonic regions. The generation of piRNA from intronic sequence is unexpected, as primary piRNA bio-

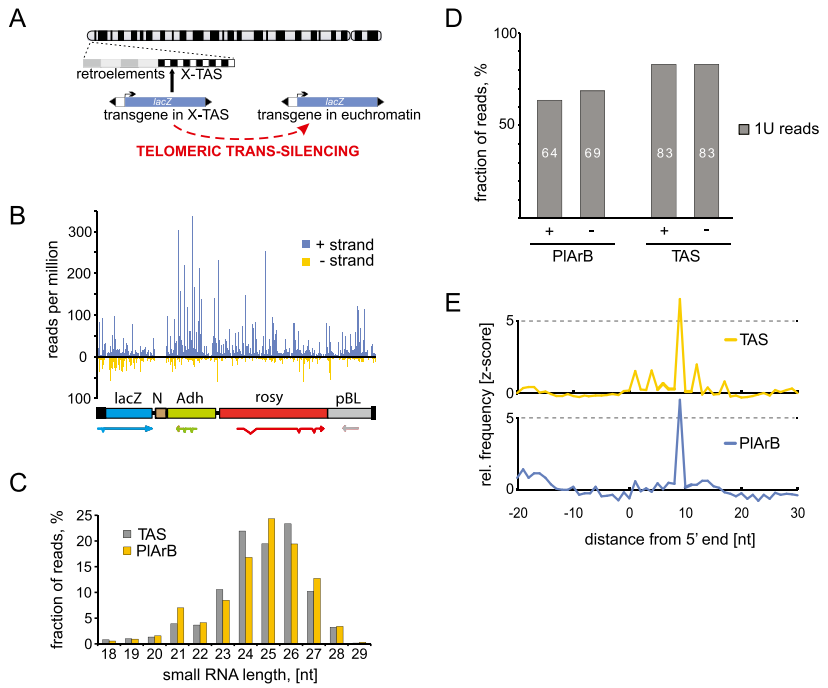


FIGURE 1. Production of artificial piRNAs (apiRNAs) from the *Drosophila* X-TAS cluster. (*A*) The P{lArB} insertion into the X-TAS cluster is shown schematically along with an illustration of *trans*-silencing. (*B*) *Below* is a schematic of the P{lArB} insert with the inferred structures of the transcripts it can produce (see text). N is an area where the sequence is unknown. *Above* is a plot of piRNA read frequencies along the plus and minus strands (indicated) of the element. (*C*) Small RNA lengths are plotted as a fraction of reads for TAS and for the inserted element. (*D*) Fractions of reads beginning with a 5′ U are plotted for the P{lArB} and TAS plus and minus strands. (*E*) The degree of 5′ overlap for piRNAs from the plus and minus strands for P{lArB} and TAS were quantified and plotted as relative frequencies (Z-scores). The spike at position 9 is a signature of the ping-pong amplification cycle.

genesis is thought to occur in the cytoplasm and has been linked to specific cytoplasmic bodies, e.g., nuage and Yb bodies, which concentrate components such as zucchini and armitage, which are implicated in piRNA processing (Tomari et al. 2004; Lim and Kai 2007; Pane et al. 2007; Malone et al. 2009; Haase et al. 2010; Olivieri et al. 2010; Saito et al. 2010; Qi et al. 2011). Furthermore, genic piRNAs that are processed from mRNA of protein-coding genes in *Drosophila* and mice are mapped almost exclusively to exonic sequences (Aravin et al. 2008; Robine et al. 2009; Gan et al. 2011). To reconcile these disparities, we searched explicitly for piRNAs that crossed predicted exon–exon junctions, since these must arise from spliced mRNAs. We did detect a few such small RNAs for *rosy* and *Adh*, coming only from the genomic strand with the intron in the appropriate orientation for splicing to occur. Considered together, these data suggest a model in which piRNA biogenesis normally occurs following intron removal, but that recognition of some RNA processing signals might be suppressed when they are present within a piRNA cluster. In this regard, strand-specific RT–PCR indicated that more than half of sense-oriented *rosy* transcripts are not spliced in P-1152 ovaries (Supplemental Fig. S2). Suppression of conventional RNA processing signals within piRNA clusters would make sense in many ways, since the insertion of a new element would often bring at least a polyadenylation signal, which under normal circumstances could negate the production of piRNAs downstream from that site by terminating transcription or preventing the export of piRNA precursors.

Generation of artificial piRNAs by insertion of a new sequence into a piRNA cluster provides a molecular tag that allows the monitoring of cluster function even if the native, nontagged cluster is present in the same genome. We exploited this fact to test whether the presence of piRNA clusters at precise genomic positions was important to their function.

In flies, piRNA clusters occur mainly at the boundaries between heterochromatin and euchromatin, particularly in pericentromeric regions (Brennecke et al. 2007). In mammals, piRNA clusters that are expressed in meiotic cells occur in strictly syntenic positions, even though the sequence content of these loci is not conserved (Aravin et al. 2006; Girard et al. 2006; Lau et al. 2006). These observations have strongly suggested that the genomic context of piRNA clusters might be key to their function. Precedent can be drawn from plants and fission yeast, where small RNAs are generated from loci whose function relies upon the presence of normally repressive chromatin marks (Huisinga and Elgin 2009; Lahmy et al. 2010). In turn, the repressive chromatin marks themselves are maintained by small RNA-directed complexes, closing the cycle. To determine whether specific chromatin environments, which are a property of the genomic context of piRNA clusters, are essential for piRNA production, we created ectopic insertions of tagged piRNA clusters in non-native sites.

As one test of the aforementioned hypothesis, we examined the position dependence of the *flamenco* cluster in *Drosophila* (Fig. 2A). *Flamenco* is present at the boundary between euchromatin and pericentromeric heterochromatin on the *Drosophila* X chromosome, and its position proximal to the *DIP1* gene is conserved through at least 12 M years of *Drosophila* evolution (Sarot et al. 2004; Brennecke et al. 2007; Malone et al. 2009). It produces piRNAs from only one genomic strand and is exclusively expressed in the somatic follicle cells of the ovary. We selected a P[acman] BAC clone that extended from a position ~30 kb upstream of the first annotated piRNA ~86 kb toward the X chromosome centromere (Venken et al. 2009). This encompassed ~30% of the *flamenco* cluster. To distinguish any ectopic copies of *flamenco* from the native locus, we marked the BAC by recombineering, inserting a cassette comprising a nonfunctional GFP sequence and a bacterial neomycin resistance gene (Copeland et al. 2001; Venken et al. 2006; Sharan et al. 2009). Marker sequences were inserted ~4 kb downstream from the first annotated piRNA in a site, which we had previously shown to produce abundant small RNAs.

For mice, we chose to modify a piRNA cluster on mouse chromosome 17 that is a major contributor to piRNA populations in developing male germ cells from the pachytene stage through the end of meiosis (Fig. 2B; Aravin et al. 2006; Girard et al. 2006). This cluster occurs in syntenic locations in rat and in human, indicating conservation through at least 80 M years of evolution. Like *flamenco*, each region of the ch17 cluster produces piRNAs from only one genomic strand. A mouse BAC clone comprising ~187 kb of chromosome 17 carried the complete ch17 cluster and extended 60 kb upstream of and 30 kb downstream from the locus. It was similarly marked by recombineering to insert a modified GFP/neo cassette.

In flies, we took advantage of a phiC-31 attachment site in the P[acman]-BAC to insert the modified *flamenco* cluster into a known genomic locus (Venken et al. 2006, 2009). Given that *flamenco* is normally present in a location annotated as heterochromatic (X chromosome, band 20A), we chose a gene-rich, euchromatic site to insert the transgene. Specifically, we created lines with one additional copy of *flamenco* on chromosome 3L at band 62E1 (landing pad 31) (Venken et al. 2006). For mice, we used standard pronuclear injection to create two independent founder lines (R13 and R37) with ch17 transgene insertions in presumably distinct random locations.

