Piwi-dependent transcriptional silencing and Dicer-2-dependent post-transcriptional silencing limit transposon expression in adult heads of Drosophila Melanogaster

Marius van den Beek

UPMC SORBONNE UNIVERSITÉS

These de Doctorat de l'Université Pierre et Marie Curie

Ecole Doctorale 515 - Complexité du Vivant
Laboratoire Drosophila Genetics and Epigenetics
UMR7622 Biologie du développement
9, quai saint Bernard, 75005 Paris France

Presentée par :

Marius VAN DEN BEEK

Thèse de doctorat de Biologie

Sujet de la thèse :

Piwi-dependent transcriptional silencing and Dicer-2-dependent post-transcriptional silencing limit transposon expression in adult heads of *Drosophila melanogaster*

Présentée et soutenue publiquement le 09 fevrier 2015 à Paris

Devant le jury composé de :

Pr. DEVAUX Frederic	Président
Dr. SEITZ Hervé	Rapporteur
Dr. COLOT Vincent	Rapporteur
Dr. QUESNESVILLES Hadi	Examinateur
Pr. VAN RIJ Ronald	Examinateur
Dr. PELISSON Alain	Examinateur
Dr. ANTONIEWSKI Christophe	Directeur de thèse

Piwi-dependent transcriptional silencing and Dicer-2-dependent post-transcriptional silencing limit transposon expression in adult heads of *Drosophila melanogaster*

Acknowledgements

Summary

The first small non-coding RNAs have been cloned in 1993, more than 20 years ago, but the study of their impact on diverse biological processes such as development, cellular differentiation, immune functions and fine-tuning of gene networks has only recently picked up pace with the introduction of next generation high-throughput sequencing.

We now know that there are at least 3 major classes of non-coding small RNAs in animals: the microRNAs (miRNAs), small interfering RNAs (siRNAs) and Piwi-interacting RNAs (piRNAs). Small RNAs are generally involved in negative regulation of complementary target RNA abundance, but each class has distinct mechanisms of biogenesis, target recognition, mechanisms of target regulation.

Transposable elements are major components of eukaryotic genomes and make up approximately 12% of the total *Drosophila* genome sequence, 45% of the human genome and more than 85% of the maize genome. Transposable elements have been proposed as important drivers of gene network evolution, as they can move or "transpose" in their host genome, creating gene duplications, cause gene inactivation or alter gene function. Nevertheless, uncontrolled high-rate transposition leads to DNA damage and genomic instabilities, a signature often found in tumorous tissues, and is hence counteracted by multiple mechanisms, amongst which the generation of piRNAs and siRNAs that survey the expression of transposable element transcripts.

To understand the role of piRNAs and siRNAs in the control of somatic transposons I have sequenced small RNA of wild type fly adult heads and compared these to the heads of *piwi* mutants. I found an increase of siRNAs against transposable elements, a remarkable finding considering the absence of Piwi in heads, suggesting an epigenetic effect of *piwi* mutation on the repertoire of small RNAs in the *Drosophila* head. RNA-sequencing of *piwi* mutant heads then revealed only minor changes of transposon expression, similar to the results we obtained with *dicer-2* mutants. It was only in double mutants of *piwi* and *dicer-2* that transposon levels increased significantly, leading me to suggest a model in which both piRNA and siRNA represent distinct and complementary layers of transposon repression in adult heads of *Drosophila melanogaster*. piRNAs establish transcriptional gene silencing during early development, and siRNAs may act later in the adult tissue through post-transcriptional gene silencing. This may provide a rescue mechanism in case of failure to establish Transcriptional Gene Silencing (TGS) during early development or failure to maintain TGS. Importantly, we also observed a decreased lifespan of double mutants

compared to *piwi* or *dicer-2* mutants, possibly due to an increased activity of transposons or as of yet unidentified effects on protein-coding gene expression.
These results constitute a significant advance in understanding how transposons are repressed in somatic tissues. Since piwi's main function in the repression of transposons is likely epigenetic in nature, it might be possible that age-dependent reduction of heterochromatin decreases the Piwi-mediated level of repression. Transposon-specific siRNAs then provide another layer of repression.

My results might justify the investigation of Piwi-mediated repression of transposons in somatic tissues of mammalian organisms, as age-dependent diseases might be caused by an increased transposon burden.

My work involved for a large part NGS analysis that I performed mostly within the Galaxy framework (http://usegalaxy.org). At the heart of Galaxy is a database that keeps track of data and data transactions, a web server, an architecture that allows the easy plugging of command-line tools into a user-friendly web-interface and a workflow engine that allows the execution of complex user-defined step-wise analyses. As the origin and treatment of all datasets is automatically traced, reproducibility of bioinformatic analyses is greatly facilitated, especially in the light of many papers involving complex genome-wide analyses that are difficult to reproduce partly due to undisclosed or incomplete method descriptions. A second important aspect of the Galaxy framework is the ease of use of bioinformatic utilities even for biologists that received no or little training on using the command line. Finally, Galaxy allows easy sharing of results with colleagues, thereby accelerating collaborative efforts.

Together with my supervisor, Christophe Antoniewski, I set up a Galaxy server during my first year of thesis (http://lbcd41.snv.jussieu.fr/galaxy) that is heavily used by us and our collaborators.
As this server was soon under heavy load we set up a second server, this time publicly available (http://mississippi.fr). Both servers are dedicated to small RNA and transcriptomic analysis, with many tools written by Christophe Antoniewski and myself. We will soon prepare a manuscript for the mississippi toolsuite, that is already released in the public testtoolshed of the Galaxy community (https://testtoolshed.g2.bx.psu.edu/).

A current limitation of the Galaxy framework is that users are not completely free in the type of analysis that can be performed on a given Galaxy server. If the necessary tool is not installed, users must find another way to treat their data, negating the advantage Galaxy

provides for traceability and reproducibility of results. Similarly, power-users that regularly write their own analyses are forced to become Galaxy administrators and learn how to write Galaxy tools if they want to use their scripts and programs inside Galaxy.

Lazarus and colleagues developed the Galaxy Tool Factory (Lazarus et al., 2012), which allows the execution of arbitrary scripts. However, the Galaxy Tool factory was limited to single input files, and most importantly, because any script can be executed, it is very insecure and should only be run by a Galaxy administrator.

To circumvent these limitations, I modified this tool factory in a way that a secure and transient sandbox, based on Docker (https://www.docker.com/) containers, is created, to which user data is mounted. User scripts are executed inside the sandbox and cannot harm data outside of the sandbox. Once functional, scripts may be used in workflows, or simple tools can be generated, helping Galaxy administrators in deploying new, custom tools and opening the opportunity to use Galaxy as an Integrated Development Environment (IDE). This work is still in development and led to a manuscript draft for the DockerToolFactory, which will be sent for publication shortly. The tool can be downloaded from the Galaxy toolshed, and the sourcecode is also available at https://bitbucket.org/mvdbeek.

Introduction I: small RNA silencing

Small RNAs are a relatively recent discovery for biologists, but not for evolution as small RNA expression is very widespread. Generally speaking, small RNAs are involved in negatively regulating RNA abundance. They play a fundamental role in regulating gene expression and discriminating parasitic from non-parasitic RNAs.

Drosophila melanogaster, the model organism that I have been studying during my PhD work, expresses 3 classes of small RNAs, miRNA, siRNAs and piRNAs. These are genetically well-separated, making it possible to study only certain aspects of a single pathway. The goal of my thesis was to understand how the piRNA pathway and the siRNA pathway limit transposon expression in adult tissues. To aid the reader in understanding and placing my work into the current scientific background I will first describe these 3 small RNA pathways and how they might interact.

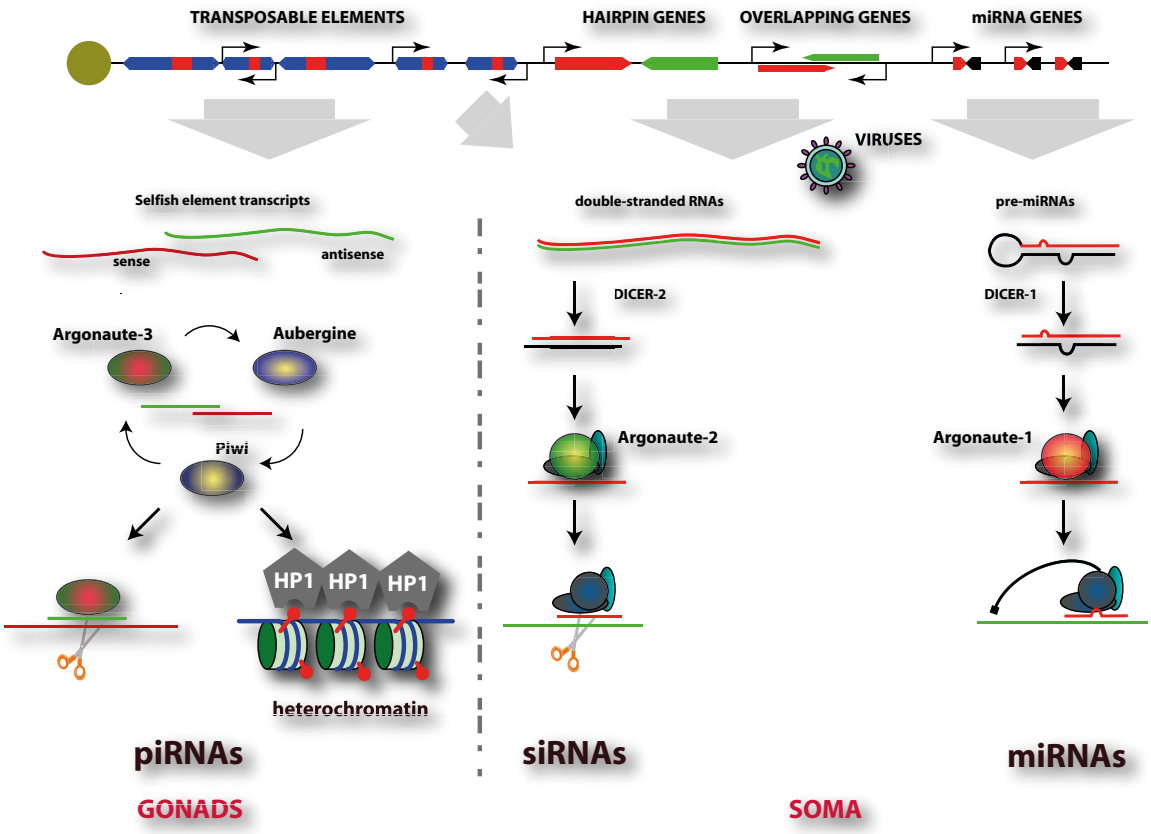


Figure 1: Drosophila small RNA pathways are genetically well-seperated. piRNA, siRNA and miRNA pathways utilize distinct protein machineries and differ in their target molecules

1. miRNA silencing

a. Discovery of *lin-4* and miRNAs

The very first miRNA locus, *lin-4*, was cloned by (R. C. Lee, Feinbaum, and Ambros 1993), while studying factors that regulate the timing of LIN-14 Protein expression in *Caenorhabditis elegans*. It was known that the LIN-14 protein was present in late-stage embryos and L1 larvae, but barely detectable in L2 larvae, while *lin-14* transcripts were constant throughout development. A *lin-14* 3' UTR mutant mirrored the failure to down-regulate LIN-14 protein seen in *lin-4* mutants. Through cloning of the *lin4* locus and genetic rescue using a small fragment of the lin4-containing intron Lee at al. ruled out the possibility that *lin-4* was a protein. Instead they identified two 22 and 61nt transcripts originating from the *lin-4* locus and predicted that they might form the now well-known miRNA stem-loop structure. They also speculated that the 22nt *lin-4* might partially base pair with repeated sequences of the *lin14* 3'UTR to inhibit translation by hindering interaction between 3' and 5' ends of *lin-14* mRNA. The base-pairing hypothesis was later confirmed by Ha and colleagues (Ha, Wightman, and Ruvkun 1996), and inhibition of translation elongation through *lin-4* RNA paired to the 3'UTR of LIN-14 was confirmed by Olsen and Ambros (Olsen and Ambros 1999). The discovery of another miRNA, let-7 (Pasquinelli et al. 2000), allowed the recognition that miRNAs also exist in vertebrate, ascidian, hemichordate, mollusc, annelid and arthropod (Pasquinelli et al. 2000).

b. miRNA biogenesis and function

As most protein-coding genes, miRNA loci (with their own promoter or as introns inside another gene) are transcribed by RNA-polymerase II (PolII) into primary miRNA transcripts (Y. Lee et al. 2004) (see Figure 2a for a graphical miRNA biogenesis summary). These transcripts contain one or more stem-loop structures, which are recognized and cleaved at the base of the loop inside the nucleus by the Drosha and Pasha-containing microprocessor complex (Denli et al. 2004) into shorter 60-70 pre-miRNAs. Intronic miRNAs are processed co-transcriptionally before splicing (Figure 2b).

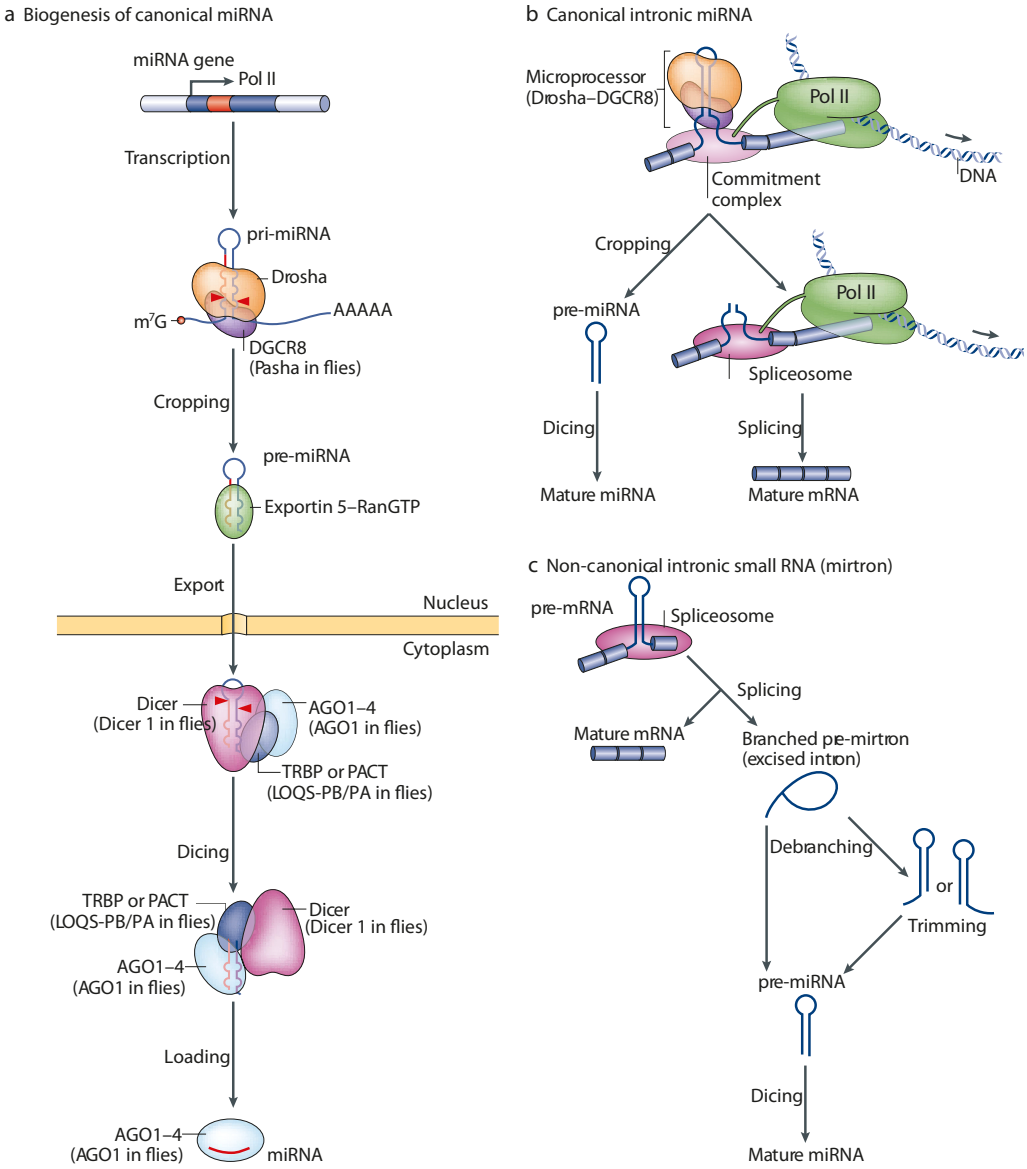


Figure 2. miRNA biogenesis pathway. (a) Canonical microRNA (miRNA) genes are transcribed by RNA polymerase II (Pol II) to generate the primary transcripts (pri-miRNAs). The initiation step (cropping) is mediated by the Drosha–DiGeorge syndrome critical region gene 8 (DGCR8; Pasha in Drosophila melanogaster and Caenorhabditis elegans) complex (also known as the Microprocessor complex) that generates ~65 nucleotide (nt) pre-miRNAs. Pre-miRNA has a short stem plus a ~2-nt 3′ overhang, which is recognized by the nuclear export factor exportin 5 (EXP5). On export from the nucleus, the cytoplasmic RNase III Dicer catalyses the second processing (dicing) step to produce miRNA duplexes. Dicer, TRBP (TAR RNA-binding protein; also known as TARBP2) or PACT (also known as PRKRA), and Argonaute (AGO)1–4 (also known as EIF2C1–4) mediate the processing of pre-miRNA and the assembly of the RISC (RNA-induced silencing complex) in humans. One strand of the duplex remains on the Ago protein as the mature miRNA, whereas the other strand is degraded. Ago is thought to be associated with Dicer in the dicing step as well as in the RISC assembly step. In D. melanogaster, Dicer 1, Loquacious (LOQS; also known as R3D1) and AGO1 are responsible for the same process. In flies, most miRNAs are loaded onto AGO1, whereas miRNAs from highly base-paired precursors are sorted into AGO2. The figure shows the mammalian processing pathways with fly components in brackets. (b) Canonical intronic miRNAs are processed co-transcriptionally before splicing. The miRNA-containing introns are spliced more slowly than the adjacent introns for unknown reasons. The splicing commitment complex is thought to tether the introns while Drosha cleaves the miRNA hairpin. The pre-miRNA enters the miRNA pathway, whereas the rest of the transcript undergoes pre-mRNA splicing and produces mature mRNA for protein synthesis. (c) Non-canonical intronic small RNAs are produced from spliced introns and debranching. Because such small RNAs (called mirtrons) can derive from small introns that resemble pre-miRNAs, they bypass the Drosha-processing step. Some introns have tails at either the 5′ end or 3′ end, so they need to be trimmed before pre-miRNA export. m7G, 7-methylguanosine. Adapted from (Narry Kim et al., 2009).

Alternatively, some pre-miRNAs are defined directly by the splicing machinery (Figure 2c, mirtron), and do not depend on processing by Drosha (Ruby, Jan, and Bartel 2007). pre-miRNAs are exported through exportin-5 into the cytoplasm (Yi et al. 2003) where they are further processed into mature, 22nt miRNAs by Dicer (Y. Lee et al. 2002) (Dicer-1 in *Drosophila, (Y. S. Lee et al. 2004)*) in conjunction with the Loquacious (Loqs) isoforms PA or PB (Fukunaga et al. 2012; Förstemann et al. 2005).

The miRNA duplex is then loaded by an ATP-dependent process into Ago1-containing ribonucleoprotein complexes (also called RNA-induced silencing complexes, RISCs), where the single stranded, mature miRNA is selected from the miRNA duplex on the basis of the thermodynamically less stable 5' end base-pairing (Kawamata, Seitz, and Tomari 2009). The mature miRNA serves as a guide to scan the cellular mRNA pool.

Once a mRNA with complementarity to the miRNA (the nt 2-8 seed region relative to the 5' of the miRNA is most important) is found, the mRNA is retained in the RNA induced silencing complex (RISC), causing translation to halt (Chendrimada et al. 2007) or destabilization of the mRNA (Wu, Fan, and Belasco 2006). miRNA-mediated halt of translation and mRNA-destabilization both require binding of GW182 protein family members to Ago1 (Eulalio, Huntzinger, and Izaurralde 2008). miRNAs are thus post-transcriptional regulators of gene expression. They function in a multitude of processes and are required for the development of *Drosophila*, as evidenced by early lethality in loss-of-function (LOF) alleles of the core miRNA components (Kataoka, Takeichi, and Uemura 2001; Y. S. Lee et al. 2004). A difficulty for identifying the targets of a miRNA, and thus their biological functions, lies in the fact that miRNAs pair imperfectly with their targets, and only short stretches of perfect complementarity in the 5' seed (nt 2-8) seem to be requirements for target repression by a microRNA, that can only partially be compensated for by additional 3' complementarity (Brennecke et al. 2005). Predicting miRNA targets from the seed sequence alone results in a large number of false positive predictions. By classifying and limiting the potential targets to matches in the 3' UTR, scoring additional matches in the 3' supplementary regions, accounting for multiple repeated target sites in the 3'UTR and additional experiments such as CLIP-seq, the False Detection Rate can be improved but final proof for target repression by miRNAs still requires experimental assessment (Hausser and Zavolan 2014).

2. piwi interacting RNA (piRNA) silencing

a. *Drosophila*-centric history of the piRNA pathway

Even though retrospectively-linked piRNA-related phenomena such as co-suppression of endogenous genes by increasing copy-numbers of transgenes had been described in plants as early as 1990 (Lemieux, Jorgensen, and Napoli 1990), and co-suppression was found to occur in *Drosophila* in 1997 (M Pal-Bhadra, Bhadra, and Birchler 1997), it was only in 2001 that endogenous ~25nt piRNAs from the Su(Ste) repeat locus were described by Aravin and Colleagues (Aravin et al. 2001). In 2003, a large-scale effort to systematically clone small RNAs during *Drosophila* development by Aravin and colleagues led to the identification of 178 repeat associated small interfering RNAs (rasiRNA), originating from transposable elements, microsatellites and the 42AB locus (Aravin et al. 2003).

RasiRNA abundance was strongest in early embryos and testes, suggestive of a biogenesis involving the germline-specific Argonaute proteins Ago3, Piwi and Aubergine.

In 2006, Vagin and Sigova et al (Vagin et al. 2006) showed that rasiRNA production was independent of Dicer-1 and Dicer-2 and that rasiRNAs were resistant to beta-elimination and bind to Piwi and Aubergine *in-vivo*.

Saito et al. also showed that Piwi does not bind miRNAs, and that siRNAs are not loaded inside Piwi in lysates of fly ovaries, but that Piwi pre-loaded with a synthetic siRNA contains target cleavage activity *in-vitro* (Saito et al. 2006).

In 2007, the Siomi and Hannon Laboratories proposed a mechanism for the formation of the 5'end of rasiRNA, hereafter called piRNAs and their role in limiting transposition. Through deep-sequencing of the Aubergine, Piwi, and Ago3-bound small RNAs, both teams independently found a preference for piRNAs with antisense complementarity to transposable element transcripts to be bound to Aubergine and Piwi, whereas Ago3-bound piRNAs in sense orientation showed a strong tendency to overlap by 10-nt from their 5' end with complementary Aubergine or Piwi bound piRNAs. The model postulates that antisense piRNA transcripts, originating from genomic loci enriched for defective transposable element fragments, provide antisense piRNA precursor transcripts that after an initial processing step (detailed in the "piRNA precursor processing" chapter) guide Aubergine or Piwi to a sense transposable element transcript, which is cleaved between the 10^{th} and 11^{th} nucleotide and becomes a new piRNA that can be loaded in Ago3 to cleave a new antisense transcript,

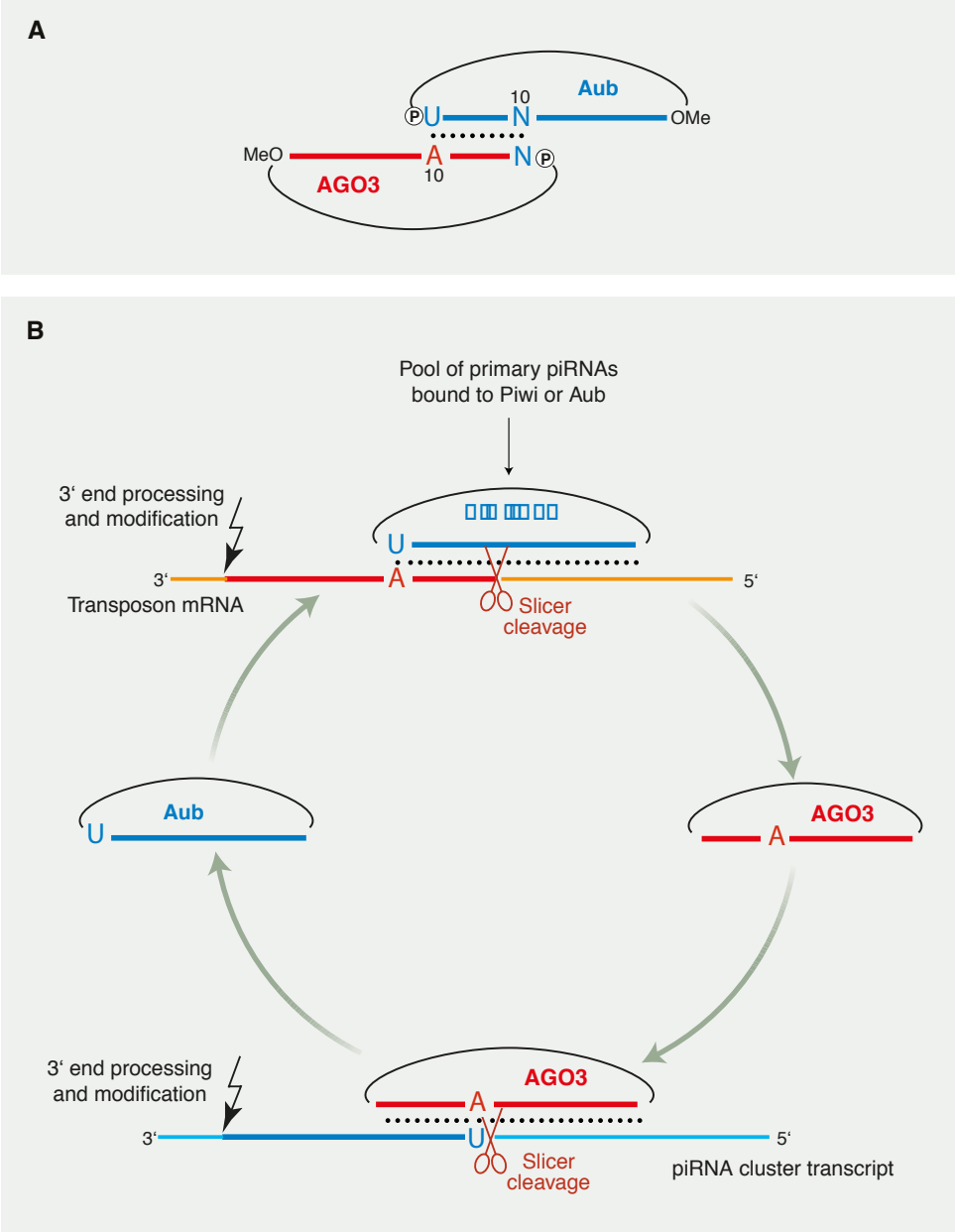


Figure 3: Secondary ping-pong piRNA properties and biogenesis mechanism. (A) Features of Aubergine and AGO3-associated piRNAs in Drosophila. Indicated are the 5′ U bias in Aub-bound piRNAs, the 10A bias in AGO3-bound piRNAs, the 5′ phosphate, and the 3′ O-methylation. (B) Ping-Pong model of piRNA biogenesis in Drosophila. Primary piRNAs are generated by an unknown mechanism and/or are maternally deposited. Those with a target are specifically amplified via a Slicer-dependent loop involving AGO3 and Aub. From (Aravin et al., 2007)

thereby forming an amplification loop that effectively provides a surplus of antisense piRNAs that can degrade transposon transcripts in a post-transcriptional manner (Aravin, Hannon, and Brennecke 2007; Brennecke et al. 2007; Gunawardane et al. 2007). The biogenesis mechanism is also referred to as the “ping-pong” or secondary piRNA biogenesis (also see Figure 3).

These observations served as a central framework for understanding the function and mode of action of the piRNA pathway in the *Drosophila* germline. It was not clear yet which other

genetic factors were important in primary piRNA biogenesis, how piRNA clusters work and what the relation of piRNA-mediated post-transcriptional silencing was with the previously described co-suppression effects that were shown to be regulated at the transcriptional level.

b. Screening efforts provided an overview of piRNA pathway genes

Knowledge from multiple, previously unrelated fields helped to define additional actors of the piRNA pathway (summarized in the introduction to Olivieri et al., 2010, (Olivieri et al. 2010)). Defects in piRNA biogenesis and function result in male and female sterility, transposon upregulation and egg patterning defects. Many genes and loci that cause these phenotypic defects turned out to be factors affecting piRNA biogenesis or function. Therefore mining of genes with similar phenotypes and (re-)characterizing mutant collections from genetic screens for these phenotypes led to the identification of many more genes involved in piRNA biology.

Especially the de-repression of transposable elements proved to be a phenotype that is easy to follow in genome-wide RNAi knockdown screens and that led to the identification of candidate piRNA genes with good sensitivity and specificity. A number of reporter constructs were used either in the discovery or validation phase of these screens. The idea behind these reporter constructs is that when transcribed, the reporter RNA contains a piece of a transposable element sequence fused to a reporter gene (lacZ or NLS-lacZ fusion are most frequently used). The transposable element sequence is then subject to piRNA mediated silencing, and reporter gene expression is only detectable when the piRNA pathway is impaired and cannot target the reporter through homology with the transposable element sequence.

Notably, the first screen (Olivieri et al. 2010; Handler et al. 2013) was performed in the Brennecke lab using the gypsy-lacZ reporter (Sarot et al. 2004), whose promoter is the 5' LTR of the *gypsy retrotransposon*, and as such recapitulates the expression pattern of endogenous *gypsy* (see Figure 4). As *gypsy* is expressed most strongly in the somatic follicular cells of the ovaries, the screen aimed at identifying factors that affect the primary piRNA pathway that operates in the somatic follicular cells. In these cells only piwi-loaded piRNAs that do not show the 10nt overlap signature (ping-pong signature) are present.

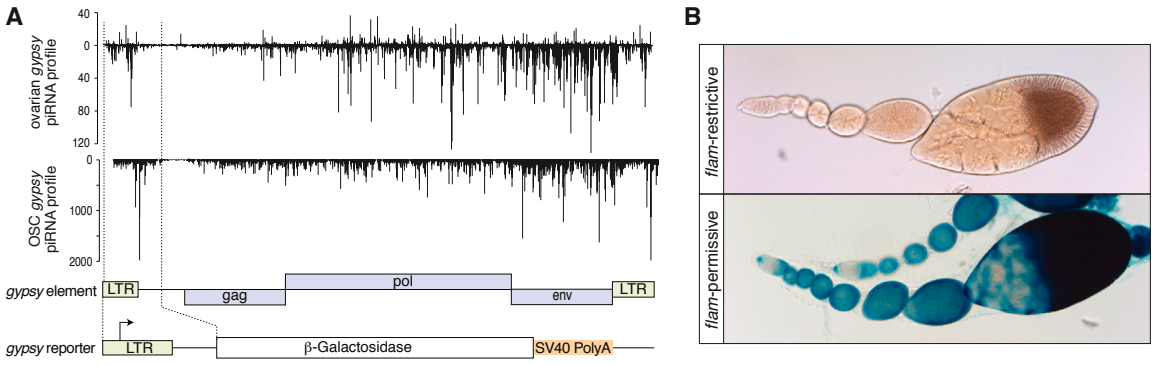


Figure 4: *gypsy*-lacZ sensor is repressed by piRNAs in OSC cells and ovaries. (A) piRNA profiles against the canonical *gypsy* element are shown for wild type (upper panel) ovaries and OSC cells (middle panel). For comparison, the *gypsy* LTR portion of the lacZ sensor is indicated (lower panel). (B) beta-galactosidase stainings of the *gypsy*-lacZ reporter in *flamenco*-restrictive and flamenco permissive background. *flamenco*-permissive strains shown strongly reduced levels of *gypsy* piRNAs.

The screen was conducted using the traffic-jam Gal4 (Tj-Gal4) driver combined with the gypsy-lacZ reporter crossed to available UAS-RNAi lines for genes expressed in the somatic ovarian

cells. The screen has been estimated to have identified about 80 % of the protein-coding genes that act in the somatic primary piRNA pathway while factors exclusively involved in the germline cells would be missed by this screen (Handler et al. 2013).

A second screen targeted specifically germline piRNA factors, by knocking down germline expressed genes using UAS-RNAi lines driven by nanos-gal4 (nos-Gal4), which is expressed in the nurse cells and the oocyte. The readout of the screen was a multiplex qPCR assay for *Het-A*, *TAHRE*, *blood* and *burdock* transposon expression. A set of 74 genes was found in this screen, among which 16 out of 17 known piRNA pathway genes (Czech et al. 2013).

In parallel, a third screen was conducted in OSC cells (Ovarian Somatic Sheet Cells). OSC cells recapitulate the primary piRNA pathway in the ovarian somatic sheet cells (Lau et al. 2009; Robine et al. 2009). Long double stranded RNA (dsRNA) that silence candidate genes were transfected into OSC cells and de-repression of endogenous *gypsy* was followed by qRT-PCR on the spliced subgenomic *gypsy* RNA (Muerdter et al. 2013). As expected, many candidate hits overlapped the screen performed using the *gypsy-lacZ* sensor.

None of these genetic screens can identify genes that are in the same time involved in the piRNA pathway and essential for cell viability, and one can wonder whether genes required in both RNAi and piRNA pathways, should they exist, be identified in these screens, as a decreased knockdown efficiency might lower the chance to observe a de-repression. Still, it seems reasonable to assume that the majority of factors specific to the ovarian piRNA pathway have been found.

c. The "life-cycle" of piRNAs

By integrating data from these screens, together with data from protein-protein interaction studies, subcellular localization studies, transcriptome and chromatin profiling, a model is emerging for each of the phases of the ovarian "piRNA life-cycle".

Multiple partly overlapping phases in piRNA biogenesis and function can be defined:

1. The "initiator phase": This includes the transcription of piRNA precursor molecules, its nuclear export and processing into a mature primary piRNA.
2. The "amplification phase": In this cytoplasmic phase antisense piRNA populations are amplified in the previously described "ping-pong" cycle.
3. The "post-transcriptional effector phase": This includes the post-transcriptional regulation of transposable elements in the cytoplasm, but also of genic targets.
4. A "transcriptional effector phase": During this phase piRNA-loaded *Piwi*-complexes enter the nucleus to guide transcriptional silencing of TEs.
5. A "feedback phase": In addition to their negative transcriptional effect on TE insertions, piRNA are likely involved in a feedforward regulatory loop that triggers or enhances transcription of piRNA producing loci.

.

When aligning piRNAs to the genome, distinct clusters with an increased density of aligned piRNAs are found, and these regions also contain a high frequency of defective transposable element pieces.

Through analysis of the small RNA content of ovaries, that contain a mix of somatic and germline tissues, early embryos, which essentially reflect the germline piRNA content, and small RNAs in the OSC cell line, which reflect the small RNA content of the somatic cells in the ovary, it has become clear that activity and biogenesis of germline piRNAs and clusters are distinct from their somatic counterparts.

While the majority of germline piRNA clusters is bidirectionally transcribed and feeds into the "ping-pong" amplification cycle, somatic piRNA clusters are transcribed from only one strand and exclusively produce piRNAs that do not show a ping-pong pattern of biogenesis.

In the following I will describe the initiator phase for somatic piRNAs and highlight differences with germline piRNA initiation.

d. Somatic piRNA precursor transcription

The prototypical somatic piRNA cluster *flamenco*, (Malone et al. 2009; Brennecke et al. 2007) had been shown to control expression and mobilization of the *gypsy* retrotransposon (Pelisson et al. 1994; Prud'homme et al. 1995). The *flamenco* locus contains many TE sequences (Robert et al. 2001) in antisense orientation with respect to its transcript (Mével-Ninio et al. 2007; Brennecke et al. 2007). *flamenco* is expressed in the ovarian somatic follicle cells, as piRNAs produced from *flamenco* are present in a cell line derived from ovarian somatic sheet cells (OSS, (Lau et al. 2009)), but *not* in early embryos (Malone et al. 2009).

flamenco transcription by RNA Polymerase II (Pol II) requires the Cubitus Interruptus transcription factor and is followed by splicing (Goriaux et al. 2014) and export to the cytoplasm, which probably involves Nup54, Nup58, CG11092, Nup214, that were identified to be required for *gypsy-lacZ* repression (Handler et al. 2013). The splicing of *flamenco* is remarkable, as one of the requirements for piRNA production from germline dual-strand clusters is the repression of splicing discussed below, but in-line with a large number of splicing factors identified in the gypsy-lacZ somatic screen (Handler et al. 2013).

e. Germline piRNA precursor transcription

In contrast to somatic piRNA clusters, bidirectional germline clusters such as the model cluster 42AB are silent in OSS cells (Lau et al. 2009), and their proper processing in ovaries depends on a heterochromatic environment, which, strikingly, does not prevent transcription (Z. Zhang et al. 2014). This is accomplished by the germline specific proteins Eggless (a Histone Methyl Transferase (HMT) specific for H3K9 methylation) (Rangan et al. 2011), Rhino (a HP1 homolog), Deadlock, Cutoff (the last 3 proteins colocalize and have been termed the "RDC complex" (Z. Zhang et al. 2014; Thomas et al. 2014; Mohn et al. 2014), and UAP56 (F. Zhang et al. 2012). The RDC likely functions to repress splicing and transcription termination that would otherwise be induced by splice and transcription termination sites sites in the TE fragment sequences. Mutations in *cuff*, *del*, *uap56* and *rhino* induce canonical, euchromatic splicing patterns and strongly reduce mature piRNA levels from germline clusters (Z. Zhang et al. 2014; Thomas et al. 2014; Mohn et al. 2014).

The current model states that the RDC complex defines germline-specific dual-strand piRNA cluster transcripts. Cluster transcripts remain bound by UAP56, are exported through NUP154 and are handed over to the processing machinery in the nuage (see next section).

Mutations in UAP56 and Rhino strongly reduce piRNA levels from dual-strand clusters and ablate the ping-pong signature, but a significant amount of piRNA-sized, cluster-mapping

small RNAs remains, so it is not clear whether UAP56 is also required for primary piRNA biogenesis.
Additionally, Piwi might be involved in the specification of de-novo piRNA clusters (de Vanssay et al. 2012; Rozhkov, Hammell, and Hannon 2013; Thomas et al. 2014), although previous reports did not attribute a major role for Piwi in ovarian piRNA biogenesis (Malone et al. 2009; Brennecke et al. 2007).

f. Processing of piRNA precursors

i. In somatic cells of the ovary

Processing of **somatic** piRNA precursor transcripts occurs in perinuclear structures called Yb-bodies. Intermediary *flamenco* transcripts and the piRNA pathway proteins Armi, Vret, SoYb, Shu, Fs(1)Yb and Piwi have been reported to be present in Yb-bodies (Murota et al. 2014) or in nuclear space adjacent to Yb-bodies (Dennis et al. 2013), and each of its components are required for accumulation of mature piRNAs. Yb-bodies frequently co-localize with Zucchini, a single-strand endoribonuclease, whose cleavage products bear a 5' monophosphate (Nishimasu et al. 2012). Zucchini is a likely candidate for primary piRNA 5'-end formation and localizes to the outer mitochondrial membrane. Knockdown of *Zucchini* leads to dispersion of the Yb-body (Murota et al. 2014), loss of mature piRNAs and an accumulation of longer than 25nt RNAs that hybridize to transposon probes on a Northern blot, which might be intermediate products of piRNA processing (piR-IL) (Saito et al. 2010; Nishimasu et al. 2012), but that lack a 5' monophosphate. It has been demonstrated that piR-ILs undergo Mg2+-dependent 3' trimming. This activity occurs after loading into Siwi (Silkworn Piwi) in Silkworm cell lysates, but the genes involved in this processing are still unknown (Kawaoka et al. 2011). Trimming activity is carried out by an insoluble exonuclease that recognizes the 3' OH. Importantly, Siwi-loaded RNAs are more stable than naked RNAs, suggesting that their activity can be maintained for longer periods of time compared to the transcriptional activity of their genomic origin. After trimming Hen-1 catalyzes methylation of piRNA 3' ends (Hen-1 also methylates siRNAs). This modification is thought to increase stability of the piRNA and is required for efficient TE silencing, however flies remain fertile (Saito et al. 2007; Horwich et al. 2007).

ii. **In germline cells**

Yb-bodies are absent in germline cells (Szakmary et al. 2009) and processing of exported piRNA precursor transcripts occurs likely in distinct, electron-dense perinuclear structures called nuage as piRNA precursor transcripts are detected in the nuage. The nuage is juxtaposed to nuclear UAP56 and Rhino signals, which mark dual-strand piRNA clusters (F. Zhang et al. 2012; Z. Zhang et al. 2014).

Detailed studies of how primary piRNA biogenesis proceeds in germline cells are still lacking, but given that mutations in primary piRNA biogenesis factors that were studied in OSC cells generally lead to a collapse of the complete piRNA pool, it is likely that primary piRNA biogenesis occurs by a similar mechanism in the germline and somatic follicle cells.

Conceptually, **primary piRNA biogenesis** is followed by either a

- an **amplification phase** in which sense transcripts (TE or other) are consumed,
- or **post-transcriptional silencing** of transcripts with piRNA complementarity,
- or re-import of Piwi-RISC complexes into the nucleus, where **transcriptional silencing** takes place.

I will first describe the amplification phase and discuss targeting of transcripts with piRNA complementarity next. Finally I will turn to the nuclear functions attributed to Piwi.

iii. Vasa organizes the nuage for piRNA amplification

The elucidation of the piRNA amplification mechanism has been greatly facilitated by a *Bombyx mori* cell line that recapitulates the full ping-pong piRNA biogenesis. Of note, *Bombyx mori* is lacking Aubergine and ping-pong amplification is occurring between Silkworm Piwi (Siwi) and Ago3. After nuclear export, piRNA precursor molecules are scanned by piRNA-loaded Siwi. Upon recognition, the precursor enters DEAD box helicase Vasa, which acts as a scaffold for piRNA amplification (Xiol et al. 2014). Transcript-bound Vasa recruits Tudor and Papi which in turn recruit Ago3 and Aubergine (L. Liu et al. 2011). In *Bombyx mori,* a Vasa mutation in the DEAD box domain (DQAD) that prevents ATP hydrolysis stabilizes Vasa interaction with Siwi and Ago3. RNA Immunoprecipitation (RIP) from Vasa-DQAD yielded both small RNAs and longer sense transposon transcripts (Xiol et al. 2014) (Zhang et al., reported 42AB transcripts in wildtype Vasa RIP (F. Zhang et al. 2012)). The ~30nt small RNAs species are constituted by molecules mostly antisense to transposons, whereas a second ~10-12nt species is thought to be the footprint of transposon transcripts that are present in Vasa and are degraded by cellular nucleases,

implicating Vasa as central part of the piRNA amplification and processing machinery that coordinates the turnover between ping-pong partners, hence it was termed the “amplifier complex” (Xiol et al. 2014).

g. Post-transcriptional regulation of piRNA source and target loci

i. Piwi is unlikely to act by a PTGS mechanism

It is less clear how genic transcripts are targeted, and whether there is a role for Piwi-loaded primary piRNAs in the post-transcriptional regulation of both TEs and genic transcripts. When considering the evidence gathered in the OSC cell line, Piwi predominantly acts by nucleating heterochromatin on TE sequences in the genome, thereby preventing their expression. If Piwi was silencing transposons post-transcriptionally by a slicer mechanism like Aubergine and Ago3 do, mutation of the slicer motif should result in increased TE expression. Piwi slicer mutant flies however do not show increased TE expression, and are fertile (Sienski, Dönertas, and Brennecke 2012). Recombinant Piwi is capable of cleaving targets when loaded by single-stranded siRNAs (21nt and 30nt), as assayed by Saito and colleagues, however the level of detected cleaved target was comparable to Ago1, which is a weak slicer and *in vivo* rather regulates targets through the recruitment of decapping enzymes and /or translational inhibition. In addition, siRNAs are not enriched in piwi IPs from ovarian lysate incubated with siRNA duplexes (Saito et al. 2006). **Altogether these data point against a role for Piwi in PTGS directly through its slicer mechanism.**

ii. Genic piRNAs as candidates for guiding PTGS

While piRNAs are predominantly derived from TE sequences, a significant fraction also aligns to the 3’UTR of genic transcripts (22% as compared to TE-piRNAs reads, 30% as compared to microRNA reads, both in OSS cells) (Robine et al. 2009), such as piRNA originating from the 3’UTR of *traffic jam* (tj). Low levels of piRNAs are found on genes juxtaposed to TE sequences or other producers of piRNA (Mohn et al. 2014; Saito et al. 2009; Robine et al. 2009; Shpiz et al. 2014), but in this case piRNA production seems to “bleed” from the TE sequences to the genic transcription unit, and these piRNAs are dependent on ping-pong amplification and the RDC complex (Mohn et al. 2014).
Genic 3’UTR piRNAs are independent of secondary piRNA pathway factors (in fact relative 3’UTR piRNA expression increases in most secondary piRNA pathway mutants, except Aubergine), they do not correlate with the expression level of the transcript from which they

derive, they have a 5'U bias, are strongly enriched in Piwi IPs and slightly enriched in Aubergine (Robine et al. 2009). It remains unknown how the piRNA biogenesis machinery selects genes for 3'UTR piRNA production, though genes with significant 3'UTR piRNAs show higher median expression levels than control genes in OSS cells, adult and 10 day post-partum mouse testes (Robine et al. 2009), suggesting that either high expression levels favor 3'UTR piRNA expression, or that 3'UTR piRNA expression facilitates high level expression or a combination of both. Genes from which 3'UTR piRNAs are produced (>1000 transcripts, according to Lau and colleagues) do show significant enrichments in GO-terms, and these overlap significantly with mouse genes that are a source for 3'UTR piRNAs, suggesting a conserved function in the regulation of 3'UTR piRNAs (Robine et al. 2009). The enriched GO-terms for 3'UTR piRNA genes involve developmental processes.

Most 3'UTR piRNAs are in sense orientation with respect to the transcript from which they derive, indicating that the mRNA is the initial substrate (or source locus) (Saito et al. 2009; Robine et al. 2009), but leaving open the question of whether the sense-piRNA loaded Piwi can in turn target the host mRNA in trans, or another, coding- or non-coding RNA with sense or antisense sequence homology. Piwi-negative mutant follicle clones show increased traffic jam protein levels, indicating that 3'UTR piRNAs cluster do have a role in repressing the host coding gene they derive from, but whether the observed derepression stems from decreased transcriptional or post-transcriptional cis- or trans-silencing has not been investigated.
Recent work (Post et al. 2014) in OSC cells suggests that genic piRNAs have to be strongly expressed to target reporter gene silencing. Importantly reporter repression was effective above background only when the reporter fragment was in the opposite orientation to the piRNAs that induce reporter targeting. This is hence pointing against a role of genic piRNAs repressing their host gene in trans, as they share the same orientation. Furthermore, reporter expression was decreased when piRNA target sites were introduced as efficiently spliced introns, suggesting that piRNA-repression occurs on the nascent transcript prior to splicing. If one were to reconcile these finding with the increase of tj protein levels in Piwi mutant follicle cell clones, it seems likely that the cis processing of the tj transcript into piRNAs is responsible for lower tj protein levels observed, and not PTGS. Post and colleagues provide evidence for trans-silencing of their reporter through genic piRNAs, but they speculate that this is more likely to occur by TGS. These findings stem from OSC and OSS cells and likely reflect the piRNA pathway that is active in ovarian follicle cells, which do not express Aubergine and Ago3. It remains possible that genic piRNAs in germline cells and early embryos do act by PTGS.

iii. piRNA-directed PTGS through homology with TE sequences

A first example of piRNA-guided post-transcriptional regulation of protein coding, non-TE genes is the piRNA-mediated degradation of maternally deposited *nanos* (*nos*) mRNA (Rouget et al. 2010). In the early embryo (0-1h after egg laying (AEL)) *nos* mRNA is ubiquitously located in the cytoplasm, and is successively degraded, except for the posterior-most region of the embryo.

Proper Nos protein localization is required for head and thorax segmentation, as Nos, which localizes to the posterior of the embryo, locally repress *hunchback* mRNA translation, while ectopic presence of Nos at the anterior part of the embryo represses *bicoid* mRNA.