Small RNA cloning and Illumina sequencing revealed that abundant piRNAs derived from GFP were produced from ectopic clusters in both flies and mice (Fig. 2C,D). Like the native loci, these produced small RNAs from only one genomic strand. Unlike X-TAS, neither *flamenco* nor the ch17 cluster normally participate in the ping-pong

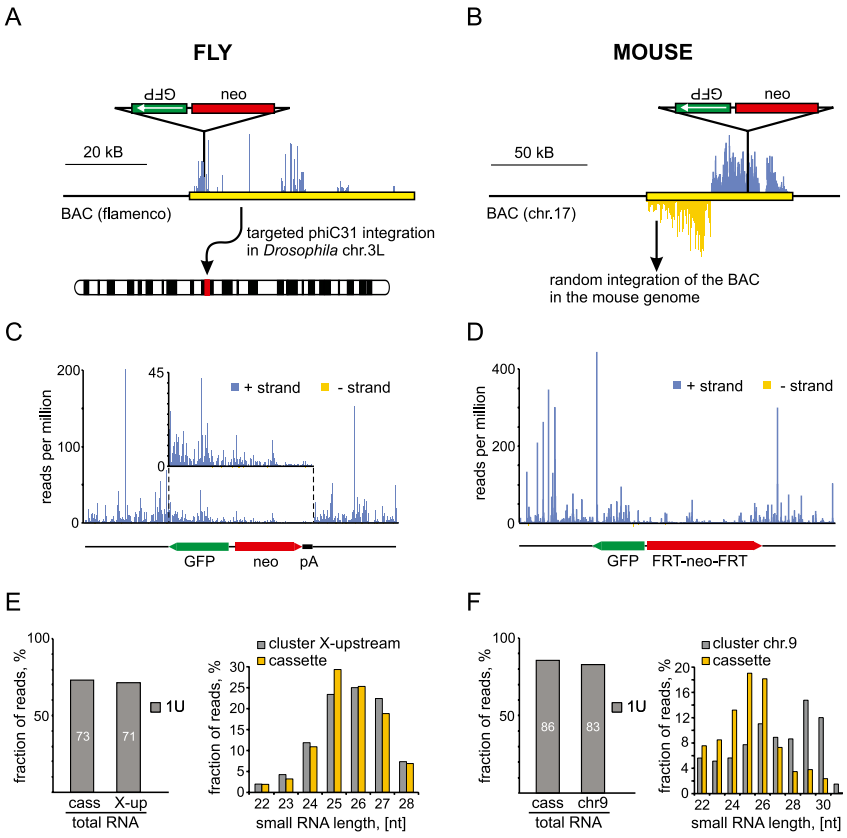


FIGURE 2. Generation of apiRNAs from ectopic clusters in flies and mice. (*A*) A schematic representation of the GFP/Neo cassette is shown along a diagram of the *flamenco* locus (in yellow, piRNA densities in blue) in the BAC used for transgenesis. *Below* is a schematic indicating that the transgene is inserted into chromosome 3L. (*B*) The GFP/neo insertion into the mouse chromosome 17 cluster is diagrammed as in *A*. (*C*) The structure of the *flamenco* GFP/Neo insertion is diagrammed *below* a plot of piRNA frequencies along the insert on the plus and minus strands (indicated). For reference, piRNAs are also mapped to flanking regions, though these represent a mixture of RNAs derived from the two native and one ectopic *flamenco* clusters. (*D*) A scheme of the GFP/Neo insert into the mouse chromosome 17 cluster is shown *below* piRNAs mapping to the insert and its context as in *C*. Again, piRNAs that flank the insert can be derived from the two native or inserted ectopic loci. (*E*) The 1U bias (*left*) and size distributions (*right*) of apiRNAs from the ectopic *flamenco* cluster are compared with another piRNA cluster (X-upstream) that also produces piRNAs from one genomic strand in follicle cells. (*F*) As in *E*, apiRNAs from the ectopic ch17 cluster in mice are compared with a similarly structured cluster on chromosome 9.

amplification loop, and the ectopic insertions also lacked signatures of the cycle, namely, small RNAs with a 10A bias and sense/antisense pairs that overlap by 10 nt. Small RNAs from the ectopic clusters did show the strong 1U bias that is a signature of primary piRNA populations (Fig. 2E,F; Supplemental Fig. S3A).

It seemed likely that the transgenic clusters would generate piRNAs both from the inserted marker gene and from sequences that represent their native content; however, it is impossible to distinguish the latter from piRNAs derived from endogenous loci. The ectopic cluster is present as a single copy in the genome, as compared with two endogenous copies. We might therefore expect piRNA levels coming from shared regions to increase by 1.5-fold if all copies were equally active. Indeed, we noted a 1.3-fold increase in piRNAs, which are derived from the portion of the *flamenco* cluster present in transgene. Similarly, the levels of MILI and MIWI piRNAs derived from the chr17 cluster in mouse increased by between 1.2- and 1.5-fold relative to a nonmodified cluster on ch9 in two independent transgenic lines. The profiles of piRNA mapped to the *flamenco* and ch17 clusters are very similar in wild-type and transgenic flies and mice (Supplemental Fig. S4). Therefore, the heterologous insertion of a marker gene does not appear to exert a strong influence on the processing of piRNAs from transgenic loci. Overall, our data indicate that transgenic piRNA clusters have similar activity to their endogenous counterparts, despite being present at non-native genomic positions.

Flamenco-derived piRNAs associate exclusively with Piwi, the only family member that is expressed in follicle cells (Sarot et al. 2004; Brennecke et al. 2007). Thus, they have a characteristic size profile, peaking at around 25 nt. piRNAs from the ectopic *flamenco* insertion shared this size distribution (Fig. 2E). piRNAs from the ch17 cluster (and other murine clusters expressed during meiosis) normally associate with both MILI and MIWI (Supplemental Fig. S3B). These complexes have distinct small RNA size profiles, with MILI to associate with a ~26-nt and MIWI harboring a ~30-nt species (Fig. 2F; Aravin et al. 2006; Girard et al. 2006). Overall, MIWI-bound species are substantially more abundant than MILI bound species (Aravin et al. 2006; Girard et al. 2006). While the ectopic ch17 cluster produced small RNAs with sizes characteristic of MILI and MIWI complexes, their ratio was very different than expected based upon the behavior of the native cluster (Fig. 2F; Supplemental Fig. S3B). RNAs with the size of MILI partners greatly outnumbered those with the size of MIWI-bound species. Thus, the ectopic cluster appeared to have a strong preference for one of its two potential Piwi-family partners (Fig. 2F; Supplemental Fig. S5).

Overall, our data indicate that piRNA clusters can function even when divorced from their normal genomic locale. With *flamenco*, the ectopic insertion behaved indistinguishably from the native locus, even though it had been substantially truncated on the centromere-proximal

side. For the ch17 cluster, piRNAs were still produced in abundance from the ectopic insertions, but the behavior of the small RNAs shifted toward preferential MILI association. This could indicate that some element of chromosomal context was important for signaling an ultimate association with MIWI or perhaps that critical signals that mark the cluster as a source of MIWI piRNAs were missing from our BAC clone, despite its extending well beyond the two ends of the cluster. Our results by no means rule out chromatin structure as a contributory element in defining piRNA clusters. However, if specific chromatin structures are important, the signals for their formation must be tightly linked to the piRNA loci themselves.

The inclusion of the same artificial sequence in piRNA clusters in multiple locations and in distinct organisms afforded the opportunity to probe the determinants of piRNA selection. In contrast to miRNAs and siRNAs, whose processing from longer precursors is informed by their specific secondary structure and is well understood, no rules that explain the selection of individual piRNAs have been defined. The only bioinformatic study that addressed this question came to the conclusion that the processing of individual piRNA from precursors is quasi-random, with only weak influences of local sequence (positions −1 to +4 relative to the 5′ end of the piRNA) (Betel et al. 2007). However, sequencing efforts from our and other groups showed that individual piRNAs are not produced uniformly along clusters. Instead, certain small RNAs appear substantially more abundant (Aravin et al. 2006, 2008; Girard et al. 2006; Brennecke et al. 2007). Characteristics underlying these inequalities could be intrinsic to the local sequence environment of each individual piRNA or could be conferred by long-distance interactions and formation of secondary structures within the precursor molecule. Alternatively, patterns could be essentially random, with the abundance of each species being determined stochastically.

As with native piRNAs, read distributions along the marker cassettes in the ectopic clusters were very uneven (Fig. 3A; Supplemental Fig. S5). Focusing on the GFP coding sequence, 1% of nucleotide residues contribute 19% of all 5′ ends of GFP-mapping piRNA reads in flies, while 10% of positions account for 70% of reads (Fig. 3B). In mouse, the distribution was even more skewed with 1% of GFP residues contributing 42% of piRNA reads (Fig. 3B). To probe the causes leading to these skewed distributions, we compared GFP-derived piRNAs in the two independent mouse transgenic lines. The correlation in the abundance of individual small RNAs was remarkable (R^2 = 0.99) (Fig. 3C), ruling out the notion that the patterns that we observe are random within each sample. Procedures for preparing small RNA libraries include steps with well-established sequence-based biases, namely, RNA adapter ligations and PCRs (Linsen et al. 2009). We therefore considered the possibility that those biases dominated apparent sequence preferences in apiRNA generation. However, very little correlation was seen between GFP piRNAs in flies and mice (R^2 = 0.01) (Fig. 3D), contrary to what one would expect if the patterns that we observed were strongly influenced by the biases of library preparation methods.

Considered as a whole, our data strongly support the existence of signals that determine the efficiency of production of individual piRNAs and raise several possibilities as to the nature of those signals. First, the biased distribution of piRNAs could be an exclusive consequence of their context within the cluster. This would imply that large-scale features, such as the structure of the transcript or preferential entry sites for the primary processing machinery determine differential piRNA production, akin to the generation of phased siRNAs from long dsRNAs in plants and animals (Zamore et al. 2000; Howell et al. 2007). Alternatively, determinants of efficient piRNA biogenesis could still be defined by the local sequence environment of each individual piRNA, with sequence determinants being interpreted differently in our two experimental models. To begin to discriminate between these possibilities, it was necessary to insert the same sequence (GFP) into different piRNA precursors that are expressed and processed in the same cell type.

The *traffic jam* (*tj*) gene encodes a basic leucine zipper transcription factor and is expressed in the follicle cells of the *Drosophila* ovary, just as is *flamenco* (Li et al. 2003; Saito et al. 2009). Importantly, *tj* generates piRNAs from a discrete segment of its 3′-UTR region (Saito et al. 2009). We created a marked, ectopic copy of *tj* by inserting a GFP coding sequence in the antisense orientation into its piRNA-producing domain and integrated this into a euchromatic site on chromosome 3L (Fig. 4A).

Sequencing of small RNAs (Fig. 4B) yielded abundant piRNAs from the inserted GFP sequence. These had the same characteristics as native *tj*-derived piRNAs, including being produced from the sense strand of the locus, having a size distribution characteristic of Piwi-associated species, and a strong bias for a 5′ terminal U residue (Fig. 4C,D). Position-dependent differences in piRNA abundance were also apparent, with the most abundant 10% of possible GFP piRNAs contributing 81% of all GFP-mapping reads (Fig. 3B).

To discriminate local- from long-distance sequence effects, we compared the abundance of individual piRNAs from the *tj* and *flamenco* transgenes. As compared with the patterns derived from independent insertions of the same transgenes in mice (R^2 = 0.99) (Fig. 3C), patterns of GFP piRNAs from *tj* and *flamenco* appeared quite different (R^2 = 0.24) (Fig. 4D). However, they were much more similar than patterns produced in mouse versus fly (R^2 = 0.01) (Fig. 3D). At the extremes, uridine positions in GFP that generate abundant piRNAs from the *flamenco* transgene tended also to generate abundant piRNAs from *tj* (Fig. 4E). Conversely, those that did not generate piRNAs from *flamenco* did not generate piRNAs from *tj*.

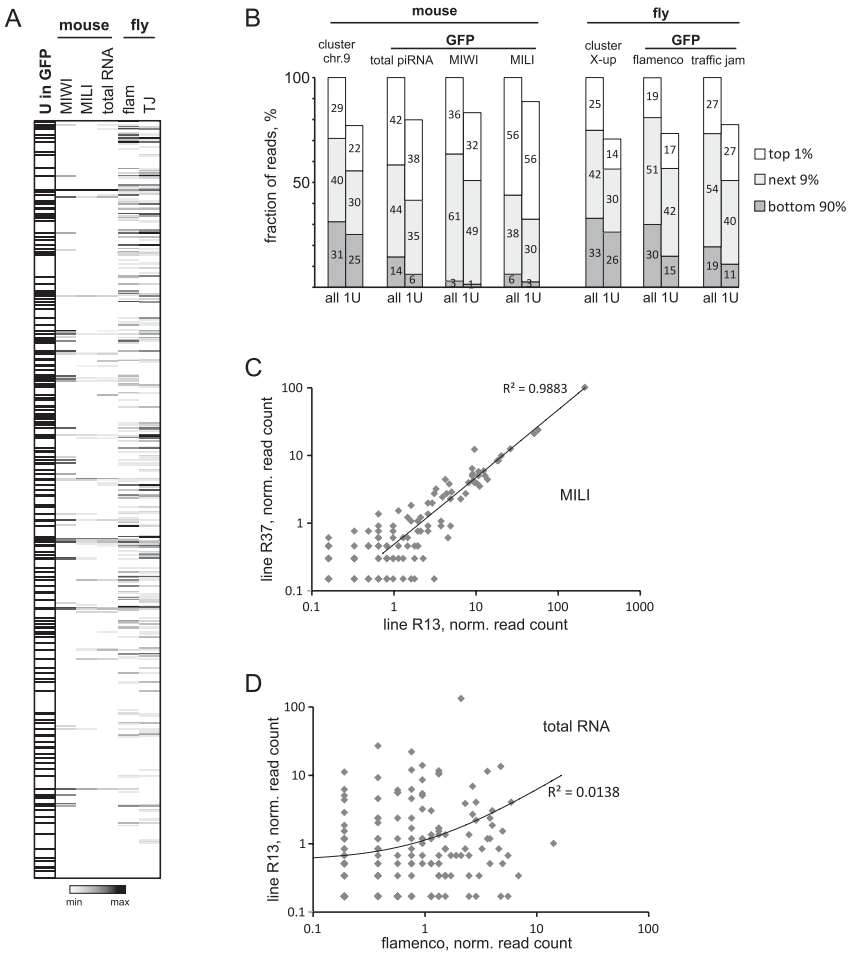


FIGURE 3. apiRNA production is not uniform along inserted sequences. (*A*) A heatmap of piRNA abundance is displayed for all positions in the GFP insert carried in ectopic piRNA clusters as indicated. Sequence measurements were from total RNAs except in mouse, where MIWI and MILI immunoprecipitates (indicated) were also analyzed. The first column simply indicates U positions relative to the heatmaps. (*B*) All possible positions for piRNA production from GFP sequences inserted into ectopic clusters (all sites or only U positions, indicated) were ranked by their contribution to actual piRNA populations. The fraction of piRNAs contributed by the top 1%, the next 9%, or the remaining 90% were measured and indicated. Native clusters (indicated) were similarly analyzed for reference. (*C*) MILI-bound piRNAs were quantified by sequencing from two independent lines carrying the ectopic ch17 cluster. Correlations between read counts for GFP-derived piRNAs are shown. Libraries were normalized as described in the Materials and Methods. (*D*) A similar analysis was performed for GFP-derived piRNAs in total reads, comparing the R13 mouse line carrying the ectopic ch17 cluster and the fly strain carrying the ectopic *flamenco* cluster.

Considered together, these results indicate effects of both local and long-range sequence environment on piRNA biogenesis. Small RNAs generated from GFP embedded in different piRNA precursor transcripts in the same species were more similar than expected by chance. Influences of sequence, however, seem species or cell-type specific, since these same biases did not extend from fly to mouse. Strong effects also appear to be exerted by the context within the cluster, given the near identity in GFP piRNA populations in independent mouse lines and their dissimilarity in comparisons of marked *flamenco* and *tj* transcripts. The precise nature of such context-dependent effects is unclear, but could depend upon the overall secondary or tertiary structures of piRNA precursors.

Our data are consistent with the model in which new insertions of transposable elements become incorporated into the piRNA repertoire as a mechanism of acquiring resistance. Indeed, our data indicate that any sequence will probably produce piRNAs immediately upon its incorporation into a functional piRNA cluster. Furthermore, our data demonstrate that the position of the cluster in the genome is not important, and, therefore, transgenic piRNA clusters can be created in heterologous genomic locations.

Previous bioinformatic analyses described the generation of individual piRNAs from long precursor molecules as a pseudo-random process with a weak influence of the local sequence environment of individual piRNA species (Betel et al. 2007). However, the distribution of individual piRNAs within the precursor is far from being random; different Us have drastically different propensities to generate piRNAs, and some non-U positions produce substantially more piRNAs than nonprocessed U positions. Here, we showed that patterns of individual piRNAs within the precursors are highly reproducible if the sequence is present within the same context. Patterns become less reproducible if the local sequence is embedded in a different context, indicating that both local and long-range sequence environments impact processing efficiency. This result explains a failure in the identification of simple rules that would explain the production of abundant piRNAs from a given precursor molecule.