Destabilization of *nanos* mRNA at the maternal-to-zygotic transition (MZT) depends, among other signals and pathways (Smaug-recognition element in the proximal-most part of the 3'UTR, CCR4, Smg (Rouget et al. 2010)), on two sites in the *nanos* 3'UTR with partial complementary to *roo* and *412* TE piRNAs (piRNAs correspond to the antisense strand of *roo* and the sense strand of *412*, respectively).

Transgenes lacking piRNA homology sites show increased polyA-tail lengths, increased stability of *nos* mRNA in the bulk of the embryo cytoplasm and defects in head development resulting from lack of *bicoid* and *hunchback* mRNA repression in the anterior region of the embryo.

Similar phenotypes were observed in piRNA mutant embryos and embryos injected with synthetic 2'O-methyl RNAs antisense to the *roo* and nos piRNAs (Rouget et al. 2010). This indicates that TE-derived piRNAs could in principle target genic mRNAs in trans, with the caveat that *nanos* regulation also depends on other sequence elements in its 3'UTR and that nanos-regulation occurs in the early embryo. The cytoplasmic distribution of Smg and the CCR4-Not complex, the proposed effectors of piRNA mediated *nos* mRNA degradation during the MZT, are disrupted in *aubergine* mutant embryos, indicating that piRNA could be globally involved in the MZT.

Whether other genes with TE-derived piRNA complementarity could be subject to nanos-like piRNA-mediated repression is not clear. In addition it is not known whether piRNAs play a role in regulating gene expression after the MZT.

Other examples of piRNA mediated-PTGS include [1] the post-transcriptional cleavage of a sex-determination factor in *Bombyx mori* guided by a female specific piRNA (Kiuchi et al. 2014) and [2] the repression of approximately 40% of the mRNAs in elongating mouse spermatids by Miwi (Gou et al. 2014). Repression in this case is linked to the CCR4-Not complex, suggesting a similar PTGS mechanism as *nanos* mRNA degradation during the *Drosophila* MZT.

h. piRNA-mediated transcriptional gene silencing

i. Nuclear function of Piwi

The *Drosophila* piRNA pathway had been linked retrospectively to transcriptional gene silencing (TGS) via its role in co-suppression (M Pal-Bhadra, Bhadra, and Birchler 1997) suppression of variegation and Polycomb-mediated gene silencing (Grimaud et al. 2006) even before its molecular mechanisms were identified, but solid molecular evidence for this mode of silencing surfaced only recently. As mentioned before, piRNA-loaded Piwi (Piwi-RISC) shuttles into the nucleus where it is localized in close proximity to chromatin. Piwi accumulates in the nucleus only when loaded with piRNAs (Saito et al. 2010), hence loss of piRNA biogenesis factors triggers delocalization of Piwi. A slight delocalization is also observed when knocking down Aubergine, suggesting that the ping-pong loop might provide piRNAs to Piwi (S. H. Wang and Elgin 2011).

Piwi is guided to nascent RNA transcripts by its loaded piRNA and recruits HP1 in a RNase-dependent fashion. HP1 in turn recruits Su(var)-3-9 via its chromo-shadow domain for the tri-methylation of H3K9 (Huang et al. 2013). This process is independent of Piwi catalytic activity and instead depends on its NLS. Similarly, Piwi catalytic mutants are fertile and show no obvious phenotypic defects, while Piwi-**Δ**NLS mutants have strongly increased TE levels, egg laying and fertility defects.

ii. Piwi-mediated TGS of TE insertions define euchromatic H3K9me3 islands

Detailed analysis in OSC cells revealed that TE insertions are responsible for most (88%) euchromatic H3K9me3 islands (Sienski, Dönertas, and Brennecke 2012). Strikingly, PolII occupancy on these euchromatic H3K9me3 islands is very similar to random control windows regions, indicating that euchromatic H3K9me3 is compatible with substantial transcription. Upon knockdown of Piwi, a subset of TEs ("Group I", defined by Piwi-sensitivity) shows strong increases in nascent and steady-state RNA levels as well as a reduction of H3K9me3 levels and increases in PolII occupancy. Increased PolII levels were also found immediately downstream but not upstream of TE insertions, suggesting that PolII can "bleed" into flanking sequences. A low number of genes (34) in vicinity to "group I" TE insertions also showed increased expression levels, PolII occupancy and nascent RNA levels upon Piwi, Armitage or Maelstrom knockdown, suggesting that TE insertions can cause Piwi-dependent spread of a heterochromatic environment. Importantly, transcription-

defective truncated TE fragments in intronic sequences also trigger nucleation of H3K9me3, but only if the insertion is in sense orientation of an active transcription unit and antisense piRNAs against the sense TE transcript exist in OSC cells. This suggests that recruitment of Piwi to target loci occurs through nascent transcripts, paradoxically requiring transcription for transcriptional gene silencing (Sienski, Dönertas, and Brennecke 2012; Post et al. 2014). In agreement with the finding that PolII is not depleted from euchromatic H3K9me3 islands, Sienski and colleagues also showed that in OSC cells Maelstrom knockdown leads to elevated TE steady-state RNA levels and PolII bleeding out of TE sequences, without affecting piRNA production or H3K9me3 levels on the TEs itself. In *Drosophila* Maelstrom thus functions downstream of piRNA biogenesis and H3K9me3 establishment to silence TEs. This again confirms that H3K9me3 is not preventing PolII transcription, and that histone modifications are likely not the endpoint of Piwi-mediated silencing (Sienski, Dönertas, and Brennecke 2012). This study did not address the role of HP1 in Piwi-mediated TGS, which Huang and colleagues proposed to be the initial silencing signal that is recruited by Piwi. One hypothesis might thus be that Piwi, through the action of Maelstrom, recruits other chromatin modifiers that reduce PolII transcription or target/signal transcripts for degradation.

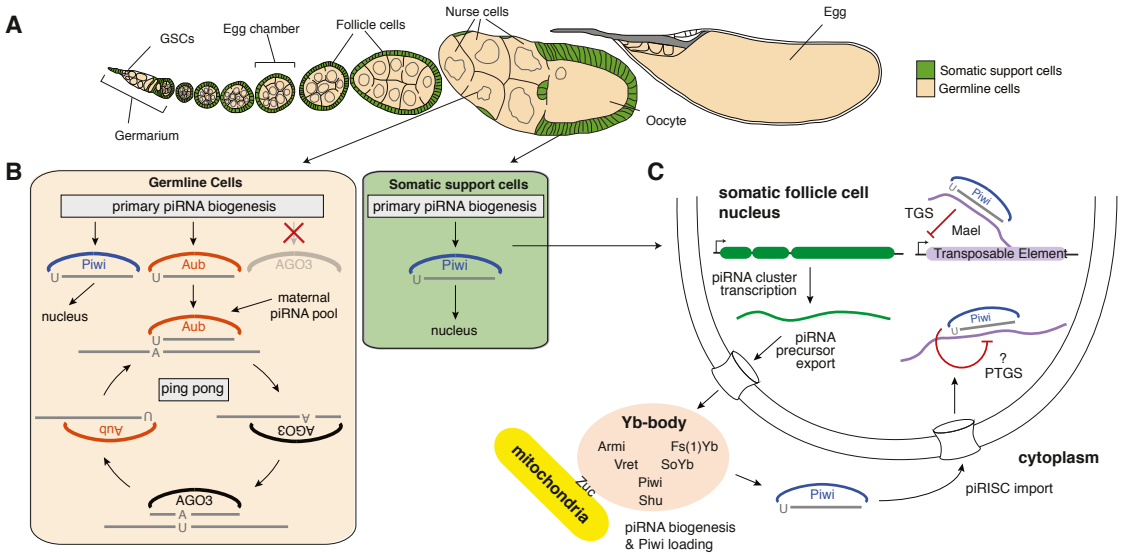


Figure 4. A) Cartoon of a Drosophila ovariole with somatic cells in green and germline cells in beige. (B) Schematic representations of the Drosophila germline and somatic piRNA pathways focusing on the three PIWI family proteins and the biogenesis routes of their bound piRNAs. From (Handler et al., 2013).

iii. Maelstrom as an adapter molecule between heterochromatin and transcript degradation

Maelstrom is a protein that contains a high mobility group [HMG] box, that is required for DNA binding in a number of proteins, and mutations in this domain lead to increase TE levels (Sienski, Dönertas, and Brennecke 2012). Maelstrom is conserved in mice and required for transposon silencing and fertility. In mice Maelstrom is in complex with Miwi and the Tudor-domain protein Tdrd6. Mouse Malestrom specifically binds pachytene piRNA precursors (Castañeda et al. 2014), and loss of Maelstrom strongly reduces pachytene piRNAs. Maelstrom function in this case was proposed to be a nucleo-cytoplasmic shuttling chaperone for piRNA cluster transcript, required for precursor handover to the nuage.
A hypothesis on how to reconcile these apparently different roles (loss of pachytene piRNAs versus a step downstream of Piwi-mediated TGS) for Maelstrom in flies and mice was put forward by Pandey and Pillai, proposing that Maelstrom binds nascent TE transcripts and targets them to cytoplasmic degradation granules, which in mouse would result in piRNA production, whereas in fly this would result in decay of TE transcripts (Pandey and Pillai 2014).
This would however not explain the increased PolII transcription inside TE and flanking sequences upon Piwi knockdown (Sienski, Dönertas, and Brennecke 2012), unless nascent transcript binding by Maelstrom would terminate transcription. Clearly further work is required to understand how Maelstrom is targeted to nascent transcripts, and why it is required for pachytene piRNAs in mouse but not for piRNA production (neither 3'UTR or ping-pong) in *Drosophila*.

iv. Loss of Piwi function affects rate of TE transcription and stability

TGS installed by Piwi was also found to be the predominant mode of TE silencing in ovarian somatic follicular cells, as tissue-specific RNAi using the *traffic jam* Gal4 driver caused loss of H3K9me3, increased steady-state and nascent RNA levels (Rozhkov, Hammell, and Hannon 2013). While this is also true for *nanos*-Gal4 driven germline knockdown, steady-state TE RNA levels in this case increased much stronger than nascent RNA levels. It is thus likely that in germline cells both transcriptional rate and transcript stability of TEs increases. This difference between somatic and germline cells might be linked to the role that Piwi might play on th specification of piRNA clusters (discussed in "**Germline piRNA precursor transcription**"). Thus, if the clusters fuel less piRNA biogenesis upon piwi knockdown, Ago3 and Aubergine are less active on TE transcripts, explaining the apparent increase of TE transcript stability. In line with this, piRNAs specific for TEs that show

stronger increases in steady-state RNA decreased in abundance. The effect was particularly strong for the telomeric TEs HetA, TART and TAHRE. These are not found in the classical piRNA clusters described by Brennecke (Brennecke et al. 2007), and instead acts as “mini-clusters” whose primary piRNAs are derived from the same loci that are targeted for repression.

Remarkably, none of these studies (Thomas et al. 2013; Rozhkov, Hammell, and Hannon 2013) reported significantly up regulated protein-coding genes except for those involved in cellular stress and DNA-damage signaling. These genes are likely activated as a consequence of increased transposition, as opposed to being direct targets of transcriptional silencing by Piwi.

3. siRNA silencing

The *Drosophila* siRNA pathway (often synonymously called RNAi) is well separated from the miRNA pathway, as the key genes involved are distinct from those of the miRNA pathway, contrary to the situation in mammals, aiding in dissecting the function of each pathway

Soon after Mello and Fire's demonstration of RNAi in *C.Elegans* (Fire et al. 1998), long exogenous dsRNA was injected into *Drosophila* embryos to study gene function (Kennerdell and Carthew 1998; Misquitta and Paterson 1999) and their principal molecular actors Dicer-2 and Ago2 were identified (Bernstein et al. 2001; Hammond et al. 2001)).

The discovery that soaking long dsRNA in the supernatant of cultured *Drosophila* cells could knock down homologous genes in a fast, cheap and easy fashion allowed very powerful reverse genetic RNAi screens. These were later extended to *in -vivo* screens when transgenic collection of hairpin lines were established by a number of consortia (VDRC and DRSC, (Dietzl et al. 2007; Ni et al. 2008))

As RNAi is cell-autonomous in Drosophila, even large scale tissue and cell-type specific screens are possibe (Roignant et al. 2003) as long as a Gal4 driver for the cells of interest are available. Strikingly, this also led to the identification of additional miRNA, siRNA and piRNA factors (Zhou et al. 2008; Carré et al. 2013; Czech et al. 2013; Olivieri et al. 2010; Handler et al. 2013)

The *Drosophila* siRNA pathway has been first and foremost involved in the defense against viral infections (van Rij et al. 2006; Zambon, Vakharia, and Wu 2006), maintenance of TE repression (Ghildiyal et al. 2008), establishment and/or maintenance of heterochromatin (Fagegaltier et al. 2009), heat shock responses (Lucchetta, Carthew, and Ismagilov 2009), promoter pausing (Cernilogar et al. 2011), mitotic progression, chromatin looping (Moshkovich et al. 2011), splicing (Taliaferro et al. 2013) and polycomb-mediated silencing (Grimaud et al. 2006; Moshkovich et al. 2011), although many of these reports are contradictory and often lack a clear molecular mechanism.

I will first summarize the well-defined mechanism of PTGS induced by canonical long dsRNAs and then discuss the various function attributed to the siRNA pathway.

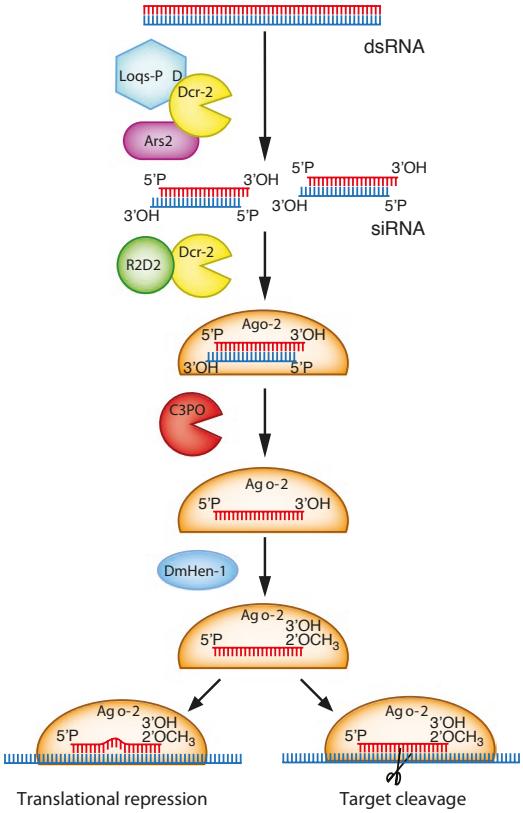


Figure 5: Overview and properties of the *Drosophila* siRNA pathway. Adapted from (van Mierlo et al., 2011)

a. Molecular view of the siRNA pathway

The principal source molecules of the siRNA pathway are stretches of complementary and annealed RNA, or more simply put, double-stranded RNA molecules (dsRNA). Long dsRNA is

processed into 21-nt siRNA duplexes by the endoribonuclease Dicer-2 (Dcr-2). In contrast to Dicer-1 which lacks the DExD/H domain, Dicer-2 contains a subdivided amino-terminal helicase domain with a DExD/H helicase domain and a HELIc helicase domain, followed by dsRNA binding domain (dsRBD), a Piwi-Argonaute-Zwille (PAZ) domain, two tandem RNAseIII domains and a C-terminal dsRBD. Dicer-2 cleaves both short (30nt) and long (>30nt) dsRNA molecules. Dicer-2 is a processive enzyme, with the initiation being the rate-limiting step. Its translocation along longer dsRNA molecules hydrolyzes approximately 21 molecules of ATP per siRNA duplex and depends on the helicase activity (Cenik et al. 2011; Welker et al. 2011). The two RNAseIII domains pair intramolecularly to form the active site necessary to cleave the two strands of the dsRNA with 2nt overhangs at the 3'end. The PAZ domain binds the 3' overhangs, and the distance by between the PAZ domain and the active site determines the precise 21 nt length of siRNA duplexes (Zamore et al. 2000; Elbashir et al. 2001; Nykänen, Haley, and Zamore 2001; MacRae, Zhou, and Doudna 2007). Finally, the

dsRBD domains are thought to enhance the affinity for dsRNAs. *In vivo*, the loquacious isoform D (loqs-PD) and Ars2 enhance processing of dsRNA substrates, while R2D2 interacts with the helicase domain of Dicer-2 and is required for efficient loading of 21nt dsRNA duplexes into Ago2 (known as the RISC-loading complex or RLC). In the absence of R2D2 siRNAs are preferentially loaded in Ago1, indicating that R2D2 is required for proper sorting of siRNAs in Ago2, as siRNAs loaded in Ago1 have decreased capability to cleave targets and might cause silencing in a more miRNA-like mechanism. R2D2 might do so by sensing the 1st nucleotide (strong bias for U) and central structure of the siRNA duplex, as R2D2 mutants show altered nucleotide biases at position 1, 9 and 10 (Marques et al. 2010; Okamura et al. 2011; Mirkovic-Hösle and Förstemann 2014).

b. Processing of dsRNA is enhanced in D2 bodies

Assembly of the RLC probably occurs in distinct cytoplasmic structures termed D2 bodies, whose principal components are Dicer-2, R2D2, while Ago2 is found transiently in D2 bodies. D2 bodies partially colocalize with P-bodies but are distinct from them, and they are also distinct from stress granules. D2 bodies are present in S2 cells and OSC cells, ovarian follicle cells but are absent from nurse cells and oocytes (Nishida et al. 2013). Interestingly, efficient RNAi knockdown in nurse cells and and oocytes requires overexpression of Dicer-2 (S. H. Wang and Elgin 2011) possibly due to the absence of D2 bodies.
Knocking down R2D2 leads to disruption of D2 bodies and misloading of endo-siRNAs into Ago1, while artificial siRNA duplexes are still loaded into Ago2 independently of D2 bodies. As Dicer-2 mutants lacking the dsRNA binding domain do not localize to D2 bodies, Nishida and colleagues proposed a model in which D2 bodies are the sites of Ago2 loading. Dicer-2 with siRNA duplexes would localize to D2 bodies, where the local concentration of loading factors are high, allowing efficient loading in Ago2, that can nevertheless take place outside of D2 bodies. Upon loading, Dicer-2/R2D2 complexes would leave the cytosol to associate with dsRNA duplexes and relocalize to D2 bodies (Nishida et al. 2013).

c. Strand selection and target cleavage

Dicer-2/R2D2 complexes as part of the RLC are thought to bind the thermodynamically less stable 3'end of the siRNA duplex, followed by an exchange of the Dicer-2/R2D2 complex with Ago2 (Tomari, Du, and Zamore 2007). ATP-mediated "stretching" of the Ago2 conformation by the chaperones HSP90 and Hsc70 is required for efficient transfer of the duplex into Ago2 (Iwasaki et al. 2010; Miyoshi et al. 2010). Ago2-loaded siRNA duplexes are termed pre-RISC, and conversion of pre-RISC into RISC requires hydrolysis of the

phosphodiester bond between nucleotides 9 and 10 of the passenger strand by the endonucleolytic activity of the Piwi domain. Nicked passenger strands are subsequently degraded by the endonucleolytic activity of the translin-TRAX complex (known also as C3PO), yielding active Ago2-RISC loaded with a single strand 21 nt siRNA, ready to cleave complementary target RNAs (Tian et al. 2011). Hen-1 further increases the stability of the mature siRNA by transferring a methyl group from S-adenosyl methionine (SAM) to 3' OH group, thereby protecting the siRNA from uridylation and other modifications that decrease small RNA stability (Sement et al. 2013; Ji and Chen 2012). Target RNA cleavage is a multiple-turnover catalytic process, whose rate-limiting step is the release of cleaved mRNA which is facilitated by the autoantigen protein La (Haley and Zamore 2004; J. Liu et al. 2004; Martinez et al. 2002; Rivas et al. 2005).

d. The siRNA pathway controls viral infections.

The primary function attributed to the siRNA pathway is in the defense against RNA viruses and TEs. The function of the siRNA pathway in antiviral immunity is linked to the cleavage of viral dsRNA, as well as targeting of viral single-strand RNA by virus-derived siRNAs (vsiRNA) loaded in Ago2, leading to destabilization and/or degradation of viral RNA. There is a multitude of potential viral dsRNA sources, depending on whether the viral genome is double-stranded, single-stranded with messenger RNA transcribed from the RNA that is packaged in the virion (+ strand) or whether the messenger RNA is transcribed after RdRP transcription (- strand). During infection with double-stranded RNA viruses, the viral genome itself is the target of Dicer-2 cleavage, wheres for + strand RNA viruses the most likely substrate are the dsRNA replication intermediates. This is supported by approximately equal abundance of vsiRNA mapping to the + and - strand, even though the + strand is 10-100 fold more abundant. Similar to + strand viral infections, during - stand viral infections vsiRNA abundance is also balanced between + and - strand, with no enriched regions, indicating again that the replication intermediates serve as substrates for the RNAi machinery. Recent reports also suggest targeting of the DNA viruses IIV-6 by the siRNA machinery, with vsiRNA matching to regions transcribed from both strands of the viral genome (Kemp et al. 2013; Bronkhorst et al. 2014).

The importance of antiviral RNAi in *Drosophila* is highlighted by the increased sensitivity of siRNA pathway mutants (*dicer-2, r2d2 and ago2*) to viral infections which show higher viral titers as well as increased mortality (van Rij et al. 2006; Galiana-Arnoux et al. 2006; X.-H. Wang et al. 2006). Indeed viruses have independently evolved strategies to evade their hosts' antiviral siRNA system by means of proteins that either inhibit siRNA pathway proteins or that shield double-stranded RNA from processing by Dcr-2 or Ago2 (Van Rij and

Andino 2008). As opposed to plants, the *Drosophila* RNAi pathway is lacking a RdRP that could amplify siRNAs and acts cell autonomously (Roignant et al. 2003; Lipardi and Paterson 2009; Yoshikawa et al. 2005; Xie et al. 2005; Gasciolli et al. 2005). During viral infections however, a protective vsiRNA response can be detected in uninfected cells, which depends on endocytic uptake of long dsRNA emitted from infected cells, possibly from lysed cells during infection or an unknown trigger sensitive to viral infections (Saleh et al. 2009).

e. The siRNA pathway represses TEs

The siRNA pathway has also been implicated in the defense against TEs. Unlike piRNA pathway mutants, which are generally sterile, siRNA pathway mutants do not show strong fertility defects (Wen et al. 2014). Nevertheless, strongly increased steady-state TE RNA levels have been reported for *Dcr-2*, *r2d2*, *Hen-1* and *Ago2* mutants in S2 cells, adult heads and carcasses (Horwich et al. 2007; Ghildiyal et al. 2008; Czech et al. 2008; Ameres et al. 2010; Mirkovic-Hösle and Förstemann 2014; Li et al. 2013). Most transcriptionally active TEs are targeted by 21-nt sense- and antisense siRNAs along their sequence, suggesting active degradation similar to viral restriction. Depending on the abundance of piRNAs in the tissue under study, different TE families may be targeted predominantly by siRNAs, piRNAs, or a combination of siRNAs and piRNAs.

As for viruses, multiple potential sources of dsRNA exist. This includes bidirectional transcription of transposable element sequences, pairing of transcripts from piRNA clusters with active full length copies distributed throughout the genome and secondary structures of transposable element RNA.

Importantly, the increases in transposable element RNA in mutants of the siRNA pathway are rather modest (<20 fold) (Ghildiyal et al. 2008; Mirkovic-Hösle and Förstemann 2014) as compared to knockdown of Piwi in ovaries (>100 fold) (Rozhkov, Hammell, and Hannon 2013) or OSC cells (Sienski, Dönertas, and Brennecke 2012), but the actual fold changes depend on the TE family, tissue, genotype and technique (qRT-PCR or cDNA sequencing) under consideration. In general it appears that measured fold changes of TE expression upon knockdown of RNAi factors in cell lines are stronger than mutants of the same factor in ovaries. It also appears that TE levels in Piwi NLS mutants ovaries are only moderately increased, while follicle and germline knockdown result in strongly increased TE levels (Klenov et al. 2014). This is in apparent contradiction with (Sienski, Dönertas, and Brennecke 2012), however Sienski and colleagues have tested only 2 TEs with a single normalization to *act5C*, a gene that appears downregulated in my datasets.