The general approach we describe here, using marked ectopic piRNA clusters to produce apiRNA species, provides a path toward further dissection of elements that discriminate piRNA clusters and marks corresponding transcripts for piRNA processing. The ability to program the piRNA pathway to produce artificial piRNAs has implications for harnessing this system for controlling gene expression. In particular, in mammals this approach may present advantages over harnessing the miRNA pathway, since piRNAs can induce epigenetic silencing of loci through the recruitment, directly or indirectly, of the de

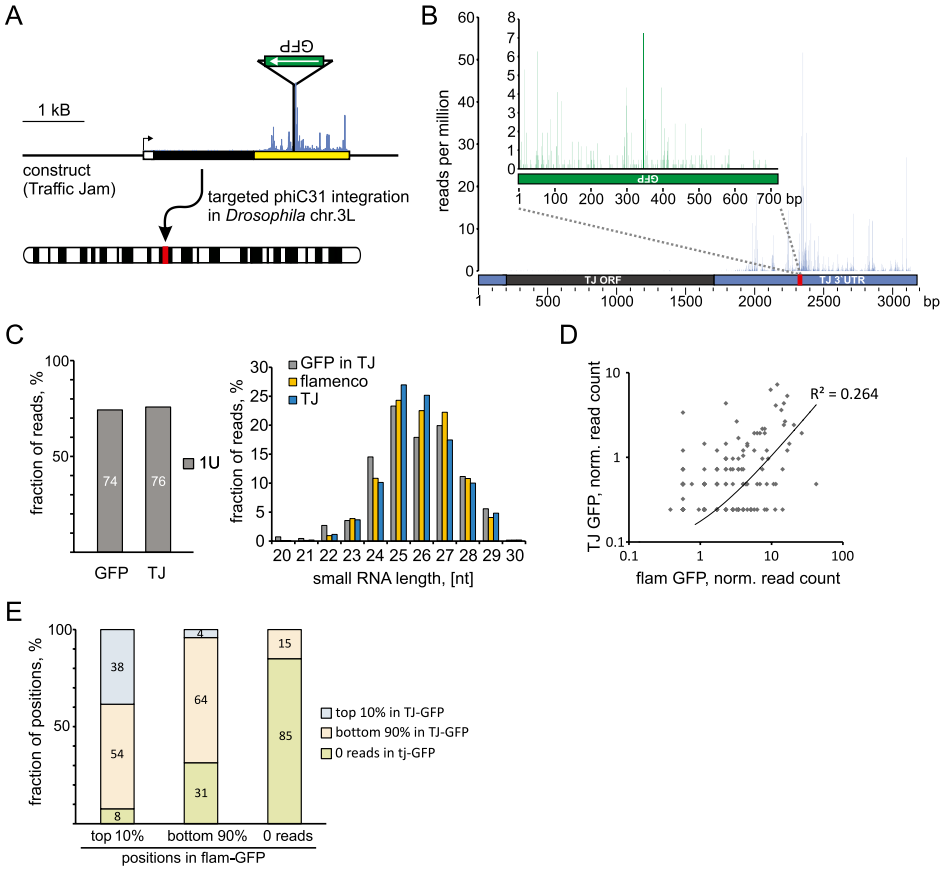


FIGURE 4. apiRNA production from the 3′ UTR of *traffic jam*. (*A*) A schematic of the GFP insertion into the 3′ UTR of the *traffic jam* gene indicates the transcriptional start site (arrow), the coding sequence (black box), and the 3′ UTR (yellow box). *Below*, a diagram indicates site-specific insertion into 3L. (*B*) piRNA read counts are plotted along the inserted GFP sequence (green *inset*) and the surrounding areas of the *tj* 3′ UTR. Note that sequences mapping outside of GFP could be produced from the ectopic insert or the two endogenous copies of *tj*. (*C*) The 1U bias (*left*) and the size distribution of piRNAs mapping to the GFP insert are shown with reference to piRNAs from the *flamenco* cluster. (*D*) Normalized piRNA read counts (see Materials and Methods) were compared for the GFP insertions into the ectopic *flamenco* or *tj* piRNA clusters. (*E*) Read counts are calculated for all possible piRNAs that start with uridine derived from the GFP insertion into *flamenco*. These were divided into the top 10%, the next 90%, and the subset that contributed no reads. For each subset, the number that were present in the top 10%, the next 90%, or the noncontributors for the GFP insertion into *tj* were plotted.

novo DNA methylation machinery (Carmell et al. 2007; Aravin et al. 2008; Kuramochi-Miyagawa et al. 2008; Siomi et al. 2011).

MATERIALS AND METHODS

D. melanogaster strains and crosses

The line, P-1152, which carries an insertion of the P{lArB} construct in telomeric sequences of X chromosome (site 1A) is described in Roche and Rio (1998). To test *trans*-silencing with P-1152, females of this line were crossed with males that have *lacZ* expressed from a euchromatic location on chromosome 2L (line BC69, site 35B10–35C1) (Lemaitre et al. 1993).

Cloning and recombineering—*D. melanogaster*

The *flamenco* transgene was created using P[acman] clone CH321-35A24, which contains an interval from chromosome X that includes ~20 kb of upstream sequence and the 5′ portion of the *flamenco* piRNA cluster (Venken et al. 2009). An antisense EGFP sequence was introduced into the BAC by recombineering as described in Sharan et al. (2009). The GFP-Neo insertion cassette was built by overlapping PCR based on a FRT-PGK-gb2-neo-FRT cassette (Gene Bridges). The position of the insertion within the *flamenco* cluster was selected based on uniqueness and high frequency of piRNA production from the surrounding region. The cassette was introduced into the the BAC using a pSim6 plasmid described in Datta et al. (2006). To promote recombination, *Escherichia coli* containing pSim6 were transferred to a 2-mL Eppendorf tube and induced at 42°C in an Eppendorf tabletop shaker. The linear DNA substrate was introduced by electroporation using the Gene Pulser XCell. Using exponential decay as a pulse-type, the cells were electroporated at 3000 V, 25 μF, and 200 Ω for 5 msec. After outgrowth and selection of cells, recombinant clones were screened for by PCR, sequencing and restriction digestion, followed by pulse-field gel electrophoresis.

The *D. melanogaster traffic jam* gene with 2 kb upstream and 0.5 kb downstream genomic regions was amplified from the CH322-145O22 P[acman] clone and inserted between the BspHI

and ClaI sites of the pIZ-V5-His vector (Invitrogen). A sequence ATTATTCTGATTGCGACAATAAATTCCGAT in the *TJ* 3′ UTR was substituted with the sequence CTTAAGCTGATTGCGACAA TAAATACCGGT by overlap PCR to introduce unique AflII and AgeI sites, which were used to insert the inverted EGFP sequence. The modified *traffic-jam* sequence was transferred into the pCasper5-attB vector (a modified *P-element* pCaSpeR5 vector (Le et al. 2007) with a phiC31 attB site to allow site-specific integration).

Cloning and recombineering—mouse

The transgene containing the modified chr17 piRNA cluster (Chr17: 27427600–27488899) was created using BAC clone RP23-131B16, which contains ~180 kb of genomic sequence that includes the whole chr17 piRNA cluster. We used the FRT-PGK-gb2-neo-FRT cassette (Gene Bridges) and a purified vector containing the EGFP sequence, to construct the GFP-Neo insert for recombineering. After three steps of overlapping PCR (KOD hot start DNA polymerase, Novagen), the recombineering inserts were cloned in a 2.1-TOPO vector (Invitrogen, Version U) according to the manufacturer's protocol. Homology arms for recombineering were added by PCR of purified plasmid.

Recombineering was carried using the Red/ET plasmid-expressing recombination proteins under an arabinose-inducible promoter (Counter-Selection BAC Modification Kit, Gene Bridges, 2007). We followed the manufacturer's protocol, except that recombined clones were selected on Kanamycin and the counter-selection step was skipped. The integrity of modified BAC DNAs were verified by restriction digests and sequencing.