Given the proposed function of Piwi in TGS, it might be reasonable to assume that TEs remain transcriptionally repressed, and only few TE insertions are actively transcribed and

repressed post-transcriptionally by the siRNA pathway, hence the lower de-repression observed in siRNA pathway mutants. Whether the siRNA pathway is involved in the repression of TEs by establishment or maintenance of TGS has not been explored extensively and results are contradictory (Fagegaltier et al. 2009; Moshkovich and Lei 2010). For Ago2, Dicer-2, and r2d2 mutations and viral proteins that suppress nuclear RNAi Fagegaltier and colleagues found a strong redistribution of HP1 and H3K9me2 marks along chromosomes, while Moshkovich and Lei showed little changes of HP1 on the status of piRNA clusters in Ago2 mutant heads. This is not necessarily a contradiction, as piRNA clusters were not specifically tested by Fagegaltier and colleagues.

f. Multiple links between piRNA and siRNA pathway

Both piRNA and siRNA pathway have been reported to suppress variegation of variegation reporters (Manika Pal-Bhadra et al. 2004; Fagegaltier et al. 2009; Gu and Elgin 2013), such as white-mottled-4 (wm4) (Muller 1930) and stubble-variegated (SbV) (Lewis, E.B. (1956). This is interesting also as a demonstration of Piwi having an effect on an adult somatic tissue.

In addition, piRNAs and siRNAs frequently co-occupy transposable element insertions and piRNA clusters. piRNA and siRNAs share methylation at their 3' end by Hen1, and loss of both pathways leads to an increased of TE expression. Furthermore, Ago2 and piwi co-immunoprecipitate with CP190 in early embryos, suggesting that Ago3 and Piwi in part occupy similar genomic regions (Moshkovich et al. 2011). However chromatin association of Ago2 has been found even for catalytic mutants of Ago2 that are unable to cleave targets, so its chromatin binding function is likely independent of guidance by siRNAs, as Ago2 binding sites also show an inverse correlation with siRNA abundance (Moshkovich et al. 2011). Furthermore an increase of siRNAs against TEs was found when piwi was knocked down either in ovarian somatic cells or germline cells (Rozhkov, Hammell, and Hannon 2013). One TE that showed increased levels (6-fold) of siRNAs upon piwi knockdown was ZAM. Strikingly, nascent ZAM RNA increased 3 fold stronger than ZAM steady-state RNA, suggesting that this siRNA response limits further increases in ZAM transposon levels.

Lastly, piRNAs and siRNAs can both target viral infections (Sindbis Virus and Rift Valley Fever Virus (RVFV), respectively) (Vodovar et al. 2012; Léger et al. 2013), with the a stronger siRNA response against RVFV during early infection time points and increasing levels of piRNA production during later infection.

Results I

Before starting the project that led to the article draft below, I tried to perform Chromatin Immunoprecipitation followed by DNA sequencing (ChIP-seq) for Dicer-2. This project was a follow-up to a previous publication of our laboratory that showed a redistribution of silent chromatin marks in mutations of siRNA pathway genes and when nuclear-targeted viral suppressors of RNAi were expressed (Fagegaltier et al. 2009). As Ago2 had been shown to associate to Chromatin independently of its catalytic activity (Moshkovich et al. 2011), and Dicer-2 had been shown to be chromatin-associated at the Hsp70 locus (Cernilogar et al. 2011) and to colocalize with Polycomb protein Polyhomeotic (PH) (Grimaud et al. 2006), I focused on Dicer-2.
I did however not manage to consistently detect transgenic Flag-tagged Dicer-2 at the Hsp70 locus in embryos and the S2 cell line. Doubt arose when immunofluorescence microscopy suggested that Dicer-2

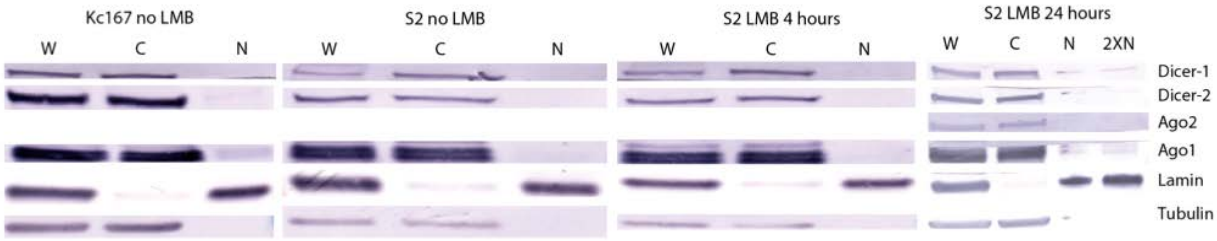


Figure 6: Subcellular fractionation experiments reveal preferential localization of Dicer-2 in the cytoplasm. Equal amounts of whole cell lysate (W), Cytoplasmic (C) and Nuclear (N) extract were loaded, except in the last panel for the lane 2XN, where twice the amount of nuclear extract was loaded. Tubulin served as a cytoplasmic marker and Lamin served as a nuclear marker.

was more likely to be cytoplasmic (confirmed by (Nishida et al. 2013)). Finally I performed subcellular fractionation experiments (which were shown in Supplemental Figure S1a to (Cernilogar et al. 2011)) and found the exact opposite result with the same cell line and with KC cells (Figure 7). Treatment with Leptomycin B, an inhibitor of nuclear export (Kudo et al. 1998), did not cause nuclear retention of Dicer-2, which does have a potential Nuclear Export Sequence (NES). While this is not a definitive demonstration that Dicer-2 does not have a nuclear function, I abandoned the project after consultation with my thesis committee in January 2013 and I started to focus on the somatic, non-gonadal function of the piRNA pathway. It is currently debated whether piRNAs exist outside of the gonads and what their function might be. It appears that piRNA-like molecules can be detected in the heads of adult flies (Yan et al. 2011; Ghildiyal et al. 2008; Mirkovic-Hösle and Förstemann 2014), but also in the central nervous system (CNS) of *Aplysia* (Rajasethupathy et al. 2012) and the *mouse* hippocampus (E. J. Lee et al. 2011). In *Aplysia* neurons piRNAs appeared to be abundant with a specific set of piRNA enriched as compared to the ovotestis. Moreover, overexpressed Piwi proteins accumulated in the nucleus of sensory neurons and piRNAs are implicated in memory formation via

methylation of the CREB2 promoter. In the *mouse* hippocampus piRNAs represented about 12% (11% >24 nt, and 1% corresponding to known Miwi-bound piRNAs) of the total sequenced small RNAs, Miwi-bound piRNAs could be immunoprecipitated from brain, and Miwi was found to colocalize with piRNAs in the cytoplasm of cultured hippocampal neurons. In contrast, *Drosophila* piRNA-like molecules in the head have not yet been ascribed a specific function, and have not been immunoprecipitated with a Piwi-class proteins. While the secondary piRNA pathway members Ago3 and Aubergine can be detected by Immunofluorescence in the brain (Perrat et al. 2013), they colocalize only in a limited subset of neurons. In addition, Piwi could not be detected in the brain, possibly because the Piwi locus is actively repressed by the l(3)mbt transcriptional repressor (Blanchard et al. 2014) .

On the other hand, mutants of the piRNA pathway are clearly affecting somatic tissues, as embryos devoid of maternally deposited piRNA complexes are arrested early in development (Mani, Megosh, and Lin 2014; S. H. Wang and Elgin 2011). Furthermore mutations in piRNA pathway genes have been shown to affect Position Effect Variegation (PEV) of the wm^4 sensor and distribution of chromatin marks (Manika Pal-Bhadra et al. 2004; Gu and Elgin 2013). The observed phenotypes of *piwi* mutations in the head were the starting point of the project that led to the following article draft.

We started the project by sequencing small RNAs in the embryos, early larvae and heads of flies in the wm^4 background, to determined whether closeby small RNAs might be responsible for the observed suppression of variegation in *dicer-2 (Fagegaltier et al. 2009)* and *piwi* (Gu and Elgin 2013; Manika Pal-Bhadra et al. 2004) mutants. Unfortunately we did not find evidence for small RNAs that might be at the root of wm^4 variegation, but by the summer of 2013 we saw an effect of *piwi* mutation on *gypsy*, and to a lesser extent on other TEs derived siRNA. This was the first hint that *piwi* was implicated in TE regulation in adult non-gonadal soma. In September 2013 we discovered that (Gu and Elgin 2013) had also worked on the somatic effect of *piwi* mutation. In their paper, Gu and Elgin found that maternal deposition of *piwi* was required for heterochromatin formation over PEV loci and TE sequences, as HP1 levels decreased slightly in *piwi* mutants. We then continued and asked whether *piwi* and *dicer-2* mutation would also affect TE transcript levels, given that the siRNA levels changed in both mutants. As both had only minor effects, except for *gypsy* expression, we hypothesized that both siRNA and piRNA pathways might represent complementary layers of TE repression, and so we recombined both mutant alleles and performed RNA-seq in heads, and indeed, the TE transcript levels were significantly elevated in double mutants. These results are presented in the following manuscript draft.

Article: Piwi-dependent transcriptional silencing and Dicer-2-dependent post-transcriptional silencing limit transposon expression in adult heads of *Drosophila melanogaster*

Supplemental Tables are in Annex 1

Title

Piwi-dependent transcriptional silencing and Dicer-2-dependent post-transcriptional silencing limit transposon expression in adult heads of *Drosophila melanogaster*

Marius van den Beek, Bruno da Silva, Christophe Antoniewski
Drosophila Genetics and Epigenetics ; Université Pierre et Marie Curie 9, Quai St Bernard
Building C – 5th floor – Room 517 ; 75252 Paris cedex 05 ; Phone: +33 1 44 27 34 39

Abstract

Whether the piRNA pathway plays a functional role in adult, non-gonadal tissues has not been definitively answered to date. We have sequenced the small RNA content of adult *Drosophila melanogaster* heads of wild type and *piwi* mutants to address whether *piwi* loss of function would affect piRNA-like molecules that can be detected in wild type heads. We find that loss of *piwi* does not affect these molecules. Instead, we observe increased siRNA levels against the majority of *Drosophila* transposable element (TE) families. To determine the effect of this siRNA response to *piwi* loss, we sequenced the transcriptome of wild type, *piwi*, *dicer-2* and *piwi, dicer-2* double-mutants. We find that the expression levels of the majority of TE families in *piwi* and *dicer-2* mutants remain unchanged and that TE expression increases significantly in *piwi, dicer-2* double-mutants. Concordantly, we observed a significantly decreased lifespan for *piwi, dicer-2 double-mutant*s. These results lead us to suggest a dual-layer model for TE repression in somatic tissues. *piwi*-mediated transcriptional silencing (TGS) established during early development constitutes the first level of TE repression. In addition, dicer-2-dependent siRNA-mediated post-transcriptional gene silencing (PTGS) provide a backup mechanism to repress TEs that escape silencing by *piwi*-mediated *TGS*.

Introduction

Transposable element (TEs) activity is thought to be an important force in genome evolution, as TE integration and excision can result in gene duplication, deletion, and the modification of gene signaling networks (Tubio et al. 2014). However, uncontrolled integration into host genes might be detrimental to individuals, potentially creating harmful mutations that decrease lifespan and fertility. Therefore a low frequency of TE mobilization is beneficial to both host and TE, whereas high TE activity decreases host fitness and adversely affects vertical transfer of the TE.

In *Drosophila melanogaster*, the siRNA and the piRNA pathways are important negative regulators of TE expression in somatic (Li et al. 2013; Ghildiyal et al. 2008) and gonadal

tissues (Vagin et al. 2006), respectively. Both pathways are active in the gonads, while the siRNA pathway is thought to be active in all somatic tissues. Dicer-2, the central endonuclease of the siRNA pathway operates on double-stranded RNA molecules by processively sliding along the molecule and cutting every 21st nucleotide. After the initial processing, these 21nt duplexes with 3'OH overhangs are rebound by Dicer-2 and r2d2 to be loaded into Ago2. Ago2 then cleaves the passenger strand and can engage in multiple rounds of endonucleolytic cleavage of transcripts with mature siRNA complementarity. Loss of functional siRNA pathway results in increased levels of TE expression and mobilization (Li et al. 2013; Xie, Donohue, and Birchler 2013; Czech et al. 2008; Ghildiyal et al. 2008), but fertility is not affected. Ago2 has also been implicated in splicing (Taliaferro et al. 2013) and identified as a TrxG group gene, but this function appears to be independent of its cleavage activity (Moshkovich et al. 2011).

The *Drosophila* piRNA pathway is well characterized for its role in maintaining genome integrity in the gonads and is therefore required for fertility. Its key components are 3 germline Argonaute-family proteins Piwi, Aubergine (Aub) and Ago3. piRNA production is initiated from piRNA cluster transcripts, which contain antisense TE sequences. After processing of these transcripts by the piRNA biogenesis machinery, primary antisense piRNAs are loaded into Piwi or Aub in germline cells. piRNA-loaded Aub slice sense TE transcripts between the tenth and 11th position, which will become the 5' end of a new TE sense piRNA. Sense piRNAs are then loaded into Ago3, which in turn slices complementary antisense TE sequences between the 10th and 11th position. This cytoplasmic cyclic PTGS process (also termed secondary piRNA amplification) leaves a detectable "ping-pong" signature, in which 5' and 3' ends of piRNAs tend to overlap by 10 nucleotides.

piRNA-loaded Piwi complexes can re-enter the nucleus, where piRNAs guide Piwi towards complementary nascent transcripts. Piwi then recruits factors (presumably Maelstrom, SuVar3-9, dSETDB1 and HP1) that establish H3K9me3 at the surrounding genomic vicinity of TE insertion sites, that may or may not, include protein coding genes and hence functions in TGS (Sienski, Dönertas, and Brennecke 2012; Brower-Toland et al. 2007; Wang and Elgin 2011; Rangan et al. 2011).

Of note, zygotic piwi expression has been detected ubiquitously in early embryos up to the 14th nuclear division (~2h after egg laying) (Mani, Megosh, and Lin 2014; Rouget et al. 2010), depletion of Piwi in nurse cells and oocytes results in early arrest of embryonic development (Mani, Megosh, and Lin 2014; Wang and Elgin 2011), and *piwi* acts as a suppressor of variegation in the eye (Gu and Elgin 2013; Pal-Bhadra et al. 2004), suggesting an important function for Piwi during early development of somatic tissues. However the function of Piwi in other somatic tissue is quite ambiguous as Piwi has both been reported to be present (Brower-Toland et al. 2007) and absent (Thomas et al. 2013) in 3rd instar larval

salivary glands. In contrast, Aubergine and Ago3 have been found in non-overlapping cells of the adult central nervous system (Perrat et al. 2013).

In order to unravel the role of piRNA pathway in TE control in somatic adult tissues, we analyzed small RNA profile in wild type, *piwi* and *dicer-2* mutant fly. We provide evidence that previously reported ping-pong pairs in adult heads likely result from contamination with testicular RNA, suggesting that secondary piRNA amplification does not play a major role in adult heads. Small RNA sequencing of piwi mutant heads did not reveal a reduction of piRNA-like molecules. Instead increased levels of siRNAs against most TE families were detected. RNA-sequencing of *piwi* mutant heads and *dicer-2* mutant heads showed only minor upregulation of transposable elements, whereas double-mutants of *piwi* and *dicer-2* showed increased TE levels and a strong decrease of lifespan. Our results suggest a dual-layer model of TE repression in somatic tissues. The first layer of TE repression is established by Piwi at the chromatin level during early development. When TEs escape to epigenetic piwi silencing, PTGS triggered by *dicer-2* and siRNAs, mediates TE degradation to decrease TE burden and allow flies survival.

Results

Secondary piRNA biogenesis is not detectable in adult heads

It is well established that a “ping-pong signature”, a strong tendency for individual piRNA-sized reads to overlap by 10 nucleotides with another piRNA-sized read of the opposite orientation, can be found when analyzing small RNA libraries prepared from ovaries (Brennecke et al. 2007; Gunawardane et al. 2007) and testes (Nagao et al. 2010). This signature collapses in *Ago3* and *Aub* mutants. Of note, Zheng and colleagues, Ghildiyal and colleagues as well as Mirkovic and Förstemann (Yan et al. 2011; Ghildiyal et al. 2008; Mirkovic-Hösle and Förstemann 2014) previously reported the existence of piRNA-like molecules in the head, based on their size and ping-pong-signature which is in line with the fact that Ago3 and Aubergine can be detected in the optic lobe of the *Drosophila* central nervous system (Perrat et al. 2013). If Ago3 and Aubergine were actively producing piRNAs, as they do in the germline, we reasoned that we should be able to find a ping-pong pattern in small RNA libraries of adult heads. To do so we re-analyzed 24 small RNA sequencing libraries prepared from adult male heads (Reinhardt et al. 2012) and aligned them to the *Drosophila* genome. The majority of reads were in the size range of miRNAs (22-23 nt) and siRNAs (21 nt), but we observed a fraction of reads in each library (mean 7.5%, interquartile range 5.6-10.4%; Supplemental Fig. S1A,S1B) compatible with the size of piRNAs in *Drosophila*.

In the 24-28 nt fraction of reads we searched for ping-pong partners for each of the 24 small RNA sequencing libraries. We detected a clear but infrequent ping-pong signature (>20 pairs, zscore >2 compared to 5-15 nt overlap) in 5 libraries (Supplemental Fig. S1C).
The genotype of any of the analyzed libraries was not expected to differentially affect piRNA metabolism, suggesting that the signature we found might be due to contaminating RNA introduced during the RNA extraction. To investigate this possibility, we performed differential expression testing of miRNAs between ping-pong positive libraries and ping-pong negative libraries, under the assumption that RNA introduced from contaminating tissues would include tissue-specific miRNAs that are not normally expressed in heads, and that should thus be absent in ping-pong negative libraries and present in ping-pong positive libraries.
A set of 27 miRNAs was significantly enriched in ping-pong positive libraries relative to ping-pong negative libraries at an adjusted P-value (Benjamini-Hochberg) of 0.01 (Supplemental Table S1). Since we analyzed male heads, contaminating RNA might stem from the testis. To investigate this possibility we randomly added reads from testicular small RNA libraries (Toledano et al. 2012; Rozhkov et al. 2010) to ping-pong negative head libraries. We find that a contamination of ~ 2% for a total library size of $2.5*10^6$ reads is sufficient to detect a ping-pong signature (Supplemental Fig. S1D). If the observed ping-pong signature was due to contamination with testicular RNA, we should find a strong overlap of differentially expressed miRNAs between between ping-pong positive and ping-pong negative libraries on the one hand, and differentially expressed miRNAs in the above simulation.
Therefore we performed differential expression testing between the ping-pong negative libraries and ping-pong negative libraries with 2% of testicular reads added. The 10 most significantly changed miRNAs found in this manner are also significantly changed (9 miRNAs with $p<0.01$, 1 miRNA with $p<0.025$; Supplemental Table S1) when comparing ping-pong positive and ping-ping negative libraries (see also Supplemental Figure 1E), confirming that the most likely origin of the ping-pong signature that we can detect is RNA contamination by testicular RNA.
The ping-pong signature is thus currently not detectable in male adult head small RNA libraries that do not show signs of contamination with gonadal small RNAs, suggesting that Aubergine and Ago3 are either not actively producing secondary piRNAs, or do so in very low amounts or in isolated cell-populations.

piRNA-sized small RNAs in adult heads are not *piwi*-dependent

Since somatic ping-pong pairs of small RNAs were not detectable in our initial analysis, we next tested whether Piwi is required for the piRNA-sized reads we observed. It was shown

before that Piwi, but not the secondary piRNA biogenesis machinery, is required for piRNA biogenesis in ovarian follicle cells, in which ping-pong pairs are not detectable (Lau et al. 2009; Robine et al. 2009; Saito et al. 2009). As the main function of Piwi has been proposed to induce transcriptional repression of TEs in ovaries (Sienski, Dönertas, and Brennecke 2012; Rozhkov, Hammell, and Hannon 2013), we focused on small RNA reads aligning to TEs. We sequenced small RNA libraries of wild type (+/+), *piwi* heterozygous (*piwi* +/-) and *piwi* homozygous (*piwi* -/-) mutant heads as well as dicer-2 helicase (dicer-2 G31R/-) and dicer-2 loss of function (*dicer-2* -/-) mutant heads.

We observed that for wild type, *piwi* heterozygous and *piwi* homozygous mutant heads the 24-28 nucleotide reads show a slight bias for aligning to the sense strand of TEs (Figure 1A, inset), that is even more pronounced in *dicer-2* mutant flies. This is in contrast to piRNAs in gonads that predominantly align to the antisense strand of TEs. Importantly, the fraction of 24-28 nucleotide antisense to TEs seemed to increase slightly in piwi homozygous mutant heads, suggesting that these reads are not *piwi*-dependent (Fig 1B).

Piwi mutation unmasks a TE-specific siRNA response

We next examined whether loss of *piwi* would affect siRNAs that target TEs in the head. As expected, in both wild type and *piwi* mutant heads we detect a substantial amount of 21 nucleotide sense and antisense reads, which appears to increase in **homozygous** *piwi* -/- mutant heads (Figure 1A, 1B). Importantly, these reads are strongly reduced in *dicer-2* mutants, confirming that these TE-aligned reads are siRNAs. We consistently observed the same trend in symmetric increase of 21 nucleotide sense and antisense RNAs in piwi homozygous mutant heads when considering TE families that had on average more than 20 aligned reads per 10 million matched reads ("cp10m"). To quantify the increase of siRNAs we only considered the fraction of 21 nucleotide reads that aligns to the TE complementary strand, in order to minimize quantification of partially degraded TE transcripts, which would be expected to align to the sense strand with little size specificity (discussed in (Malone et al. 2009)). We further restricted analysis to TE families that had on average 5 or more 21 nucleotide antisense reads per library after correction for sequencing depth. By doing so, we determined that in piwi mutants, 27 out of 62 transposable element families show >2 fold increases of siRNAs, with an overall median fold change of 1.69 compared to a 0.92 fold change for piwi heterozygous mutant heads (P-value 1.03e-05, Mann-Whitney U).

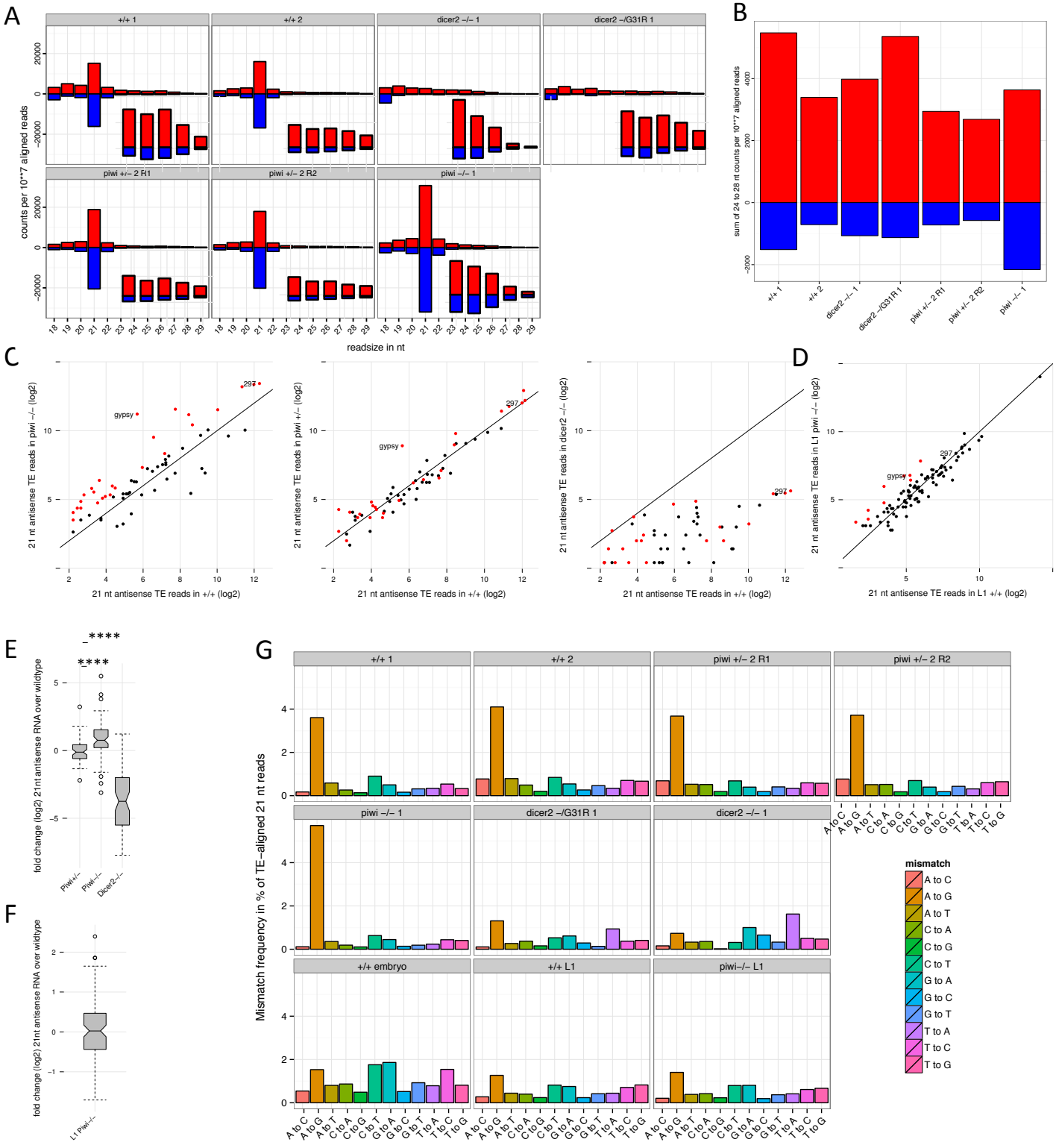


Figure 1. Loss of Piwi repression leads to an increase of TE-targeting siRNAs. (A) Overview of the size (x-axis) and amount (y-axis, in counts per 10 million mapped reads) of small RNA reads that align to TEs in adult heads of the indicated genotype. Zoom of the 24-28 nt fractions are shown as insets. (B) Sense or antisense reads of 24 to 28 nt were summed for each indicated genotype (x-axis). (C). Scatterplots displaying the abundance of 21 nt antisense reads in mutant (y-axis) and wild type (x-axis) heads. Red dots in the first panel indicate the transposon-specific 21 nt antisense reads that increased more than 2 fold in *piwi* homozygous mutant heads. These dots are shown for comparison in the second and third panel. (D) Boxplots showing the distribution of 21 nt antisense fold changes (y-axis) between wild type and the indicated mutants (x-axis). Significance of differences between the distributions was assessed with Mann-Whitney U test. (E) as (C), but for wildtype 1st instar larvae (homozygous piwi mutant vs wild type). (F) as (D), but for wild type 1st instar larvae (homozygous piwi mutant vs wild type). (G) Mismatches of 21 nt reads aligning to reference genome TE insertions with 1 mismatch allowed. Identity of mismatch is indicated on the x-axis, and the fraction of all reads with this mismatch identity over all TE matches is indicated on the y-axis. In all panels, the index of sample duplicate is indicated after the genotype when appropriate.

While most TE siRNAs were hence moderately increased, gypsy siRNA expression was increased more than 45 fold (Figure 1C).

We also investigated if the observed changes would occur at earlier time points in development. We observed a 2.8 fold increase of gypsy-specific siRNA in *piwi* -/- mutant first instar larvae as compared to +/+ first instar larvae, however the majority of TE family siRNAs remained unchanged at this stage (p-value 0.83, Mann-Whitney U, Figure 1D).

Together, these results suggested that increased TE siRNA levels upon Piwi loss is a response that is strongest in adult heads at the adult stage. In addition to the previously reported suppression of variegation of wm^4 in adult eyes and mild decrease of HP1 occupancy at TEs in 3rd instar larvae, somatic loss of Piwi also induces an increased production of TE-specific siRNAs.

The siRNA response likely originates from Dicer-2-mediated TE transcript processing in adult heads

We then investigated the origin of the increased of TE derived siRNA in *piwi* mutant. One simple hypothesis is that siRNA response might be a direct consequence of increased TE transcription upon absence of Piwi, that would in turn result in increased Dicer-2-mediated processing of TE transcript into siRNAs. Alternatively, the elevated levels of TE-siRNAs might originate from piRNA clusters, which produce both piRNAs and siRNAs in the germline. If this were to be the case we should be able to detect an increase of specific 21-nucleotide RNAs originating from piRNA clusters.

As most piRNA clusters are transcribed bi-directionally we quantified both 21 nt sense and antisense RNAs that map exclusively to piRNA clusters (defined by (Brennecke et al. 2007)). We thereby excluded sequences shared with TE insertions elsewhere in the genome, allowing us to separate production of siRNAs originating from clusters and those originating from TE insertions. We observed a low quantity of cluster-specific 21 nt reads in both *piwi* -/- and control heads (between 2% and 3,5% of total TE reads, Supplemental table 2). These cluster-derived 21 nucleotide RNAs increased slightly in piwi mutant heads (Supplemental Figure 2), the increase however was lower than the increase of 21 nt antisense TE reads (1.44 fold vs 2.15 fold, Supplemental table 2). The increase of TE-specific siRNAs in *piwi* -/- mutant heads was thus unlikely to be caused by an increase in Dicer-2-mediated processing of piRNA cluster transcripts, favoring a hypothesis in which increased transcription of euchromatic TE insertions leads to an increase in TE-specific siRNA production. In principle the increase in TE-specific siRNAs could be maternally inherited or stably maintained from early development. We took advantage of the fact that mature Ago2-loaded siRNAs are single-stranded and that stretches of base-paired RNAs, among which the substrate of

Dicer-2, are frequently deaminated through the action of adenosine deaminase acting on RNA (ADAR) enzyme, which converts adenosine (A) to inosine (I) (Keegan et al. 2005; Palladino et al. 2000; Wu, Lamm, and Fire 2011). This change manifests in a A to G mismatches in RNA sequencing datasets as compared to the reference DNA sequence. We therefore determined the frequency of all nucleotide mismatches for all FlyBase-listed TE insertions in our small RNA sequencing datasets (Figure 1E). The amount of A to G mismatches is not elevated over other mismatches in early embryos (1.5 % of all 21 nucleotide reads matching to TE insertions). We detect a similar frequency of A to G mismatches (1.2 to 1.4%) in first instar larvae, however no other mismatches were preferred, suggesting that ADAR might be active at low level in 1st instar larvae. We do detect a higher frequency of A to G mismatches in wild type heads (3.6-4.1%) that, despite a >2-fold increase of 21 nt RNA, further increases in *piwi* mutants (5.7 %), indicating that processing of the double-stranded precursor by Dicer-2 occurs after the onset of ADAR activity, suggesting that TE-targeting siRNAs are actively produced from double-stranded RNAs in adult heads and are not inherited from earlier developmental stages.

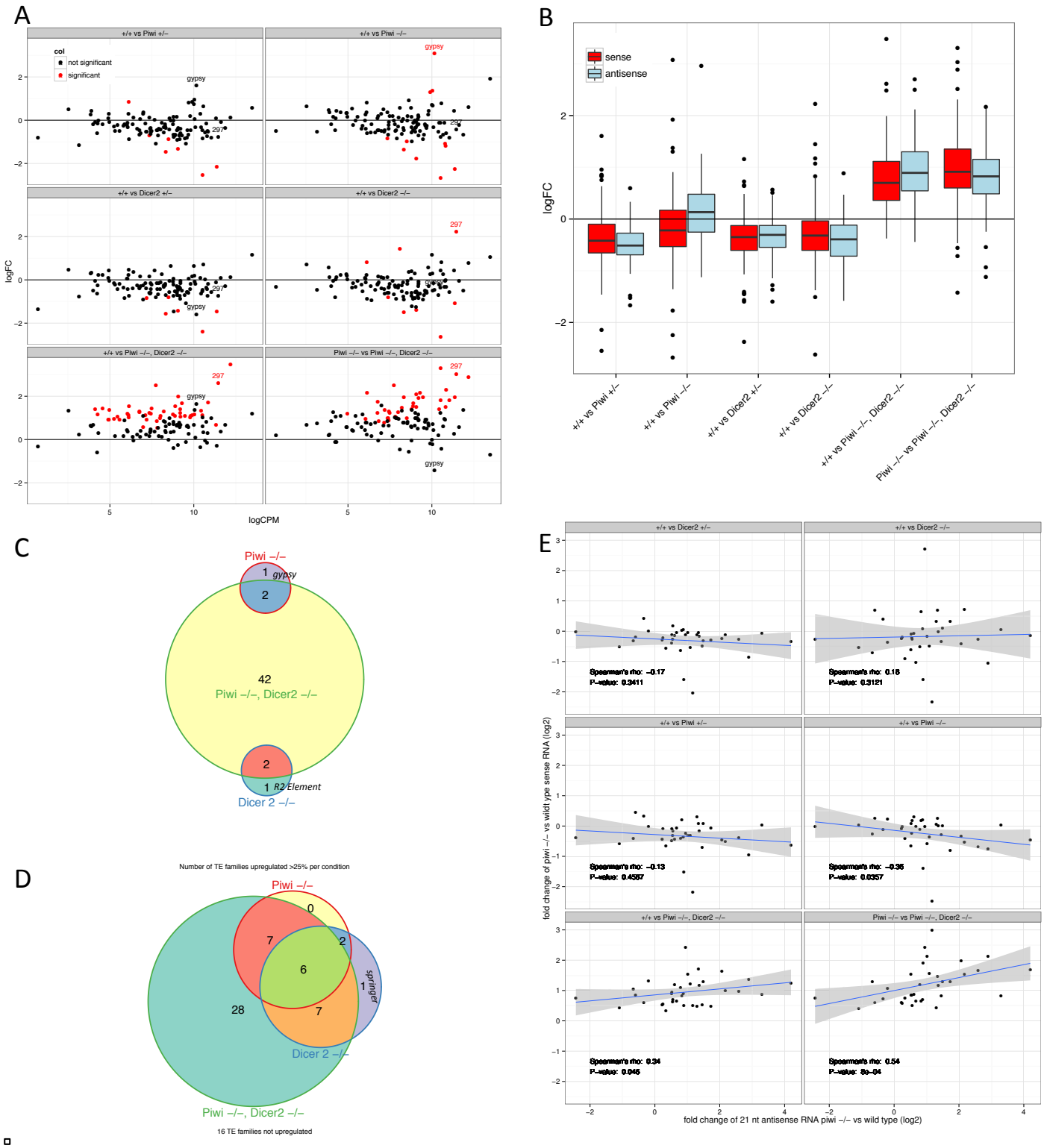


Figure 2. Piwi and Dicer-2 are complementary factors for the repression of TEs in adult heads (A) Scatterplots displaying the log2 fold changes of sense TE transcript expression on the y-axis and the mean expression strength on the x-axis. log2 fold changes were calculated between the indicated mutant and the wild type TE levels, except in the last panel where we calculate changes between the double mutant and the *piwi* mutant. (B) Boxplots showing the distribution of log2 fold changes as in (A), but considering changes of sense (red) and antisense TE transcripts (blue) separately for each comparison (x-axis). (C) Venn Diagram showing the overlap of significantly (p.adj <0.05) up regulated TEs for the indicated mutants as compared to the wild type control. (D) As (C), but taking TEs whose transcript abundance increases more than 25% over the wild type. (E) Scatterplot displaying the correlation between log2 fold changes of 21 nt antisense RNA in *piwi* homozygous mutant heads compared to wild type heads on the x-axis and log2 fold changes of sense TE transcripts for the genotype comparisons indicated above each panel. Only TEs that passed a threshold of on average five 21 nucleotide antisense reads over all small RNA libraries were analysed. Futher, only TEs whose sense transcript level increased in *piwi*, *dicer-2* while insensitive to *piwi* loss are shown. A similar Scatterplot including all TEs that passed the siRNA threshold can be found in Supplemental Figure 2. The blue line is a fit produced by the lmfit function, and the grey area delimits the corresponding confidence interval.

Neither loss of Piwi nor Dicer-2 leads to strong upregulation of TEs

To determine whether the observed increase of siRNA production efficiently counteracts any increased TE transcription caused by a lack of Piwi-mediated TGS, we sequenced the head transcriptome of *piwi* mutants, *dicer-2* mutants and *piwi, dicer-2* double-mutants and compared these to wild type head transcriptome. In piwi mutants, transcript levels of most TEs remain unchanged, with the notable exception of *gypsy,* whose level increases about 5-fold (Figure 2A). *gypsy* is also the TE against which we observed the strongest increase of siRNA levels, suggesting that the transcription of TEs is indeed increased in *piwi* mutant heads and correlates with siRNA production.

Similarly, most TEs are are not ureguted in *dicer-2* mutants, except for *297*, which produces a significant amount of siRNA in wild type heads, perhaps indicating inefficient Piwi-mediated TGS for this TE family. We conclude that *piwi* and *dicer-2* are redundant in adult heads for the maintenance of TE repression for most TE families and that Piwi-mediated TGS and Dicer-2-mediated PTGS both can efficiently repress TEs.

Piwi and Dicer-2 compensatory mechanism revealed in double-mutant

To confirm our hypothesis that the siRNA response compensates for *piwi* loss, we analyzed RNA libraries from *piwi, dicer-2* double-mutant heads. We find that the majority of TE families is significantly upregulated in *piwi, dicer-2* double-mutant heads compared to single *piwi* or *dicer-2* mutants (Figure 2 A,B, C).

Furthermore we detect a significant positive correlation of siRNA production changes and increase in steady-state TE transcripts in *piwi, dicer-2* double-mutant heads (Supplemental Figure 2)**.** By excluding the TE families that are upregulated already in piwi mutant heads, we also find a significant negative correlation between the siRNA production and the TE transcript level, highlighting the efficiency of the siRNA response (Figure 2E). In agreement, we did not detect any significant increase of sense TE transcripts in piwi mutant heads, but we did detect a significant increase in antisense transcripts (Figure 2B), that might form transient duplexes with sense TE transcripts, serving as a substrate for Dicer-2-mediated siRNA processing.

Together this suggests a double layer expression control of Piwi-mediated TGS and Dicer-2-mediated PTGS to firmly repress TE levels in the adult soma.

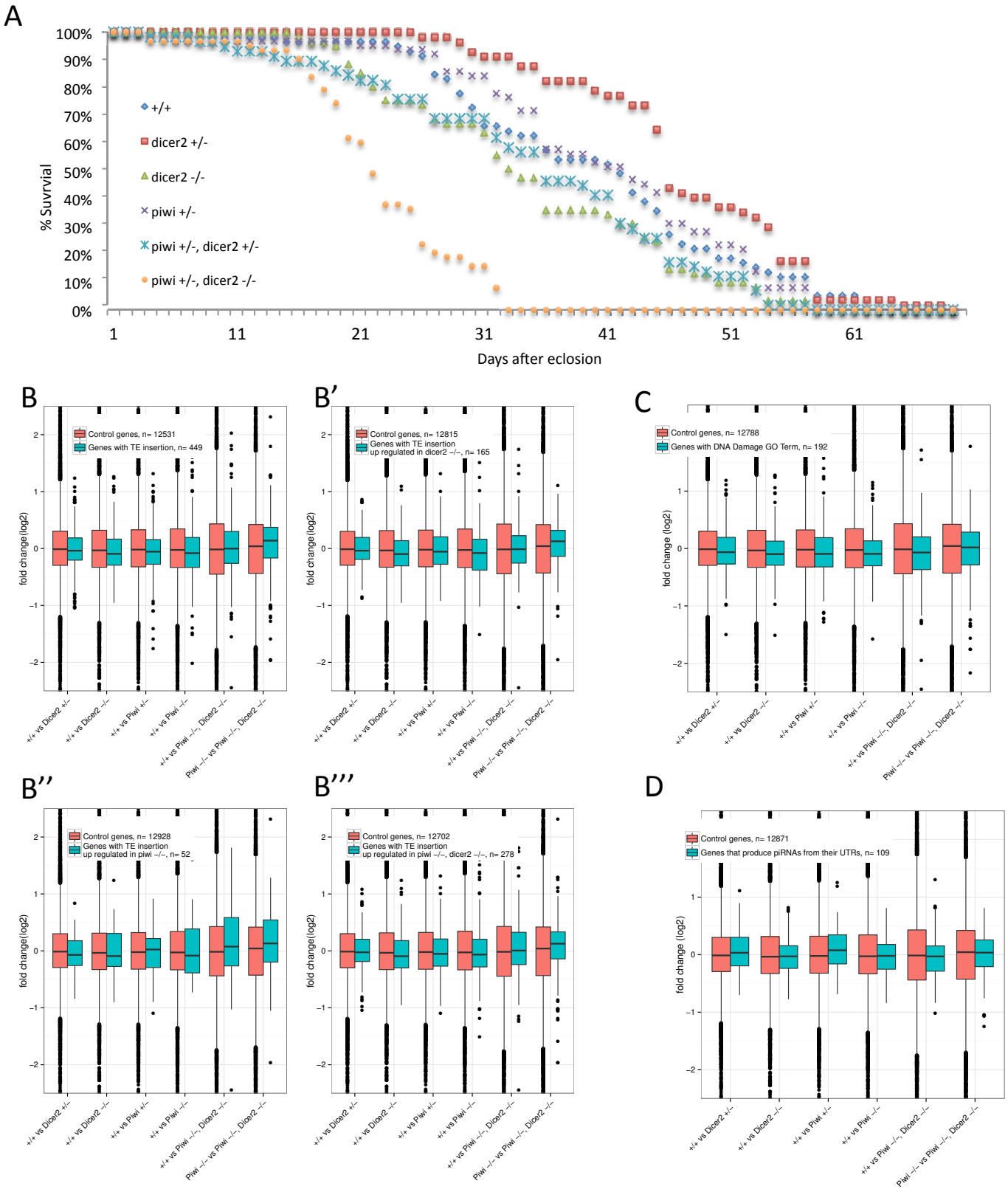


Figure3. *piwi*, *dicer*-2 mutant have a decreased lifespan, independently of previously reported *piwi*-regulated genes. (A) Lifespan assay indicates that *piwi* +/-, *dicer2* -/- mutants have a reduced lifespan as compared as compared to wild type flies. (B) Boxplots of fold change for genes with reference genome TE insertions in their genomic boundaries do not suggest a trend for these genes to become de-repressed across any of the tested mutant condition (x-axis). We compared genes without TE insertions (red boxes) with genes with insertion of TE of any family (blue boxes, B), or with TE insertions that are up regulated in *piwi* homozygous (blue boxes, B'), *dicer2* homozygous (B'') or *piwi*, *dicer-2* double homozygous mutants (B'''). Boxplots of the distribution of fold changes for genes with DNA Damage Gene ontology terms (C) and genes that have been reported to host piRNA production from their UTRs (D).

Piwi and Dicer-2 do not restrict expression of genes that contain TEs in their genomic loci

Piwi has previously been shown to silence genes adjacent to TE insertions and genes that carry TEs in their boundaries (Sienski, Dönertas, and Brennecke 2012). Our transcriptomic data does not support a significant trend of upregulation of genes that contain TE insertions in their genomic boundaries, whether we consider all TE families, or only those that are upregulated in *piwi*, *dicer-2* or *piwi*, *dicer-2* double-mutants (Figure 3A). Piwi has also been shown to repress genes that produce *traffic jam* class piRNAs from their 3'UTR (Robine et al. 2009). Again, we detect no expression bias for these genes in any of our mutant conditions.

A hallmark of TE activation in ovarian piwi mutations is DNA-damage and the resulting increase of DNA-damage signalling is easily detectable in transcriptome sequencing. We do not find an upregulation of the DNA damage signaling pathway, suggesting that either TE mobilization is not as severe as in the germline, or that an increased DNA damage response occurs with aging (we sequenced 1 day old adults).

The lifespan of *piwi*, *dicer-2* double-mutants is severely decreased

To test whether the fitness of *piwi*, *dicer-2* double-mutants would be affected, we tested their survival. *Piwi* and *dicer-2* mutations do not affect lifespan strongly while the lifespan of *piwi*, *dicer-2* double-mutants is shorter than any of the other conditions tested (Figure 3A). This might be due to an increased load of TE transcription and deleterious TE mobilization, the effect of misregulated gene expression (1490 genes differentially expressed exclusively in *piwi*, *dicer-2* double-mutants) or an unknown genetic interaction between piwi and the siRNA pathway. Additional efforts are required to determine the cause of the observed lifespan reduction.

Discussion

Origin and function of piRNA-like molecules in somatic tissues

Through the analysis of a large number of small RNA sequencing libraries we are questioning previous reports of the presence the secondary piRNA biogenesis signature in adult heads (Yan et al. 2011; Ghildiyal et al. 2008; Mirkovic-Hösle and Förstemann 2014). Based on our analysis it is likely that these observations result from a small contamination with gonadal RNA, that can easily occur when collecting heads by vortexing frozen flies and filtering fly-parts by sieving. A contamination with gonadal RNA would be in agreement with

the miRNA signature we detected between head libraries that correlate with the presence of ping-pong signature. Our results also highlight the importance of controlling the purity of RNA preparations. Considering the enormous increase of TE expression in piRNA mutant ovaries, often reaching more than 1000 fold upregulation, best practice for analysis should be a control that clearly shows the degree of tissue specificity of the RNA preparation. We show that we cannot detect the secondary piRNA biogenesis pattern in heads, which is a very sensitive readout. This is not completely unexpected, since Perrat and colleagues show that Aubergine and Ago3 do not colocalize extensively, though both are expressed in gamma-neurons. While this is not a proof for the absence of secondary piRNA amplification, we would not expect this process to contribute much to the global repression of TEs in the adult head. We further show that the 24-28 nucleotide fraction of small RNA reads is not sensitive to loss of Piwi. This means that either these piRNA-like molecules are primary piRNAs produced independently of Piwi or that they are degradation products. Their sense bias and the low activity of piRNA clusters in heads indicates that most likely these piRNA-like molecules are degradation products.

Piwi likely exerts its function in adult heads through chromatin compaction

Piwi has been known to act as a suppressor of variegation, providing a first link between the piRNA pathway and heterochromatin formation (Pal-Bhadra et al. 2004). A recent report by the Elgin laboratory (Gu and Elgin 2013) confirmed this and also showed that Piwi mutant 3rd instar larvae had slightly decreased levels of HP1 chromatin binding over some, but not all TE sequences. Gu and Elgin also showed that eye-lineage specific knockdown of HP1 but not Piwi suppresses variegation. Importantly, the eye-lineage-specific promoter that was used is active only after late embryogenesis, suggesting that Piwi-mediated TGS can be maintained in the absence of Piwi expression after early embryogenesis and that Piwi might not be present in the adult eye. These findings are in line with our results as we did not see a reduction of piRNA-sized molecules antisense to TEs, we did not detect significant levels of piwi RNA and we did not succeed in detecting Piwi by Western blotting (data not shown). We thus favor a hypothesis wherein Piwi exerts TGS during early zygotic development, but cannot exclude the possibility that very low amounts of Piwi protein, with or without loaded piRNAs, would remain in the proximity of the chromatin to continuously repress TEs by a TGS mechanism. Of note, we found that loss of *piwi* has little effect on steady state RNA levels for most TE families in our data, while depletion of su(var-3-9) and Histone H1 in salivary glands of 3rd instar larvae was shown to lead to much higher TE levels (15-1000 fold) (Lu et al., 2013). Together with reports of early embryonic lethality of embryos laid by mothers that were depleted for Piwi in the oocyte (Cox et al. 1998; Wang and Elgin 2011)

this might indicates that, while we abolish zygotic expression of Piwi, maternal inheritance of Piwi might be sufficient to setup a baseline level of TGS. This is likely to be less strong than what can be achieved in wild type flies and hence it is possible that we are underestimating the epigenetic effect that Piwi-mediated TGS might have.

Repression of TEs in the absence of siRNAs

We found that *dicer-2* mutations caused a strong loss of TE-specific siRNAs, but this loss did not result in major changes of TE expression for most TE families in adult heads, with the exception of *297*, whose expression increased about 4.8 fold compared to the wild type or 7.8 fold compared to the *dicer-2* heterozygote library. This suggests that transcriptional repression by *piwi* is highly efficient at restricting TE expression, but also that some TEs can escape repression by Piwi. We currently do not know why 297, but not other TEs can escape efficient TGS by Piwi, but our results are in line with the results of Gildyal et al., who found *297* expression to be strongly increased in heads of *dicer-2* mutants, and Xie et al., who demonstrated increased somatic transposition of 297 in *dicer-2* mutants.

A dual layer-repression by small RNAs

As defects in the repression of TEs in the soma might cause severe fitness disadvantages, TE expression should be under tight control. We demonstrate here that loss of Piwi alone does not lead to strong changes in TE expression for most TE families, and neither does loss of Dicer-2, while double-mutants showed increased expression of TEs across a large panel of TE families. Loss of Piwi is however accompanied by increased production of siRNAs and antisense TE RNA, which suggests that the transcription of TE insertion increases and leads to a feedback loop in which Dicer-2 processes double-stranded RNA-molecules to produce siRNA-duplexes that in turn efficiently reduce steady-state levels of sense TEs.