Transgenic animal production—*D. melanogaster*

Tagged BAC DNA was purified with a Plasmid Maxi Kit (QIAGEN). The DNA was used for PhiC31 integrase-mediated transgenesis, which was carried out by BestGene (http://www.thebestgene.com/). *Flamenco* and *tj* transgenes were integrated into attP docking sites on chromosome 3 (VK00031—site 62E1, and VK00033—site 65B2, respectively).

Transgenic animal production—mouse

BAC DNAs were prepared from overnight *E. coli* cultures using Nucleobond BAC 100 columns (Clontech). DNA was eluted in Injection Buffer (10 mM TRIS, 0.1 mM EDTA, 100 mM NaCl, 1X polyamines) and linearized with PI-SceI enzyme for 4 h. Following linearization, BAC DNA was dialyzed overnight on a 25-mm, 0.025-μm filter (Millipore) by floating on Injection Buffer. Transgenic animals were obtained by pro-nuclear injection into B6xSJL F1 hybrids oocytes. Founder animals were crossed to C57BL/6J mice. R37 and R13 transgenic lines were initiated from two independent founder mice.

Immunoprecipitation of PIWI proteins

Immunoprecipitations from *D. melanogaster* ovaries were carried out according to previously described procedures (Brennecke et al. 2007). For mice, MILI and MIWI were immunoprecipitated from adult testis using antibodies and procedures previously described (Aravin et al. 2007b; Vagin et al. 2009). Briefly, testis were dounced in lysis buffer (10 mM Hepes at pH 7.0, 100 mM KCl, 5 mM MgCl2, 0.5% NP-40, 1% triton X-100, 10% Glycerol, 1 mM DTT, proteinase and RNAase inhibitors). Antibodies (MILI-N2 and MIWI-N2) were then added to the cleared lysates and binding reactions were allowed to proceed overnight at 4°C. Protein A beads are then added to the solution and incubated 3–4 h at 4°C with rotation. After three to four washes in NT-2 buffer (5 mM Tris at pH 7.4, 150 mM NaCl2, 1 mM MgCL2, 0.05% NP-40, RNAase inhibitors, 1 mM DTT), antibody complexes were proteinase K treated and RNAs ethanol precipitated following phenol/chlorophorm extraction. A fraction of the precipitated RNAs was radiolabeled and size profiles verified on 15% urea polyacrylamide gels.

Small RNA cloning

Small RNAs from IPs and total RNA extracts were cloned as previously described in Brennecke et al. (2007) and Aravin et al. (2008). Briefly, small RNAs within a 19–33-nt window for mouse samples or a 19–28-nt window for *D. melanogaster* were isolated from 12% polyacrylamide gels. 3′ and 5′ linkers were ligated, and products were reverse transcribed using Superscript III (Invitrogen). Following PCR amplification, libraries were submitted for sequencing using the Illumina GA2x platform.

Detection of β-galactosidase activity in *D. melanogaster* ovaries

Dissected ovaries from 3–5-d-old flies were fixed in freshly prepared 2% glutaraldehyde in PBS for 20 min, washed twice in PBS, and stained for several hours at 37°C in Fe/NaP buffer (3.1 mM $K_3Fe(CN)_6$; 3.1 mM $K_4Fe(CN)_6$; 10 mM $NaH_2PO_4xH_2O$; 0.15 M NaCl; 1 mM $MgCl_2$) with 0.25% X-Gal. Stained ovaries were mounted in 70% glycerol/PBS.

Bioinformatic analysis of small RNA libraries

After FASTQ to FASTA conversion, the Illumina dapter (CTGTAGGCACCATCAATTC) was clipped from the 3′ end of the read and sequences shorter than 16 nt were discarded from further analysis. The remaining sequences were collapsed into a nonredundant list and mapped to the *D. melanogaster* genome (*D. melanogaster* Apr. 2006 [BDGP R5/dm3]) or the mouse genome (mm9) using the short read aligner bowtie (Langmead et al. 2009). Up to two mismatches were allowed. Sequences that failed to map to the genome were mapped against the artificially introduced sequences. The multiplicity count of mapped sequences was normalized to the total number of reads that mapped to the genome. All further bioinformatic analysis on mapping sequences was done using Unix-based text utilities. Details of those scripts can be obtained upon request. Small RNA sequencing data are deposited at GEO, accession number GSE32435.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We thank members of the Hannon and Aravin labs for helpful discussion and comments on the manuscript. We thank members

of the McCombie lab (CSHL) and Igor Antoshechkin (Caltech) for help with RNA sequencing. We thank Andres Canela (CSHL) for technical assistance and Simon Knott (CSHL) and Alex Zahn (Caltech) for help with statistical analysis. Sang Yong Kim (CSHL) created the transgenic mice used in this study. F.M. was supported by the Volkswagen Foundation and B.C. by the Boehringer Ingelheim Fonds. This work was supported by grants from the National Institutes of Health (DP2 OD007371A and R00 HD057233 to A.A.A.; 5R01GM062534 to G.J.H.), by the Ellison Medical Foundation (A.A.A.), and by a kind gift from Kathryn W. Davis (G.J.H.).

Received August 8, 2011; accepted September 26, 2011.

REFERENCES

Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, Morris P, Brownstein MJ, Kuramochi-Miyagawa S, Nakano T, et al. 2006. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442:** 203–207.

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

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007b. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316:** 744–747.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ. 2008. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31:** 785–799.

Betel D, Sheridan R, Marks DS, Sander C. 2007. Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol* **3:** e222. doi: 10.1371/journal.0030222.

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

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322:** 1387–1392.

Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ. 2007. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12:** 503–514.

Copeland NG, Jenkins NA, Court DL. 2001. Recombineering: a powerful new tool for mouse functional genomics. *Nat Rev Genet* **2:** 769–779.

Cox DN, Chao A, Lin H. 2000. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127:** 503–514.

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

Datta S, Costantino N, Court DL. 2006. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379:** 109–115.

Gan H, Lin X, Zhang Z, Zhang W, Liao S, Wang L, Han C. 2011. piRNA profiling during specific stages of mouse spermatogenesis. *RNA* **17:** 1191–1203.

Girard A, Sachidanandam R, Hannon GJ, Carmell MA. 2006. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442:** 199–202.

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC. 2007. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* **315:** 1587–1590.

Haase AD, Fenoglio S, Muerdter F, Guzzardo PM, Czech B, Pappin DJ, Chen C, Gordon A, Hannon GJ. 2010. Probing the initiation and effector phases of the somatic piRNA pathway in *Drosophila*. *Genes Dev* **24:** 2499–2504.

Harris AN, Macdonald PM. 2001. Aubergine encodes a *Drosophila* polar granule component required for pole cell formation and related to eIF2C. *Development* **128:** 2823–2832.

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov DV, Blaser H, Raz E, Moens CB, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129:** 69–82.

Howell MD, Fahlgren N, Chapman EJ, Cumbie JS, Sullivan CM, Givan SA, Kasschau KD, Carrington JC. 2007. Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in *Arabidopsis* reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell* **19:** 926–942.

Huisinga KL, Elgin SC. 2009. Small RNA-directed heterochromatin formation in the context of development: what flies might learn from fission yeast. *Biochim Biophys Acta* **1789:** 3–16.

Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, et al. 2008. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev* **22:** 908–917.

Lahmy S, Bies-Etheve N, Lagrange T. 2010. Plant-specific multisubunit RNA polymerase in gene silencing. *Epigenetics* **5:** 4–8.

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

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363–367.

Lau NC, Robine N, Martin R, Chung WJ, Niki Y, Berezikov E, Lai EC. 2009. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res* **19:** 1776–1785.

Le T, Yu M, Williams B, Goel S, Paul SM, Beitel GJ. 2007. CaSpeR5, a family of *Drosophila* transgenesis and shuttle vectors with improved multiple cloning sites. *Biotechniques* **42:** 164–166.

Lemaitre B, Ronsseray S, Coen D. 1993. Maternal repression of the P element promoter in the germline of *Drosophila melanogaster*: a model for the P cytotype. *Genetics* **135:** 149–160.

Li MA, Alls JD, Avancini RM, Koo K, Godt D. 2003. The large Maf factor Traffic Jam controls gonad morphogenesis in *Drosophila*. *Nat Cell Biol* **5:** 994–1000.

Li C, Vagin VV, Lee S, Xu J, Ma S, Xi H, Seitz H, Horwich MD, Syrzycka M, Honda BM, et al. 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137:** 509–521.