The double-stranded RNA molecules that are substrates for Dicer-2 could originate from secondary structures of the TE transcript or alternatively from pairing of sense transcripts with antisense transcripts or piRNA cluster transcripts or from promoter-proximal RNAs that are produced by erroneous transcription initiation. We consider the latter to be unlikely, as we would expect strong enrichments of siRNA production near the transcriptional start of TEs, but visual inspection does not support elevated levels of promoter-proximal siRNA mapping. It is a mystery to us why we observe a symmetric increase of sense- and antisense siRNAs in piwi mutants, yet the level of TE antisense transcripts appears to increase significantly. We would expect the increased accumulation of both transcript

senses to be symmetrical in piwi mutants or to favor sense transcription, since in general TEs produce higher levels of sense than antisense transcripts. One explanation could be that siRNA-mediated PTGS occurs mainly through single-strand siRNA loaded in Ago2. Dicer-2-mediated processing of sense and antisense transcripts into 21nt RNAs leads to symmetric amount of 21nt sense and antisense RNAs, but only antisense RNAs efficiently degrade sense RNA, while for an unknown reason, sense siRNAs do not reduce steady-state levels of antisense TE transcripts. This might be a mechanism to ensure a steady production of siRNA duplexes, as siRNAs that efficiently degrade antisense RNAs also limit the amount of double-strand transcripts that can be processed into siRNAs.

Altogether, we propose a model in which Piwi ensures transcriptional silencing guided by piRNAs. This repression is strong and maintained independently of Piwi. TE-specific siRNAs and TE transcripts can nevertheless be detected in wild type fly heads. We propose that one function of the siRNA pathway is to prevent accumulation of sufficient levels of TEs that lead to transposition events, by targeting TE transcripts through PTGS. This might be especially important for TEs that can escape Piwi-mediated repression, such as 297, or TEs that do not produce piRNAs as they have not yet been integrated into a piRNA cluster. In evolutionary terms, this failsafe mechanism provides an advantage to both TE and host: a TE that causes sterility of it's host will not propagate to the next host generation. We propose that in somatic tissues the siRNA pathway recognizes the invading TE, either through stretches of homologous sequences that pair with endogenous transcripts, or more akin to viral infections, through the secondary structure of the TE transcripts. siRNA-mediated PTGS then maintains a tolerable load of the TE, so that the organism can survive until the TE has integrated into a piRNA cluster and developing germ cells survive until the point fertility is restored, hence benefiting both TE and host.

Decreased life-span due to transposition or gene expression changes?

Compared to *piwi* knockdown in gonadal tissues (Rozhkov, Hammell, and Hannon 2013) or OSC cells (Sienski, Dönertas, and Brennecke 2012), somatic TE expression increases are rather weak even in the double-mutant. Increases in the double-mutant rather resemble the increases found in *piwi*-mutants defective for nuclear localization of Piwi. In general, whether a TE will be transcribed in any given cell depends both on the ability of the set of available transcription factors to support transcription from the TE's promoter and the strength by which its expression is counteracted. This implies that the net contribution to the steady-state RNA level is difficult to detangle. In addition, zygotic loss of Piwi will probably not fully reveal the effect that Piwi has on adult somatic tissue homeostasis (maternal depletion results in early embryonic arrest of development), but still allow us to understand its

epigenetic function. We demonstrate increased TE expression in *piwi, dicer-2* double-mutants, which and itthis correlates with shortened life-span, raising the possibility that increased TE transcript levels lead to harmful transposition events. While we did not attempt to quantify increased transposition in this paper, it will be an important future step to confirm the relevance of the complementarity of piRNA-guided TGS and siRNA-mediated PTGS that we described in this paper. We note that we have performed all RNA-sequencing in heads of 1 day young adult male flies, and expression changes of TEs might be more significant in aged mutant individuals. We further did not find any evidence for the activation of DNA damage signalling pathways, a signature very abundant in RNA-sequencing of piwi mutants ovaries. We can however not rule out that DNA damage occurs later in aged individuals, or that only few cells accumulate DNA damage by transposition, but that loss of few cells in the adult brain leads to the observed lethality.

Material and Methods

Stocks and fly husbandry

Flies

Flies were grown on standard *Drosophila* food at 25°C. All flies were brought into the w^{m4} background (Muller 1930). dicer-2^{R416X} and dicer-$2^{L811fsx}$ alleles were described in (Lee et al. 2004). $piwi^2$ and $piwi^3$ alleles were published in (Cox et al. 1998). Double-mutants were generated by crossing virgin female $dcr2^{R416X}$/CyO-GFP to male $piwi^3$/CyO-GFP flies. Offspring virgin $dcr2^{R416X}/piwi^3$ flies were then crossed to male wm4;Ln^{2R} Gla, wgGla^1, Bc^1/ CyO-GFP to establish wm4;$piwi^2$, dicer-2^{R416X}/CyO-GFP stocks. Stocks were then screened by PCR for the presence of the *piwi* mutation. The same procedure was applied to generate wm4;$piwi^3$, dicer-$2^{L811Fsx}$/CyO-GFP stocks. The table below provides the detailed genotype of all mutant combinations used.

Indicated name	Full genotype	Maternal genotype	Paternal genotype
+/+	wm4;+/+	wm4;+/+	wm4;+/+
dicer-2 +/-	wm4;$dcr2^{L811Fsx}$/CyO-GFP	wm4;$dcr2^{L811Fsx}$/CyO-GFP	wm4;+/+
dicer-2 -/-	wm4;$dcr2^{L811Fsx}/dcr2^{R416X}$	wm4;$dcr2^{L811Fsx}$/CyO-GFP	wm4;$dcr2^{R416X}$/CyO-GFP
piwi +/-	wm4;$piwi^2$/+	wm4;$piwi^2$/CyO-GFP	wm4;+/+
piwi -/-	wm4;$piwi^2/piwi^3$	wm4;$piwi^2$/CyO-GFP	wm4;$piwi^3$/CyO-GFP

dicer-2, piwi -/+	wm4;piwi^{2}, dicer-2^{R416X}/CyO-GFP	wm4;piwi^{2}, dicer-2^{R416X}/CyO-GFP	wm4;+/+
dicer-2, piwi -/-	wm4;piwi^{2}, dicer-2^{R416X}/piwi^{3}, dicer-2^{L811Fsx}	wm4;piwi^{2}, dicer-2^{R416X}/CyO-GFP	wm4;piwi^{3}, dicer-2^{L811Fsx}/CyO-GFP

RNA extraction and sequencing

One to two day old flies were anesthetized, sorted by sex and genotype, transferred into 15 ml Falcon tubes and frozen in liquid nitrogen. The procedure was repeated multiple days until pools of 50 to 100 flies were obtained per biological replicate. Heads were separated from bodies by vortexing, followed by sieving and selecting heads from splintered thoraces and legs on a cooled metal plate. Heads were collected into 2 ml Precellys tubes for hard tissues and covered by 1ml Trizol. Heads were homogenized in two rounds of 5000 rpm for 30 seconds using a Precellys24 Tissue Homogenizer. Homogenate was centrifuged for 30 seconds at 13000 rpm and supernatant transferred into a new 2 ml tube, 200µl of Chloroform was added and tubes were thoroughly vortexed. Further purification was as in (Rio et al. 2010). Remaining DNA was removed using Fermentas DNase I, RNase-free following the manufacturers instructions.

Small RNA library preparation and sequencing was performed on an Illumina HiSeq 2500 at Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) using the *Drosophila* small RNA track based on the Illumina TruSeq protocol.

RNA-seq was performed in biological triplicates for +/+, *piwi* -/+, *piwi* -/-, *dicer-2* -/+ and *dicer-2* -/-, with one replicate per condition sequenced in paired-end mode (2 * 101) and two replicates sequenced in single-read mode (1*51). *piwi* -/-, *dicer-2* -/- samples were sequenced in biological duplicates in single-read mode. Total RNA was depleted of rRNA using Ribo-Zero™ Gold Kit (Epicentre). Directional RNA-seq library preparation and sequencing was performed at the Genomic Paris Centre (Paris, France) using the Epicentre ScriptSeq™ v2 RNA-Seq Library Preparation Kit on an Illlumina HiSeq 2000 instrument.

Computational Analysis

The complete computational analysis pipeline was run on our in-house Galaxy server. All necessary workflows and tools will be publicly available at http://mississippi.fr/galaxy .

All small RNA libraries were quality controlled, sequencing adapter-clipped and converted to fasta reads. All reads that aligned to ribosomal RNA were discarded. All small RNA alignments were done using bowtie 0.12.7, allowing 1 mismatch between sequenced read and reference sequence (Langmead et al. 2009). To produce Supplemental Figure 1, fasta

reads were aligned to the *Drosophila* genome (FlyBase release 5.49) (St Pierre et al. 2014), randomly placing reads that align equally well in multiple genomic locations (multimapper) using the bowtie option "-M 1". Size distribution and ping-pong signature were calculated using the mississippi toolsuite (https://testtoolshed.g2.bx.psu.edu/view/drosofff/mississipi_toolsuite_beta). The ping-pong signature was calculated by counting the number of pairs that overlap between 5 to 15 nucleotides between sense- and antisense aligned reads and transforming the obtained counts into z-scores (each count subtracted by the mean and divided by the standard deviation). Ping-pong positive libraries were selected by having a z-score higher than 2 and more than 20 pairs overlapping by 10 nucleotides. Ping-pong negative libraries were selected by having a negative z-score. To obtain a list of differentially expressed miRNA between ping-pong positive and ping-pong negative libraries reads were matched to the *Drosophila* pre-miRNAs of the miRBase 20 release (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008; Kozomara and Griffiths-Jones 2011; Kozomara and Griffiths-Jones 2014). Differential expression profiling between ping-pong positive and ping-pong negative libraries was performed using edgeR_3.8.2 (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012) with standard settings. For simulating contamination with testis RNA 2 testis-libraries (accessions SRX135547, SRX023726) were downsampled to 10 million reads, pooled and 50000 randomly selected reads were added to 2.45 $*10^6$ randomly selected reads from ping-pong negative libraries. piRNA signature was calculated as before. Differential miRNA expression was calculated between simulated libraries and ping-pong negative libraries of equal size (randomly downsampled to 2.5 * 10^6), with libraries that were sampled from the same initial ping-pong negative library paired as a blocking factor. This allows for obtaining an accurate list of miRNAs (contamination signature) that should be expected to be to significantly change in abundance if a contamination occurred. Size distribution for small RNAs that align to TEs (Figure1A, 1B) was calculated from reads that matched any of the canonical TE sequences with 1 mismatch allowed, excluding reads that matched to ribosomal RNA, tRNA or abundant insect viruses. Abundance of 21 nt antisense RNA for each TE family was calculated by filtering reads to 21 nt length and aligning reads to canonical TE sequences, allowing only unique reads using the bowtie option "-v 1". Only antisense reads were counted, and only TEs with on average 20 reads per library were analyzed. Between-library normalized 21 nucleotide antisesene TE counts were obtained by pooling these with miRNA reads (obtained as before) and calculating a normalization factor using the DESeq (Anders and Huber 2010) function estimateSizeFactors. log2 fold changes were calculated by dividing normalized reads of mutants by normalized reads of controls and taking the logarithm. Difference of the population of log2 fold changes were tested using a two-tailed Mann-

Whitney-U test. To calculate mismatch frequencies for 21 nucleotide small RNA, ribosomal, non-coding RNA and viral reads were filtered out. Remaining reads were aligned to the collection of TE insertions (FlyBase version 5.49), allowing 1 mismatch. Each possible mismatch was counted and divided by the total number of 21 nucleotide reads aligned to the collection of TE insertions.

For gene expression profiling, reads were quality-filtered using the FASTX toolkit with a Quality cut-off of 30 for 90% of the read. For the paired-end libraries only the R1 read of the pair was used and trimmed to 51 nucleotides. Reads were then aligned to the *Drosophila* genome release 5 (dm3) using Tophat2 (Kim et al. 2013). Default parameters were used, except that we supplied Gene Model annotations from UCSC genome browser for dm3 (http://support.illumina.com/sequencing/sequencing_software/igenome.html). Read counting was performed using featureCounts (Liao, Smyth, and Shi 2014) guided by the aforementioned Gene Model file.

For TE expression profiling reads were further trimmed to 30 nucleotides and aligned to canonical TEs using bowtie 0.12.7, allowing 2 mismatches and only uniquely matching reads. Sense and antisense reads were counted and merge with gene counts. Differential expression profiling was performed using edgeR (Robinson, McCarthy, and Smyth 2010; McCarthy, Chen, and Smyth 2012). Genes with less than 5 reads on average across libraries were discarded from the analysis. Diverging from the default, we used Full Quantile between-library normalization as implemented by the EDAseq package (Risso et al. 2011) and removed unwanted variation using replicate samples with the RUVs function (choosing k=2) implemented in the RUVseq package (Risso et al. 2014). Library sequencing method (paired-end vs. single-read) was introduced together with gene-wise Full Quantile normalization offsets and gene-wise RUVs offsets as covariates in the edgeR design formula. All libraries were tested for differential gene expression against the wild type, and in addition the double-mutant was also tested against the piwi -/- mutant. Proportional Venn Diagrams in Figure 2C and Figure 2D were drawn using the Vennerable package (http://r-forge.r-project.org/projects/vennerable). The spearman rank correlation and corresponfing P-value between log2 fold changes in TE 21 nucleotide antisense RNA (log2 fold change calculated from data underlying Figure 1C) and sense TE transcript expression was calcuated with the rcorr function in the Hmisc R package (http://cran.r-project.org/web/packages/Hmisc/). All graphs were plotted using ggplot2 (http://ggplot2.org/). GO terms for DNA damage and genes with TE insertions were retrieved from FlyBase (St Pierre et al. 2014).

Survival test

60 flies of the indicated genotype were split in 6 tubes of 5 males and 5 females. Tubes were flipped every second day and dead flies were counted.

References

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106.

Brennecke, Julius, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. 2007. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila.” *Cell* 128 (6): 1089–1103.

Brower-Toland, Brent, Seth D Findley, Ling Jiang, Li Liu, Hang Yin, Monica Dus, Pei Zhou, Sarah C R Elgin, and Haifan Lin. 2007. “Drosophila PIWI Associates with Chromatin and Interacts Directly with HP1a.” *Genes & Development* 21 (18): 2300–2311.

Cox, D N, A Chao, J Baker, L Chang, D Qiao, and H Lin. 1998. “A Novel Class of Evolutionarily Conserved Genes Defined by Piwi Are Essential for Stem Cell Self-Renewal.” *Genes & Development* 12 (23): 3715–27.

Czech, Benjamin, Colin D Malone, Rui Zhou, Alexander Stark, Catherine Schlingeheyde, Monica Dus, Norbert Perrimon, et al. 2008. “An Endogenous Small Interfering RNA Pathway in Drosophila.” *Nature* 453 (7196): 798–802.

Ghildiyal, Megha, Herve Seitz, Michael D Horwich, Chengjian Li, Tingting Du, Soohyun Lee, Jia Xu, et al. 2008. “Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells.” *Science* 320 (5879): 1077–81.

Griffiths-Jones, Sam. 2004. “The microRNA Registry.” *Nucleic Acids Research* 32 (Database issue): D109–11.

Griffiths-Jones, Sam, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. 2006. “miRBase: microRNA Sequences, Targets and Gene Nomenclature.” *Nucleic Acids Research* 34 (Database issue): D140–4.

Griffiths-Jones, Sam, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. 2008. “miRBase: Tools for microRNA Genomics.” *Nucleic Acids Research* 36 (Database issue): D154–8.

Gu, Tingting, and Sarah C R Elgin. 2013. “Maternal Depletion of Piwi, a Component of the RNAi System, Impacts Heterochromatin Formation in Drosophila.” *PLoS Genetics* 9 (9): e1003780.

Gunawardane, Lalith S, Kuniaki Saito, Kazumichi M Nishida, Keita Miyoshi, Yoshinori Kawamura, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. 2007. “A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5’ End Formation in Drosophila.” *Science* 315 (5818): 1587–90.

Keegan, Liam P, James Brindle, Angela Gallo, Anne Leroy, Robert A Reenan, and Mary A O’Connell. 2005. “Tuning of RNA Editing by ADAR Is Required in Drosophila.” *The EMBO Journal* 24 (12): 2183–93.

Kim, Daehwan, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. 2013. “TopHat2: Accurate Alignment of Transcriptomes in the Presence of Insertions, Deletions and Gene Fusions.” *Genome Biology* 14 (4): R36.

Kozomara, Ana, and Sam Griffiths-Jones. 2011. “miRBase: Integrating microRNA Annotation and Deep-Sequencing Data.” *Nucleic Acids Research* 39 (Database issue): D152–7.

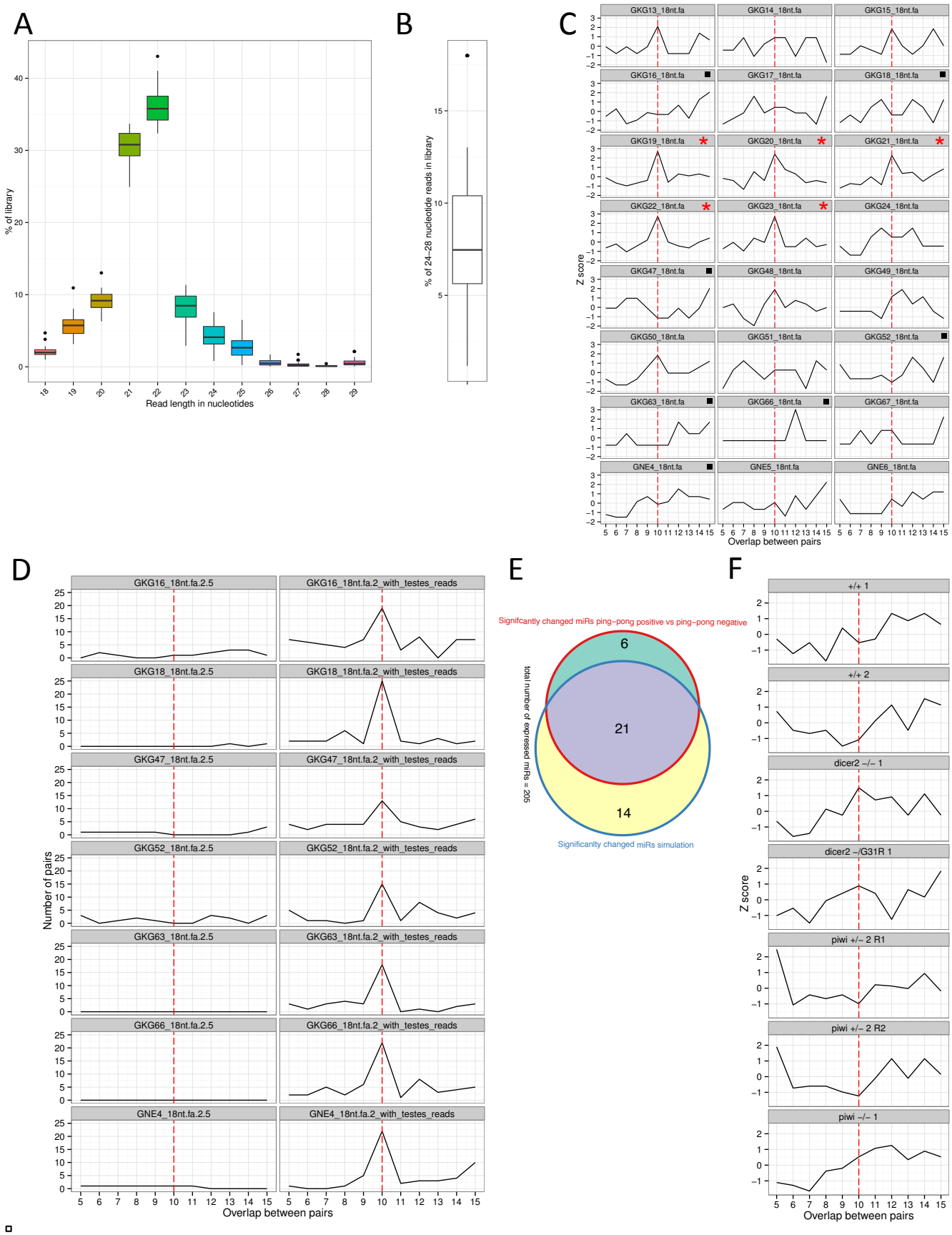
———. 2014. “miRBase: Annotating High Confidence microRNAs Using Deep

Sequencing Data." *Nucleic Acids Research* 42 (Database issue): D68–73.
Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L Salzberg. 2009. "Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome." *Genome Biology* 10 (3): R25.
Lau, Nelson C, Nicolas Robine, Raquel Martin, Wei-Jen Chung, Yuzo Niki, Eugene Berezikov, and Eric C Lai. 2009. "Abundant Primary piRNAs, Endo-siRNAs, and microRNAs in a Drosophila Ovary Cell Line." *Genome Research* 19 (10): 1776–85.
Lee, Young Sik, Kenji Nakahara, John W Pham, Kevin Kim, Zhengying He, Erik J Sontheimer, and Richard W Carthew. 2004. "Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways." *Cell* 117 (1): 69–81.
Li, Wanhe, Lisa Prazak, Nabanita Chatterjee, Servan Grüninger, Lisa Krug, Delphine Theodorou, and Josh Dubnau. 2013. "Activation of Transposable Elements during Aging and Neuronal Decline in Drosophila." *Nature Neuroscience* 16 (5): 529–31.
Liao, Yang, Gordon K Smyth, and Wei Shi. 2014. "featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features." *Bioinformatics* 30 (7): 923–30.
Malone, Colin D, Julius Brennecke, Monica Dus, Alexander Stark, W Richard McCombie, Ravi Sachidanandam, and Gregory J Hannon. 2009. "Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary." *Cell* 137 (3): 522–35.
Mani, Sneha Ramesh, Heather Megosh, and Haifan Lin. 2014. "PIWI Proteins Are Essential for Early Drosophila Embryogenesis." *Developmental Biology* 385 (2): 340–49.
McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." *Nucleic Acids Research* 40 (10): 4288–97.
Mirkovic-Hösle, Milijana, and Klaus Förstemann. 2014. "Transposon Defense by Endo-siRNAs, piRNAs and Somatic pilRNAs in Drosophila: Contributions of Loqs-PD and R2D2." *PloS One* 9 (1): e84994.
Moshkovich, Nellie, Parul Nisha, Patrick J Boyle, Brandi A Thompson, Ryan K Dale, and Elissa P Lei. 2011. "RNAi-Independent Role for Argonaute2 in CTCF/CP190 Chromatin Insulator Function." *Genes & Development* 25 (16): 1686–1701.
Muller, H J. 1930. "Types of Visible Variations Induced by X-Rays inDrosophila." *Journal of Genetics* 22 (3). Springer India: 299–334.
Nagao, Akihiro, Toutai Mituyama, Haidong Huang, Dahua Chen, Mikiko C Siomi, and Haruhiko Siomi. 2010. "Biogenesis Pathways of piRNAs Loaded onto AGO3 in the Drosophila Testis." *RNA* 16 (12): 2503–15.
Pal-Bhadra, Manika, Boris A Leibovitch, Sumit G Gandhi, Madhusudana Rao, Utpal Bhadra, James A Birchler, and Sarah C R Elgin. 2004. "Heterochromatic Silencing and HP1 Localization in Drosophila Are Dependent on the RNAi Machinery." *Science* 303 (5658): 669–72.
Palladino, M J, L P Keegan, M A O'Connell, and R A Reenan. 2000. "dADAR, a Drosophila Double-Stranded RNA-Specific Adenosine Deaminase Is Highly Developmentally Regulated and Is Itself a Target for RNA Editing." *RNA* 6 (7): 1004–18.
Perrat, Paola N, Shamik DasGupta, Jie Wang, William Theurkauf, Zhiping Weng, Michael Rosbash, and Scott Waddell. 2013. "Transposition-Driven Genomic Heterogeneity in the Drosophila Brain." *Science* 340 (6128): 91–95.
Rangan, Prashanth, Colin D Malone, Caryn Navarro, Sam P Newbold, Patrick S Hayes, Ravi Sachidanandam, Gregory J Hannon, and Ruth Lehmann. 2011. "piRNA Production Requires Heterochromatin Formation in Drosophila." *Current Biology: CB* 21 (16): 1373–79.
Rio, Donald C, Manuel Ares Jr, Gregory J Hannon, and Timothy W Nilsen. 2010. "Purification of RNA Using TRIzol (TRI Reagent)." *Cold Spring Harbor Protocols* 2010 (6): db.prot5439.

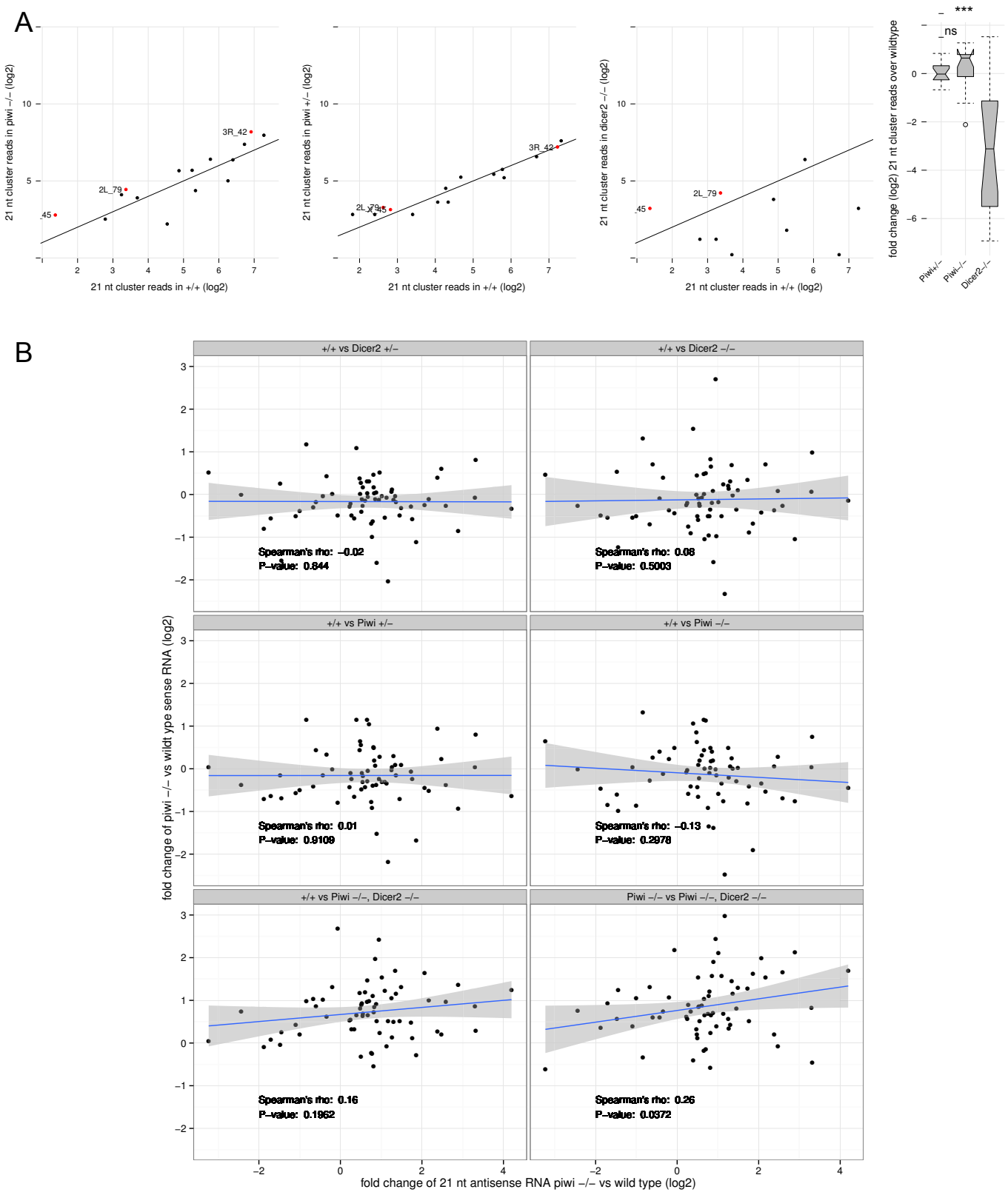
Risso, Davide, John Ngai, Terence P Speed, and Sandrine Dudoit. 2014. "Normalization of RNA-Seq Data Using Factor Analysis of Control Genes or Samples." *Nature Biotechnology* 32 (9): 896–902.
Risso, Davide, Katja Schwartz, Gavin Sherlock, and Sandrine Dudoit. 2011. "GC-Content Normalization for RNA-Seq Data." *BMC Bioinformatics* 12 (December): 480.
Robine, Nicolas, Nelson C Lau, Sudha Balla, Zhigang Jin, Katsutomo Okamura, Satomi Kuramochi-Miyagawa, Michael D Blower, and Eric C Lai. 2009. "A Broadly Conserved Pathway Generates 3′UTR-Directed Primary piRNAs." *Current Biology: CB* 19 (24): 2066–76.
Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics* 26 (1): 139–40.
Rouget, Christel, Catherine Papin, Anthony Boureux, Anne-Cécile Meunier, Bénédicte Franco, Nicolas Robine, Eric C Lai, Alain Pelisson, and Martine Simonelig. 2010. "Maternal mRNA Deadenylation and Decay by the piRNA Pathway in the Early Drosophila Embryo." *Nature* 467 (7319): 1128–32.
Rozhkov, Nikolay V, Alexei A Aravin, Elena S Zelentsova, Natalia G Schostak, Ravi Sachidanandam, W Richard McCombie, Gregory J Hannon, and Michael B Evgen'ev. 2010. "Small RNA-Based Silencing Strategies for Transposons in the Process of Invading Drosophila Species." *RNA* 16 (8): 1634–45.
Rozhkov, Nikolay V, Molly Hammell, and Gregory J Hannon. 2013. "Multiple Roles for Piwi in Silencing Drosophila Transposons." *Genes & Development* 27 (4): 400–412.
Saito, Kuniaki, Sachi Inagaki, Toutai Mituyama, Yoshinori Kawamura, Yukiteru Ono, Eri Sakota, Hazuki Kotani, Kiyoshi Asai, Haruhiko Siomi, and Mikiko C Siomi. 2009. "A Regulatory Circuit for Piwi by the Large Maf Gene Traffic Jam in Drosophila." *Nature* 461 (7268): 1296–99.
Sienski, Grzegorz, Derya Dönertas, and Julius Brennecke. 2012. "Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression." *Cell* 151 (5): 964–80.
St Pierre, Susan E, Laura Ponting, Raymund Stefancsik, Peter McQuilton, and FlyBase Consortium. 2014. "FlyBase 102--Advanced Approaches to Interrogating FlyBase." *Nucleic Acids Research* 42 (Database issue): D780–8.
Taliaferro, J Matthew, Julie L Aspden, Todd Bradley, Dhruv Marwha, Marco Blanchette, and Donald C Rio. 2013. "Two New and Distinct Roles for Drosophila Argonaute-2 in the Nucleus: Alternative Pre-mRNA Splicing and Transcriptional Repression." *Genes & Development* 27 (4): 378–89.
Thomas, Adrien Le, Alicia K Rogers, Alexandre Webster, Georgi K Marinov, Susan E Liao, Edward M Perkins, Junho K Hur, Alexei A Aravin, and Katalin Fejes Tóth. 2013. "Piwi Induces piRNA-Guided Transcriptional Silencing and Establishment of a Repressive Chromatin State." *Genes & Development* 27 (4): 390–99.
Toledano, Hila, Cecilia D'Alterio, Benjamin Czech, Erel Levine, and D Leanne Jones. 2012. "The Let-7-Imp Axis Regulates Ageing of the Drosophila Testis Stem-Cell Niche." *Nature* 485 (7400): 605–10.
Tubio, Jose M C, Yilong Li, Young Seok Ju, Inigo Martincorena, Susanna L Cooke, Marta Tojo, Gunes Gundem, et al. 2014. "Extensive Transduction of Nonrepetitive DNA Mediated by L1 Retrotransposition in Cancer Genomes." *Science* 345 (6196): 1251343.
Vagin, Vasily V, Alla Sigova, Chengjian Li, Hervé Seitz, Vladimir Gvozdev, and Phillip D Zamore. 2006. "A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline." *Science* 313 (5785): 320–24.
Wang, Sidney H, and Sarah C R Elgin. 2011. "Drosophila Piwi Functions Downstream of piRNA Production Mediating a Chromatin-Based Transposon Silencing Mechanism in Female Germ Line." *Proceedings of the National Academy of Sciences* 108 (52): 21164–69.
Wu, Diane, Ayelet T Lamm, and Andrew Z Fire. 2011. "Competition between ADAR and RNAi Pathways for an Extensive Class of RNA Targets." *Nature Structural & Molecular*

Biology 18 (10): 1094–1101.
Xie, Weiwu, Ryan C Donohue, and James A Birchler. 2013. “Quantitatively Increased Somatic Transposition of Transposable Elements in Drosophila Strains Compromised for RNAi.” *PloS One* 8 (8): e72163.
Yan, Zheng, Hai Yang Hu, Xi Jiang, Vera Maierhofer, Elena Neb, Liu He, Yuhui Hu, et al. 2011. “Widespread Expression of piRNA-like Molecules in Somatic Tissues.” *Nucleic Acids Research* 39 (15): 6596–6607.

Supplemental Material



Supplemental Figure 1. *Drosophila* small RNA head libraries have variable amount of piRNA-sized reads and ping-pong signature in heads, which correlates with a contamination signature. (A) Boxplot showing the distribution of read lengths for the libraries analyzed in (C). (B) similar to (A), but for the fraction of 24-28 nucleotide length reads. (C) Tendency for small RNAs to overlap. The number of overlaps was transformed to Z scores to take into account the "cleanness" of the 10 nucleotide overlap as compared to other lengths of overlaps. Libraries with red asterisks were selected as ping-pong positive, and black squares indicate libraries selected as ping-pong negative libraries. (D) Ping-pong signature for $2.5*10^6$ ping-pong negative libraries with (right group of panels) and without (left panels) the addition of 2% of a testis-library. Number of pairs are shown instead of Z scores, as some down-sampled libraries had 0 overlapping pairs. (E) Venn Diagram showing the overlap of differentially expressed miRNAs between ping-pong positive compared to ping-pong negative libraries (red circle) and ping-pong negative libraries compared to ping-pong negative libraries with the addition of 2% of a testis library. (F) ping-pong signature for small RNA libraries analyzed in the remainder of the article.



Supplemental Figure 2. (A) Scatterplots displaying the abundance of cluster-derived 21 nt reads in mutant (y-axis) and wild type (x-axis) heads. Red dots in the first panel indicate the cluster-specific 21 nt antisense reads that increased more than 2 fold in *piwi* -/- mutant heads. These dots are shown for comparison in the second and third panel. (D) Boxplots showing the distribution of 21 nt read fold changes (y-axis) between wild type and the indicated mutants (x-axis). Significance of differences between the distributions was assessed with Mann-Whitney U test. (B) Related to Figure 2E. Scatterplot displaying the correlation between log2 fold changes of 21 nt antisense RNA in *piwi* homozygous mutant heads compared to wild type heads on the x-axis and log2 fold changes of sense TE transcripts for the genotype comparisons indicated above each panel. All TE families that passed a threshold of on average five 21 nucleotide antisense reads over all small RNA libraries were analysed. The blue line is a fit produced by the lmfit function, and the grey area delimits the corresponding confidence interval.

a. Conclusion and future work

The above article represents a first draft of my work, and has to be changed in a number of key points prior to publication.

[1.] Most importantly, there is a replicate missing for the small RNAs in *piwi* -/- mutant heads. This is due to an intermediate hypothesis we had after the Publication of Gu and Elgin (Gu and Elgin 2013) and the fact that we started both small RNA sequencing and RNA-sequencing without replicates; As we saw equally strong increases of *gypsy* transcript levels in piwi +/- mutant RNA-seq (without replicate) and by qRT-PCR, we decided to focus on the effect that a reduction of piwi-dose though maternal contribution would have (we sequenced the offspring of piwi +/- mothers crossed to wild type fathers). Therefore we replicated only wild type and *piwi* +/- libraries, for which, at the time, we did not have any library at all. However, we did not see the general trend for all TEs to have higher siRNA levels in piwi +/- mutant heads as compared to wild type heads, and qRT-PCR between *piwi* +/-; *dicer-2* -/- and *piwi* -/-; *dicer-2* -/- flies showed that TE levels were further increased when zygotic expression of piwi was abolished.

[2.] Furthermore I could harden the hypothesis of testicular contamination in heads by testing whether testis-specific *suppressor of stellate* (Su(ste)) small RNAs correlate with the piRNA signature.

[3.] The manuscript could be further improved by investigating whether the observed A to G mismatches are specific to certain classes of TEs. The same is true for the increases seen in the different mutants, e.g. is there a class of TEs that responds preferentially to loss of Piwi or loss of Dicer-2? Does the number of known reference insertions correlate with TE expression in any of the mutants?

[3.] Another interesting observation is that *gypsy* appears to evade repression by Dicer-2, while *297* appears to evade repression by *piwi*. In fact, *gypsy* levels decrease in *dicer-2* -/- compared to +/+, and *piwi*, *dicer-2* -/- compared to *piwi* -/-. What could be the origin of this observation?
A very interesting hypothesis for the evasion of *297* to transcriptional repression is linked to the proposed biological function of Piwi-mediated transcriptional repression and siRNA-mediated post-transcriptional repression as redundant layers involved in TE defense. *297*, while being an exception in the above manuscript, might recapitulate what happens when a new TE invades a species. If fhe new TE is mobile it likely causes extensive DNA damage,

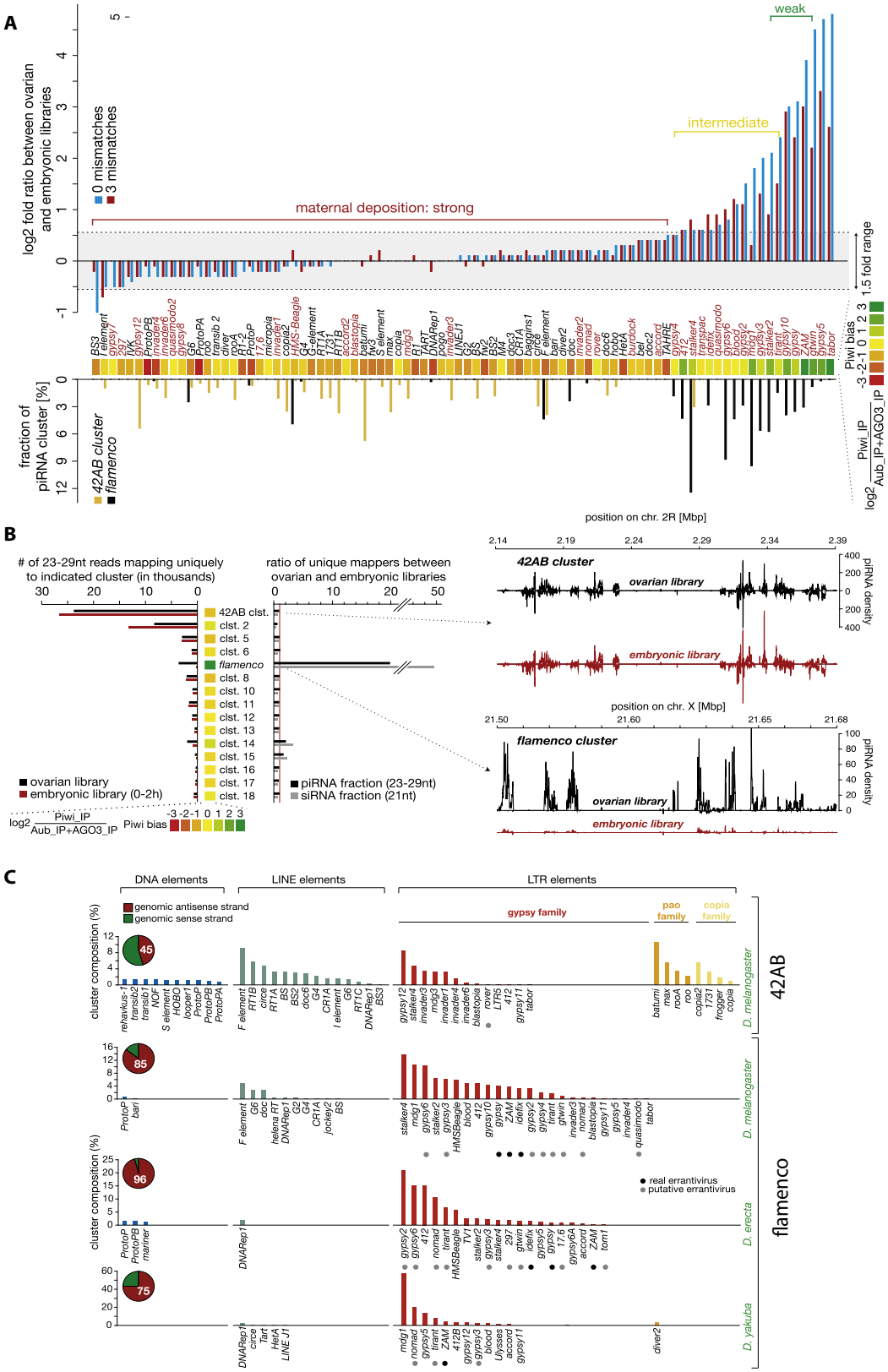


Figure 7: Maternal deposition of piRNAs based on TE family and piRNA cluster. (A) The tendency for small RNA to be maternally deposited is depicted, based on the relative abundance between ovaries and embryonic small RNA libraries. Small RNAs of TE famliies that are enriched in embryos (log2 ratio above 1) are likely to be deposited by the mother. (B) similar to (A) but for piRNA clusters. Note the bias for *297* and 42AB small RNAs to be maternally deposited, while gypsy and flamenco small RNAs appear to not be maternally deposited. (C) TE composition of 42AB and flamenco. Note the absence of 297 from 42AB and flamenco in the *Drosophila melanogaster* species group. Adapted from (Malone et al., 2009).

eventually leading to reproductive arrest. The presence of TE RNA in the cell might be sufficient to induce a siRNA response, as occurs during viral infections (*297* is classified as a putative errantivirus). This siRNA response, while not sufficient to completely clear the new TE, might be sufficient to prevent early lethality, allowing the TE to jump into a piRNA cluster and henceforth to be targeted by piRNAs.

The question then is why *297* can escape piwi-mediated TGS in the adult soma, even though ping-pong piRNAs are produced against *297*? It appears that *297* fragments are absent from the somatic *flamenco* cluster in the *Drosophila Melanogaster,* but not in the *Drosophila Erecta* group. Perhaps the deposited 297 ping-pong piRNAs are not capable to induce transcriptional silencing on *297* insertions. It is generally believed that small RNAs are deposited bound to an Argonaute-class protein, and *297* piRNAs are strongly enriched in Aubergine and Ago3, which have not been demonstrated to be invlolved directly in transcriptional repression. Lack of TGS on 297 could therefore be explained by the absence of *297* piRNAs in Piwi. This is in agreement with the finding by (Xie et al., 2014) who found 297 to be fixed in the reference genome strain (10+ years after the sequencing project), but that *297* can transpose in somatic cells of *dicer-2* -/- mutants.

gypsy on the other hand is present in the *flamenco* cluster, and appears to be efficiently repressed by Piwi also in *dicer-2* -/- mutants. A straightforward explanation would be that dicer-2 -/- mutants are devoid of functional *gypsy* insertions. This is however not likely, as *gypsy* is not strongly derepressed in *piwi* -/-, *dicer-2* -/- mutants, which were the result of recombination between *piwi* - and *dicer-2* – chromosomes.

Elucidating the underlying biology of these evasion phenoma clearly deserves further attention.

[4.] A further question that I did not address is whether piwi and/or dicer-2 do play roles in splicing and/or transcriptional pausing in the soma?

Dicer-2 had previously been implicated in promoter-proximal pausing at the Hsp70 locus (Cernilogar et al. 2011).

[5.] Finally, the significance of the manuscript could be further elevated if we could explain the decreased lifespan of *piwi* -/-, *dicer-2* -/- flies. One hypothesis put forward in the paper is that there is a constant increase of somatic transposition that ultimately leads to a decrease in life span. We would therefore need to test whether the increase in TE levels that we have seen corresponds to an increase of transposition with age.

If we were to assay this on a genome-wide scale, we could also test whether there is a correlation between the presence of a TE insertion in proximity a gene and its increase in abundance in *piwi* mutants, as (Sienski, Dönertas, and Brennecke 2012) have done for Piwi

knockdown in OSC cells. In fact, it appears that non-reference genome TE insertions had stronger effects on neighboring gene expression (Annex 2).

b. Methods for detecting TE insertion events

It would be very beneficial if we could assess whether increased TE transcript levels in our studied mutants really result in increased somatic transposition, and whether this happens during aging. Such hypothetical transposition events are expected to occur at random places in single cells, making it difficult to detect these in the DNA isolated from a large cell population, such as wings or heads.

Multiple methods have been developed to follow transposition events, such as the gypsy-TRAP reporter by Li et al. (Li et al., 2013), Southern blots, transposon-display, or fluorescent in-situ hybridizations and more recently high-throughput sequencing.

Southern blotting is the oldest of these methods, but likely lacks sensitivity for somatic transposition events that occurred in a, supposedly, small fraction of cells. A newer development is the transposon-display technique, where genomic DNA is digested, ligated with a specific adapter and amplified in one or two rounds of PCR with primers specific to the ligated adapter and an outwards pointing TE primers (Waugh et al., 1997, Melayah et al., 2001), and new insertions are identified by amplified fragment length polymorphisms (AFLPs).

In my case it would hence be possible to isolate wing DNA after eclosion (wing removal does not significantly affect lifespan in *Drosophila*), to identify a base-level of TE insertion heterogeneity in young adults, as well as DNA from aged heads, to determine whether the level of somatic TE insertions increases on a per genotype basis with aging.

I have tried the gypsy-TRAP reporter (Li et al., 2013), where a gypsy landing-site separates the promoter of gal80 cassette from its coding sequence. If gypsy integration did not occur, gal80 is active and represses and gal4-driver activity. If gypsy integrates at the landing site, gal80 transcription is shut off and a gal4 reporter is activated. I could not detect transposition events in aged adult brain as was described in the accompanying paper. This lack of transposition detection may be due to the fact that I used a different UAS-GFP and gal4 driver from what was used in the paper. In any case, due to the choice of a preferential gypsy integration site from the ovoD locus, this sensor is limited to *gypsy* insertion events.

Fluorescent *in situ* hybridization (FISH) constitutes a viable alternative to detect single cell TE insertion events for salivary gland cells isolated from 3rd instar larvae, as demonstrated by Xie and Birchler (Xie and Birchler, 2014). This is however laborious, as it requires good knowledge of polytene chromosomes and can only be performed for 2 TE families at a time. In addition, it would not be suitable for analysing the differences in transposition activity

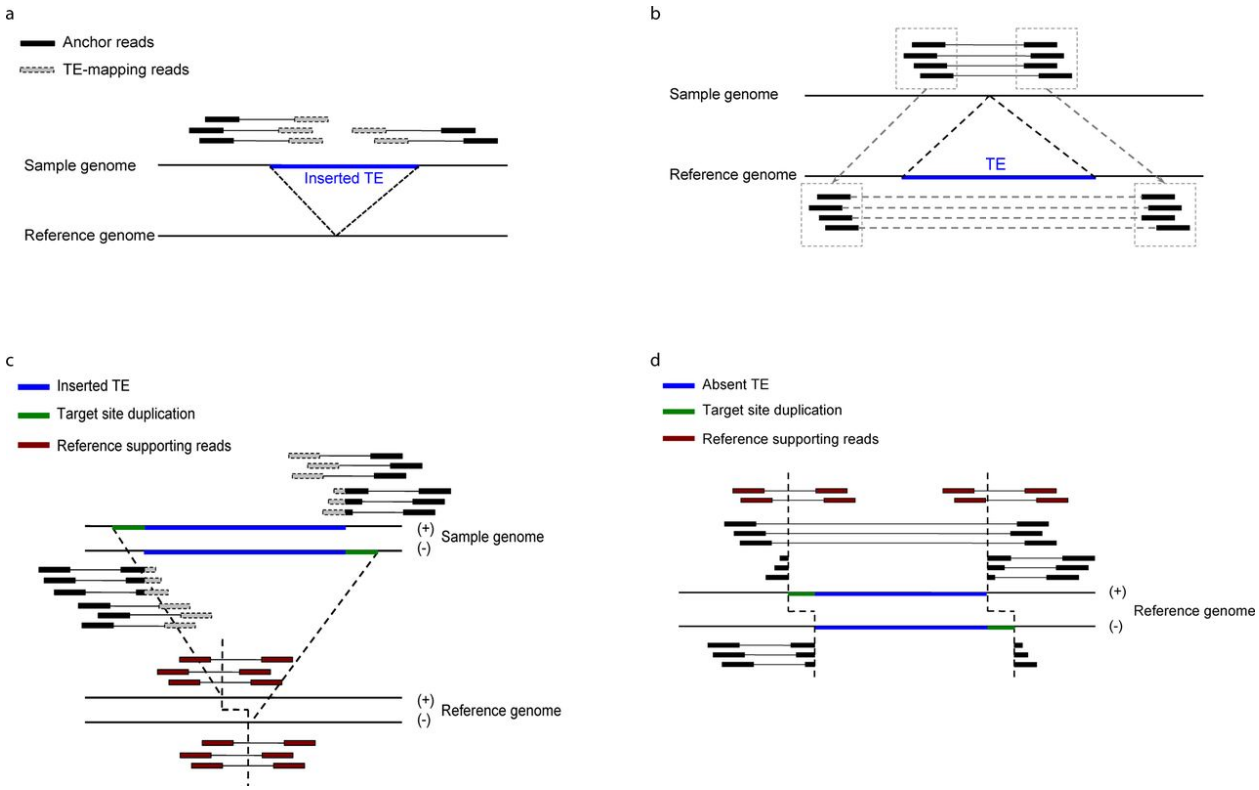
during aging. Alternatively, the procedure might be adapted to work on adult tissues, but positional information such as available in polytene chromosomes would be lost and only total fluorescence per cell could be evaluated as a surrogate for copy number alterations for the specific TE under study.