Lim AK, Kai T. 2007. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in *Drosophila melanogaster*. *Proc Natl Acad Sci* **104:** 6714–6719.

Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, et al. 2009. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6:** 474–476.

Malone CD, Hannon GJ. 2009. Small RNAs as guardians of the genome. *Cell* **136:** 656–668.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137:** 522–535.

Olivieri D, Sykora MM, Sachidanandam R, Mechtler K, 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.

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

Pelisson A. 1981. The I–R system of hybrid dysgenesis in *Drosophila melanogaster*: are I factor insertions responsible for the mutator effect of the I–R interaction? *Mol Gen Genet* **183:** 123–129.

Qi H, Watanabe T, Ku HY, Liu N, Zhong M, Lin H. 2011. The Yb body, a major site for Piwi-associated RNA biogenesis and a gateway for Piwi expression and transport to the nucleus in somatic cells. *J Biol Chem* **286:** 3789–3797.

Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC. 2009. A broadly conserved pathway generates 3′ UTR-directed primary piRNAs. *Curr Biol* **19:** 2066–2076.

Roche SE, Rio DC. 1998. *Trans*-silencing by *P* elements inserted in subtelomeric heterochromatin involves the *Drosophila* Polycomb group gene, *Enhancer of zeste*. *Genetics* **149:** 1839–1855.

Ronsseray S, Lehmann M, Anxolabéhère D. 1991. The maternally inherited regulation of *P* elements in *Drosophila melanogaster* can be elicited by two *P* copies at cytological site 1A on the *X* chromosome. *Genetics* **129:** 501–512.

Ronsseray S, Lehmann M, Nouaud D, Anxolabéhère D. 1996. The regulatory properties of autonomous subtelomeric *P* elements are sensitive to a *Suppressor of variegation* in *Drosophila melanogaster*. *Genetics* **143:** 1663–1674.

Ronsseray S, Josse T, Boivin A, Anxolabéhère D. 2003. Telomeric transgenes and *trans*-silencing in *Drosophila*. *Genetica* **117:** 327–335.

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

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

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

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

Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, Siomi H, Siomi MC. 2010. Roles for the Yb body components Armitage and Yb in primary piRNA biogenesis in *Drosophila*. *Genes Dev* **24:** 2493–2498.

Sarot E, Payen-Groschene G, Bucheton A, Pelisson A. 2004. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the *Drosophila melanogaster* flamenco gene. *Genetics* **166:** 1313–1321.

Senti KA, Brennecke J. 2010. The piRNA pathway: a fly's perspective on the guardian of the genome. *Trends Genet* **26:** 499–509.

Sharan SK, Thomason LC, Kuznetsov SG, Court DL. 2009. Recombineering: a homologous recombination-based method of genetic engineering. *Nat Protoc* **4:** 206–223.

Siomi MC, Sato K, Pezic D, Aravin AA. 2011. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* **12:** 246–258.

Todeschini AL, Teysset L, Delmarre V, Ronsseray S. 2010. The epigenetic *trans*-silencing effect in *Drosophila* involves maternally-transmitted small RNAs whose production depends on the piRNA pathway and HP1. *PLoS ONE* **5:** e11032. doi: 10.1371/journal.pone.0011032.

Tomari Y, Du T, Haley B, Schwarz DS, Bennett R, Cook HA, Koppetsch BS, Theurkauf WE, Zamore PD. 2004. RISC assembly defects in the *Drosophila* RNAi mutant armitage. *Cell* **116:** 831–841.

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

Vagin VV, Wohlschlegel J, Qu J, Jonsson Z, Huang X, Chuma S, Girard A, Sachidanandam R, Hannon GJ, Aravin AA. 2009. Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev* **23:** 1749–1762.

Venken KJ, He Y, Hoskins RA, Bellen HJ. 2006. P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. *Science* **314:** 1747–1751.

Venken KJ, Carlson JW, Schulze KL, Pan H, He Y, Spokony R, Wan KH, Koriabine M, de Jong PJ, White KP, et al. 2009. Versatile P[acman] BAC libraries for transgenesis studies in *Drosophila melanogaster*. *Nat Methods* **6:** 431–434.

Wilson C, Pearson RK, Bellen HJ, O'Kane CJ, Grossniklaus U, Gehring WJ. 1989. P-element-mediated enhancer detection: An efficient method for isolating and characterizing developmentally regulated genes in *Drosophila*. *Genes Dev* **3:** 1301–1313.

Zamore PD, Tuschl T, Sharp PA, Bartel DP. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101:** 25–33.

Muerdter et al. Supplementary Figure 1

A

strain P1152;
P{lArB} in TAS

B

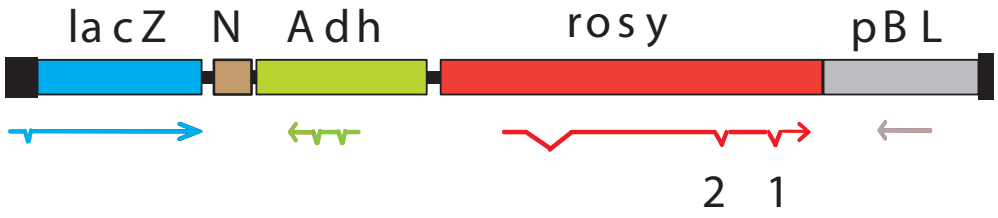
strain BC69;
lacZ transgene
in euchromatin

C

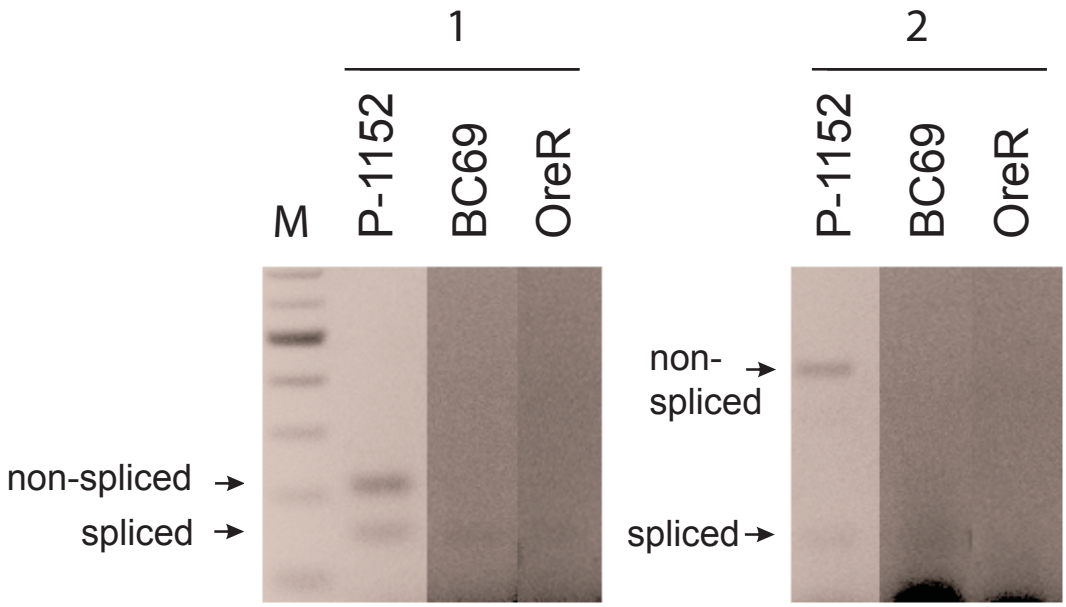
P1152 (female)
x BC69 (male)
(F1)

Muerdter et al. Supplementary Figure 2

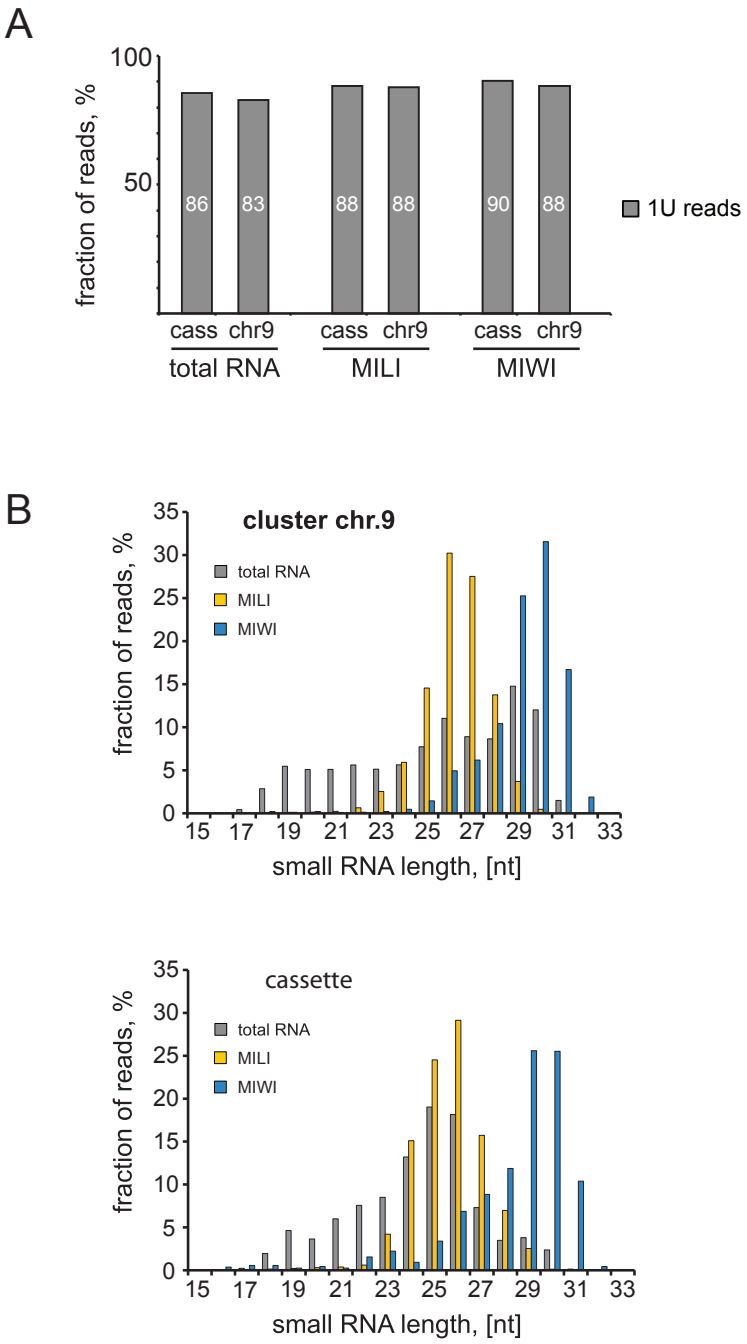
A



B

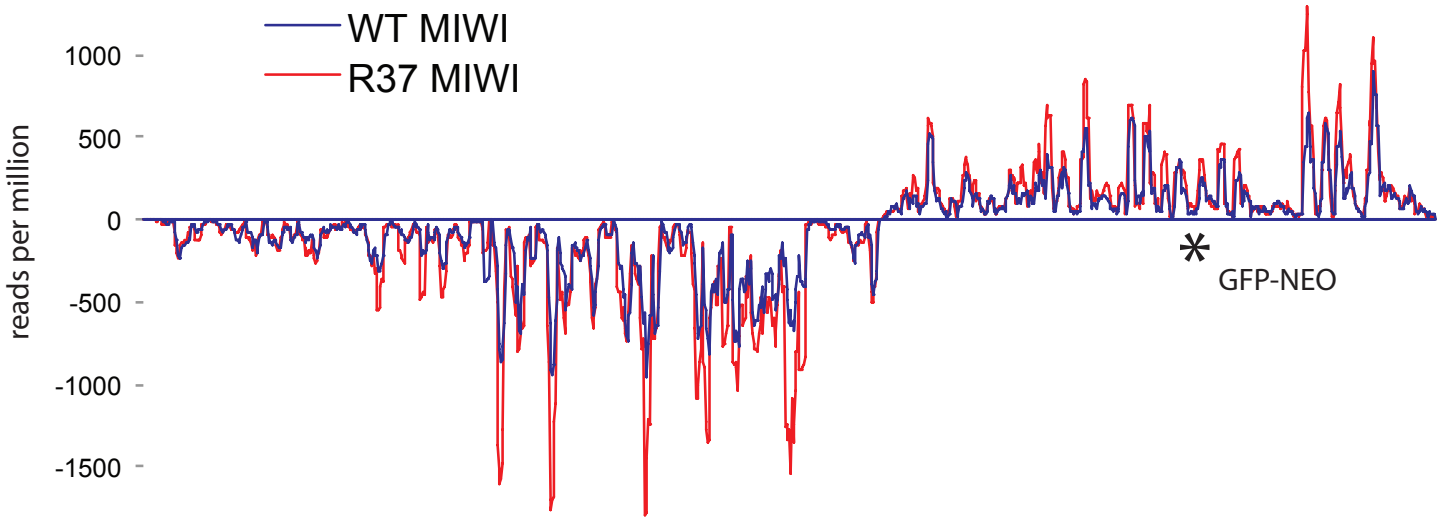


Muerdter et al. Supplementary Figure 3

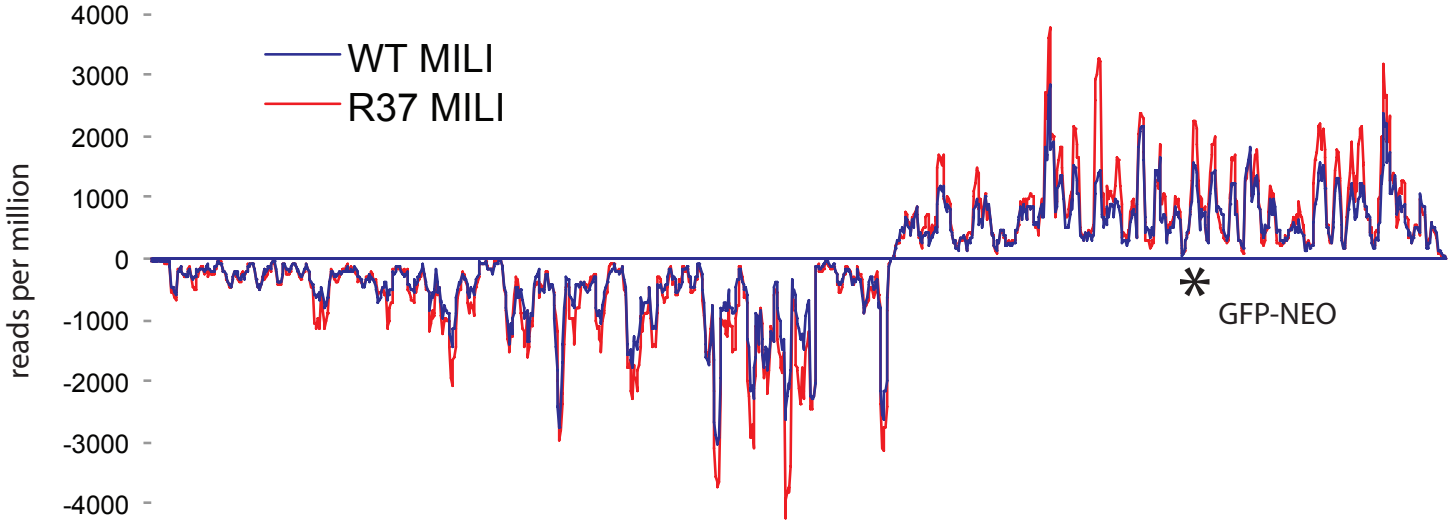


Muerdter et al. Supplementary Figure 4

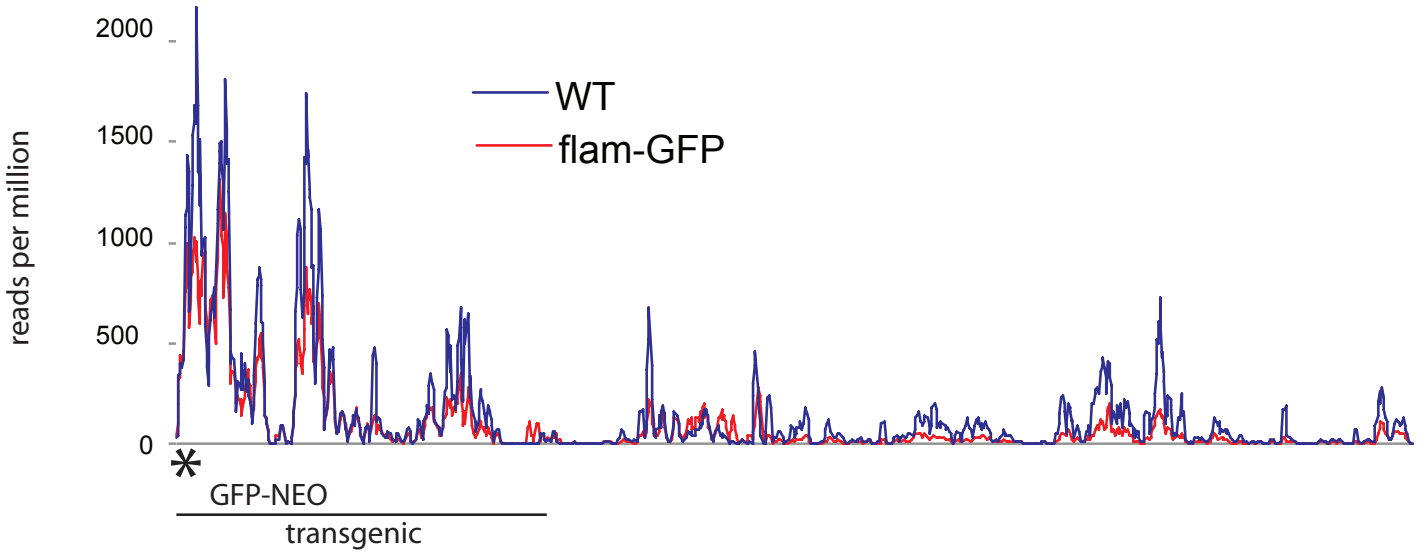
A



B

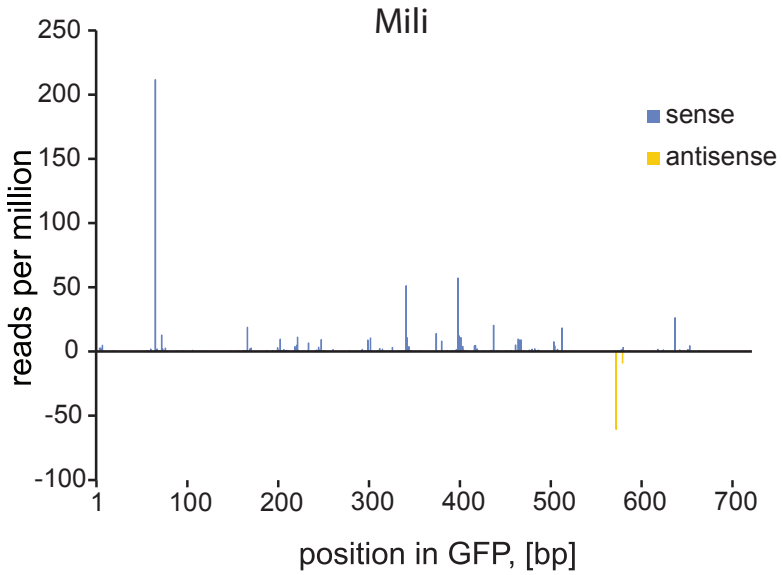


C

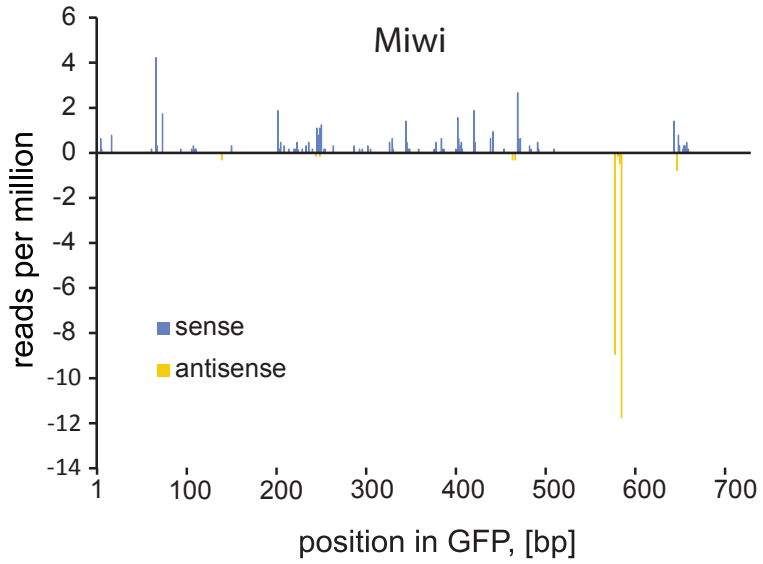


Muerdter et al. Supplementary Figure 5

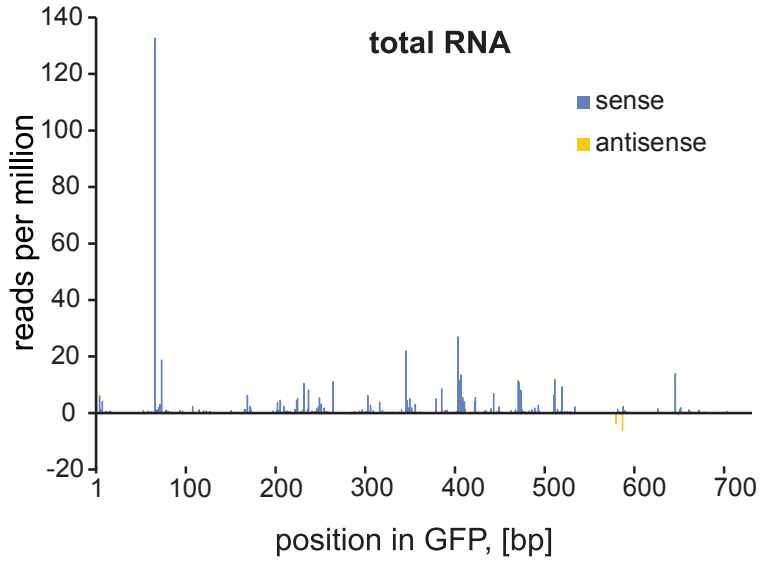
A



B



C



Supplementary Figure Legends

Figure S1. Trans-silencing of lacZ by P{lArB} derived apiRNAs. (A) Ovaries of strain P1152, which carries the P{lArB} insertion in TAS. (B) Ovaries of strain BC69 show lacZ expression from a euchromatic transgene. (C) Trans-silencing of lacZ expression in F1 ovaries of a cross between P1152 females and BC69 males. Note the slightly different levels of repression within different cells of the same ovary.

Figure S2. RT-PCR analysis of *rosy* transcripts. Reverse transcription with primer specific to sense strand of *rosy* transcripts was performed on total RNA from ovaries of strain P-1152 (TAS-inserted transgene), BC69 (same transgene inserted into euchromatin) or Oregon-R (wild type). (A) Position of PCR primers flanking 3rd and 2nd introns of *rosy* transcript. (B) Presence of longer PCR product indicates accumulation of non-spliced *rosy* transcripts in ovaries of P-1152, but not BC69 and Oregon flies.

Figure S3. Features of apiRNAs in mouse. (A) The 1U bias of apiRNAs mapping to the insertion cassette (cass) is compared to native piRNAs from another cluster on chr9. Sequences are derived from total RNA, MILI and MIWI immunoprecipitations (indicated). (B) Size distributions of native piRNAs mapping to a cluster on chr9 (upper panel) compared to apiRNAs mapping to the insertion cassette (lower panel). Sequences are derived from total RNA, MILI and MIWI immunoprecipitations (indicated).

Figure S4. piRNA profiles over wild-type and transgenic piRNA clusters in flies and mice (A) Read densities of piRNAs bound to MIWI are plotted along the cluster on chr 17 on the plus and minus strand (indicated). The site of the GFP cassette insertion is indicated with an asterix. (B) Read densities of piRNAs bound to MILI are plotted along the cluster on chr 17 on the plus and minus strand (indicated). (C) Read densities of piRNAs from total RNA are plotted along *flamenco* on the plus strand. The portion of the cluster contained in the BAC is indicated as 'transgenic'.

Figure S5. apiRNAs in mouse are preferentially bound by MILI. (A) Read counts of apiRNAs bound to MILI are plotted along the inserted GFP sequence on the plus and minus strand (indicated). (B) Read counts of apiRNAs bound to MIWI are plotted along the inserted GFP sequence on the plus and minus strand (indicated). (C) Read counts of apiRNAs from total RNA are plotted along the inserted GFP sequence on the plus and minus strand (indicated).