The newest and likely most powerful, but also most expensive method to detect TE insertions is genome re-sequencing. Multiple methods exist to detect new TE insertions of known TE family members in genome-resequencing data. A powerful approach to determine position and frequency of new TE insertion is to align reads, preferentially from paired-end libraries with large insert sizes, to the known genome (Zhuang et al. 2014). Reads that do not span a new TE insertion will align with little to no mismatches to the genome, and the genomic distance of the mate pairs will correspond approximately to the insert size of the library (~300-600 nucleotides for Illumina DNA sequencing libraries). Variations will arise when the sequenced sample contains a non-reference TE insertion (Figure 8a), or when a reference TE insertion is absent in the sequenced sample (Figure 8b). Based on reads that either span insertion or deletion sites, or reads that disparately map to a non-TE sequence and a TE sequence, one can determine the insertion (or deletion) site. By taking into account the local coverage and reads that align perfectly to the reference genome at the same position one can also determine the frequency of the insertion in the studied sample.

To confidently detect somatic transposition events from a pool of cells a sufficient sequencing coverage is needed to identify all homozygous (detection frequency near 100% for a given locus) and heterozygous (detection frequency near 50%) TE insertions in an individual, as those likely reflect TE insertions inherited through the parental germline. To classify human variants in exome sequencing data as heterozygous, an allele frequency between 14% and 86% with a minimum coverage of 20X has been suggested as criterium(Heinrich et al. 2012; Bell et al. 2011). For an estimation of somatic TE insertion frequency in Drosophila it is likely sufficient to sequence to 50X depth.

For *Drosophila melanogaster*, whose haploid genome size is about $122*10^6$ nucleotides, a 50X coverage for a single sample with paired-end reads and insert sizes of about 500 nucleotides would require a sequencing depth of $61*10^8/500=12*10^6$ reads, or about 6% of a lane on a HiSeq2000 sequencer, assuming an output of $2*10^8$ reads per lane and equal coverage. To distinguish the impact of *piwi* mutation, *dicer-2* mutation and *piwi, dicer-2* double mutation on transposition rate, we should sequence young and old wild type samples, as well as young and old sample from all mutations, preferably in replicates. For 3 replicates per genotype and time-point we would need to sequence 24 samples, requiring 2 lanes on a HiSeq2000, which would, at current, cost about 400 euros, excluding library preparation, which adds significant costs as well.

Figure 8: Diagrams depicting presence (a) and absence (b) of TEs and how the integration site of these can be identified at base-pair resolution for presence (c) and absence (d) of insertion events. Taken from (Zhuang et al. 2014)



Alternatively, we could reduce the cost of the analysis using Restriction site associated DNA-sequencing (RADseq). In RADseq genomic DNA is digested with a restriction enzyme, barcoded adapters are ligated to the restricted DNA, followed by DNA shearing to obtain typical NGS insert size of about 300 nucleotides. The sheared DNA is then end-repaired and the opposite sequencing primer is annealed. RADseq allows the interrogation of a smaller selection of genomic sites with higher sequencing depth. In addition protocols have been developed in which the DNA is digested with a 6-bp cutter and a more frequent 4-bp cutter (ddRADseq), omitting the lossy steps of DNA shearing and end-repair, as the restriction enzymes can be chosen to yield NGS library compatible insert sizes. The higher efficiency of ligation to sticky restriction digest ends also allows using lower amounts of starting material, but the double digestion protocol might lead to preferential loss of loci with or without TE insertions, making estimates of TE insertion frequency more difficult.

2. Article: Isolation of Small Interfering RNAs Using Viral Suppressors of RNA Interference

Published as

Van den Beek, Marius, Christophe Antoniewski, and Clément Carré. 2014. “Isolation of Small Interfering RNAs Using Viral Suppressors of RNA Interference.” *Methods in Molecular Biology* 1173: 147–55.

Isolation of small interfering RNAs using viral suppressors of RNA Interference.

Marius van den Beek, Christophe Antoniewski and Clément Carré.

Drosophila Genetics and Epigenetics ; Université Pierre et Marie Curie 9, Quai St Bernard
Building C – 5th floor – Room 517 ; 75252 Paris cedex 05 ; Phone: +33 1 44 27 34 39

E-mail : clement.carre@snv.jussieu.fr

Abstract

The *tombusvirus* P19 VSR (viral suppressor of RNA interference) binds siRNAs with high affinity, whereas the *Flockhouse Virus* (FHV) B2 VSR binds both long double stranded RNA (dsRNA) and siRNAs. Both VSRs are small proteins and function in plant and animal cells. Fusing a Nuclear Localization Signal (NLS) to the N-terminus shifts the localization of the VSR from cytoplasmic to nuclear, allowing researchers to specifically probe the subcellular distribution of siRNAs, and to investigate the function of nuclear and cytoplasmic siRNAs. This Chapter provides a detailed protocol for the immunoprecipitation of small interfering RNAs (siRNAs) bound to epitope-tagged VSR and subsequent analysis by 3'-end-labelling using cytidine-3',5'-bis phosphate ([5'-^{32}P]pCp) and northern blotting.

Key words: RNAi, siRNA, endo-siRNA, Immunoprecipitation, Viral supressor of RNAi, B2, P19, RNA purification, pCp labeling, RNA detection, *Drosophila*.

1. Introduction

Small interfering RNAs (siRNAs) are implicated in a variety of processes such as Transposable Elements (TE) repression[1], maintenance of pericentric heterochromatin[2] and antiviral defense[3]. In *Drosophila,* the biogenesis of siRNAs starts with the cleavage of a long double-stranded RNA (dsRNA) precursor by the Dicer-2 endonuclease into a 21 nucleotide RNA duplex structure with 2nt 3'OH overhangs[4]. Following cleavage, the siRNA duplex is loaded into an Argonaute (Ago) containing RNA-induced silencing complex, where the duplex is unwound and the strand bearing the thermodynamically more stable 5' end (passenger strand) is cleaved by the central RISC protein Argonaute (Ago2 in flies)[5,6]. The remaining strand (guide strand) is 2'-O-methylated by Hen1 at its 3' end and guides the recognition and subsequent cleavage of complementary single-stranded target RNAs[7].

As antiviral RNAi limits viral replication, many viruses evolved suppressors of RNAi[8] (VSRs). Depending on the VSR, RNAi suppression may occur through the binding of long dsRNA substrates and/or siRNA duplexes (B2, P19, DCV1A), thereby limiting substrate availability for Dicer and Argonautes, or through the direct inhibition of Argonaute proteins (VSR-1A of Cricket Paralysis Virus[9]). In general, VSRs that act through binding to dsRNA or siRNAs are active in both plants and animals.

The *tombusvirus* P19 VSR is a 19kD protein that forms a head-to-tail homodimer and localizes to the cytoplasm when expressed as a transgene in *Drosophila* S2 cells and salivary glands[2]. P19 binds 21 nucleotide dsRNAs[10,11], thereby suppressing siRNA-mediated Post-Transcriptional Gene Silencing (PTGS) in plants[12], insects[13] and mammalian cells[14,15], leading to a de-repression of endogenous siRNA (endo-siRNA) targeted transposable elements. We have shown that fusing P19 to the NLS of the transformer (tra) protein efficiently re-localizes it to the nucleus[2].

FHV B2 is 12kD in size and forms a four-helix bundle that binds to one face of an A-form RNA duplex, independent of its length, thereby both limiting the processing of long dsRNA by Dicer and siRNA duplex incorporation into RISC[16]. Additionally, B2 has been reported to bind to the conserved PAZ domain of Dicer family proteins[17]. B2 immunoprecipitates efficiently long dsRNA, but not siRNAs, likely due to the inhibition of siRNA biogenesis. When expressed in *Drosophila* salivary glands B2 localizes to nucleoli, the nucleoplasm and the cytoplasm[2].

Through the immunoprecipitation of a nuclear-targeted P19 we were able to pull down siRNAs that localize to the nucleus without prior biochemical fractionation and to compare them to siRNAs that were pulled down by immunoprecipitating cytoplasmic P19 or the dsRNAs that co-precipitate with B2[2]. RNAi against mRNAs is not suppressed by the NLS-P19 transgene, indicating that the bulk of siRNA-mediated PTGS is not occurring inside the nucleus. Instead, a redistribution of H3K9me2 and Heterochromatin Protein-1 (HP1) is observed when expressing NLS-P19 but not unmodified P19. This mirrors the redistribution of repressive chromatin marks observed in mutants of the RNAi pathway and links endo-siRNAs to the

maintenance of chromatin organization[2]. This also highlights the potential for nuclear-engineered VSRs in dissecting the contribution of RNAi to heterochromatin maintenance.

Here we describe the use of the V5-tagged VSRs P19, NLS-P19 and B2 to immunoprecipitate bound siRNAs for experiments such as 3'-end-labelling using [5'-^{32}P]pCp, northern blot detection or high-throughput sequencing (see Antoniewski. C chapter: *Computing siRNA and piRNA overlap signatures* in this book).

2. Materials

2.1 Immunoprecipitation.

1. Mouse Monoclonal anti-V5 (Invitrogen, cat. no R960-25) and/or rat anti-HA High Affinity antibodies (Roche, cat. no 11867423001).
2. Gammabind-G sepharose (GE Healthcare).
3. Phosphate Buffered saline (PBS) (10X stock): 1.37M NaCl, 27 mM KCl, 100 mM Na_2HPO_4. Adjust to pH 7.4 with HCl; autoclave. Store at room temperature.
4. Lysis buffer: 50 mM Tris-HCl pH 7.5, 150mM NaCl, 2.5 mM $MgCl_2$, 250 mM sucrose, 0.05% Nonidet P-40, 0.5% Triton X-100. Before use, adjust to 1 mM Dithiothreitol (DTT), 1 x protease inhibitor mixture cocktail (Roche). Store at 4°C. Before use add 40U per ml of RNase inhibitor.
5. RNase inhibitor: RNase OUT (Invitrogen).
6. Wash buffer: 50mM Tris-HCl pH 7.5, 150mM NaCl, 2.5mM $MgCl_2$, 250mM sucrose, 0.05% Nonidet P-40, 0.5% Triton X-100. Store at 4°C. Before use add 1 mM Dithiothreitol (DTT), 1 x protease inhibitor mixture cocktail and 40U per ml of RNase OUT.
7. RNA Loading Dye (2X) (New England Biolabs). Store at -20°C.
8. 2X Laemmli Buffer: 4% SDS, 20% glycerol, 10% 2-mercaptoethanol, 0.004% bromophenol blue, 0.125 M Tris-HCl pH 6.8 . Store at -20°C.
9. 100 mM $CuSO_4$ solution in distilled water. Store at 4°C.
10. Rotating wheel.

2.2 RNA extraction & pCp labelling.

1. Nuclease free water.
2. Isopropanol.
3. Chloroform.
4. Phenol/Chloroform/Isoamyl alcohol pH 4.5
5. 80% Ethanol (in RNase-free water).
6. TRIzol reagent (Sigma).
7. RNA carrier (glycogen or home made linear acrylamide).
8. 3M NaAcetate (NaAc), pH 5.2.
9. T4 RNA ligase (Roche).
10. pCp (Cytidine 5′-triphosphate disodium salt), [5'-^{32}P]- 3000Ci/mmol 10 mCi/ml (PerkinElmer,).
11. Dimethyl sulfoxide (DMSO).
12. G50 MicroSpin columns (GE Healthcare).
13. 15 % acrylamide denaturing gel: 1X TBE buffer (89 mM Tris-borate, 2 mM EDTA), 15% acrylamide/bisacrylamide (19:1), 7 M urea, TEMED and 10% APS fresh solution.

3. Methods

The use of epitope-tagged P19, NLS-P19 and B2 to immunoprecipitate small RNAs was described by us and others (see introduction). Depending on the amount required for the small RNA analysis one might use transient transfection or establish a stable cell line. To analyse small RNAs by northern blotting or ^{32}P-pCp 3' end-labelling, transient transfection are sufficient and take less time. To immunoprecipitate large amounts of small RNAs for deep-sequencing or to detect low-abundant RNAs, stable transfections are more suitable but require more time. We provide instructions for the immunoprecipitation from transient and stable transfections (Note 1). Entry clones and expression vectors may be obtained from the authors and were described previously by Fagegaltier et al.[2].

3.1 Immunoprecipitation of small RNAs bound to epitope-tagged VSR in *Drosophila* S2 cells.

3.1.1. Using transient expression of the VSRs.

1. Transfect 4 µg of pMT-DEST48 control plasmid, pMT-B2-V5, pMT-P19-V5, or pMT-NLSP19-V5 in 3*10^6 S2 cells with Effectene Reagent (Invitrogen) according to the manufacturer's instructions.
2. Depending on the number of conditions tested, cells can be split 2 days after transfection and cultured for an additional two days.
3. Induce construct expression with 500 µM $CuSO_4$ for 24 hours.
4. Equilibrate 70 µl of beads for 10 min in lysis buffer at 4°C on a rotating wheel, 15 rpm.
5. Harvest cells, wash twice in cold 1 x PBS and lyse on ice for 30 min in 1 ml of lysis buffer.
6. Centrifuge at 14000 rpm for 15 min at 4 °C to pellet insoluble cell debris.
7. Transfer 10% of supernatant to a new microcentrifuge tube. This will be the input sample for the analysis of the IPs.
8. Pre-clear the extract: Transfer the remaining supernatant to a new tube and add 20 µl of equilibrated beads (from step 4). Incubate for 1 h at 4 °C on a rotating wheel, 15 rpm.
9. Centrifuge the pre-clearing mix for 5 min at 4 °C, 800 rpm and transfer the cleared supernatant to a new microcentrifuge tube.
10. Add 5 µg of mouse anti-V5 antibody (Note. 3) and 50 µl of fresh equilibrated Gammabind-Plus resin slurry (from step 4) to the pre-cleared supernatant and incubate overnight at 4 °C on a rotating wheel (15 rpm).
11. Centrifuge the samples at 4 °C and 800 rpm for 5 min.
12. Keep the supernatant as the unbound fraction.
13. Wash the beads 5 times in 1 ml of wash buffer (Note. 2).

Continue with section 3.2.1, *Purification of immunoprecipitated RNAs*.

3.1.2. Immunoprecipitation from stable cell line expressing a VSR.

It is also possible to establish blasticidin or hygromycin-resistant S2 cell lines stably transformed with the appropriate vector (in our case pAWH-P19 or pAWHNLS-P19 constructs) using Effectene Reagent (Invitrogen) according to the manufacturer's instructions (see http://www.flyrnai.org/DRSC-PRL.html for specific

S2 stable cell line establishment).

1. Harvest cells from ten to fifteen 75 cm^2 plates at 80% confluency (Note. 4), wash twice in cold PBS 1 x, and lyse the cells on ice for 30 min in lysis buffer.
2. Centrifuge at 14000 rpm for 15min at 4°C to pellet insoluble cell debris.
3. Transfer 10% of supernatant to a new microcentrifuge tube. This will be the input sample for the analysis of the IPs.
4. Pre-clear the extract: Transfer the remaining supernatant to a new tube and add 50 µl of equilibrated beads for 1 h at 4 °C on a rotating wheel, 15 rpm (Note. 6).
5. After centrifugation at 800 rpm for 5 min at 4 °C add 400 µl of equilibrated Gammabind-Plus resin (see step 4 of 3.1.1) and 20 µg of the appropriate antibody (here: rat anti-HA High Affinity) to the pre-cleared supernatant and incubate for 2 h at 4 °C on a rotating wheel at 15 rpm.
6. After centrifugation at 800 rpm for 5 min at 4 °C wash beads five times in wash buffer.

3.2.1. Purification of immunoprecipitated RNAs.

1. Dilute 25 % of the beads in 1 x Laemmli buffer for protein analysis by western blotting. Store the sample at -20 °C.
2. Wash the remaining beads in wash buffer without proteinase inhibitors.
3. Incubate beads with 20 µl of proteinase K for 2 h (20 µg with an activity of 30U/mg).
4. Add 400 µl of TRIzol and 100 µl of chloroform directly to the beads.
5. Vortex the mix for 15 sec.
(at this step, sample can be stored at -80°C if needed)
6. Incubate sample at room temperature for 3 min.
7. Centrifuge at 14000 rpm at 4 °C for 15 min.
8. Carefully pipette aqueous phase (upper phase) into a clean microcentrifuge tube and discard the lower phase.
9. Add an equal volume (around 140 µl) of isopropanol and mix by gentle inversion (1 µl of glycogen (20 µg/µl) or linear acrylamide (5 µg) carrier facilitates precipitation and visualization of the RNA pellet).
10. Incubate sample at room temperature for 10min.
11. Centrifuge tubes at 14000 rpm at 4 °C for 15min.
12. Discard the supernatant without touching the pellet and wash with 80 % ethanol, vortex briefly to detach the pellet from the tube.
13. Centrifuge at 14000 rpm at 4 °C for 10 min.
14. Carefully remove the ethanol and air-dry the pellet for about 5 to 8 min (a visible white pellet should disappear during the drying).
15. Add 30 µl of RNase-free water to the pellet and resuspend by gently pipetting up and down.

RNA concentration and quality can be checked using a nanodrop spectrophotometer (expected concentration for IPs using the indicated amount of cells and proteins is around 10 (transient expression protocol) to 100 ng/µl (stable expression protocol). The RNA can be used immediately or stored at -80 °C (avoid repeated freezing/unfreezing cycles).

For total RNA analysis (input or supernatant after IPs), RNA should be extracted with TRIzol Reagent (Invitrogen) according to the manufacturer's instructions, except that RNA washes are performed in 80% ethanol.

At this point, RNA from the IPs can be used for standard RNA analysis (RT-qPCR or northern blotting for example). However, in this chapter we focus on [5'-^{32}P]pCp labelling followed by northern blotting.

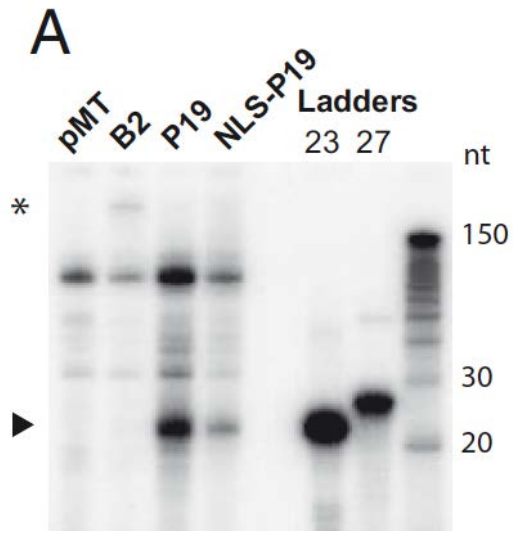
3.2.2. 32P-pCp 3' End-labelling RNA.

RNA molecules can be 3'-end-labeled using [5'-^{32}P]pCp (cytidine-3',5'-bis phosphate) and RNA ligase. The RNA to be labelled must have a free 3'-hydroxyl end for better results. It was shown however that pCp labelling is efficient enough to label small RNA (piRNA) although they are 2'-O-methylated at their 3'-OH end (Saito et al, Genes & Dev 2006). Using this protocol, we aimed at detecting endo-siRNAs IP with VSRs proteins. This class of small RNA are 2'O-methylated at their 3'-extremity after their passage into the Ago2-RISC complex. The 2'O-methylation at the 3'-OH end could affect the efficiency of the pCp 3'-end labelling due to the inaccessibility of the 2' end of small RNAs. However and importantly, the VSRs used to precipitate the siRNAs in this protocol capture them as a duplex before entry into the Ago2-RISC complex and subsequent 2'O-methylation.

1. Pipette 4µl of RNA in a new microcentrifuge tube. This corresponds to around 10% of the immunoprecipitated RNA. More RNA may be labelled if the signal obtains is too low.
2. Add 2.5µl, 100 µCi of [5'-^{32}P]pCp.
3. Add 3µl 10X RNA ligase buffer.
4. Add 3µl DMSO (Note. 7)
5. Add water to 29.5µl total volume.
6. Add 0.5µl T4 RNA ligase (10U).
7. The 30µl reactions are then incubated overnight at 4°C (Note. 8).

At this step, samples can be stored several days at -20°C.

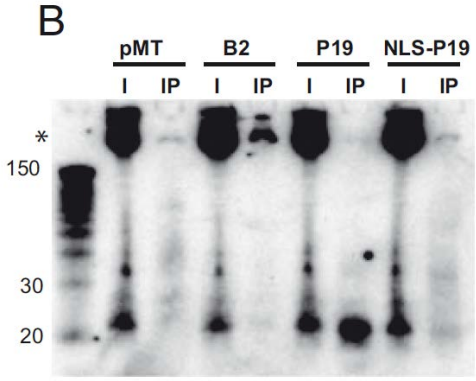
8. Add 100µl of H_2O to the pCp labelled samples.
9. Remove unincorporated nucleotides using a G50-column (GE Healthcare).
10. Pipette 40µl of 3M NaAc, 260µl H2O and 2µl of glycogen to the samples.
11. Add 400µl of Phenol/Chloroform/Isoamyl alcohol and proceed to RNA classical extraction. Carefully pipette aqueous phase (upper phase) into a clean eppendorf tube and precipitate with 1ml of EtOH for 2hr at -20°C. Centrifuge 30min at 4°C, 14000 rpm. Wash with 70% ethanol and let the tubes dry at room temperature (Note. 9).



12. Add 12µl of loading RNA buffer 1X.
13. Denature the samples for 4 min at 95°C and load them onto a 7M urea denaturing 15% polyacrylamide gel.
14. Signals are visualized by autoradiography (see Figure. A and Note. 5&6).

P19 and B2 respectively sequester endogenous TE-matching siRNAs or longer precursors in S2 cells. (A) Immunoprecipitated P19 and NLS-P19 sequester 21nt RNAs that migrate as 22–23nt species after 3'end-pCp labelling (arrowhead) whereas larger RNA species are sequestered exclusively by B2 (*). Control immunoprecipitation (pMT) (Note. 10) was performed using S2 cells transfected with the empty expression vector pMT-DEST48 (from Fagegaltier et al, PNAS, 2009).

Northern blot analysis of IP RNAs was able to confirm the presence of endo-siRNA from the HMS-Beagle retrotransposon. Briefly, twenty micrograms of total RNAs isolated from transfected cells (Input), or 90% of the immunoprecipitated RNAs (IP) were resolved by electrophoresis onto 7M urea denaturing 15% polyacrylamide gels. Classical northern blot analysis was performed in PerfectHyb Plus (Sigma) with sense 5'-^{32}P end-labelled oligonucleotide probe: HMS-Beagle 5'-^{32}P-*TCCCGACATTCCATAGGCATTTA*-3'.



(B) A sense HMS-Beagle probe revealed enriched endo-siRNAs in Northern blots of P19 and NLS-P19 RNAs immunoprecipitates (IP) and longer RNA species (*) in B2 RNA immunoprecipitate. I, corresponds to total RNA input material (From Fagegaltier et al, PNAS, 2009).

Notes.

1. IP experiments were done successfully with less material. However, this depends strongly on the efficiency of the antibody to immunoprecipitate the corresponding protein.

2. Salt concentration in the wash buffer is 150mM NaCl. If background problems occur salt concentration can be raised up to 800mM KCl.

3. Here, monoclonal anti-V5 antibody (Invitrogene) and anti-HA (Roche) were used. However, others appropriate tag antibodies such anti-Flag as well as specific antibodies against VSRs can be used.

4. The number of 75cm^2 plate used from stable cell line depends on the expression of each individual stable cell line. In our hands, RNA IP of 15 plates was always sufficient for a good pCp labelling reaction using the VSRs described in this chapter.

5. Small RNAs immunoprecipitated from S2 cells (control) and stably transformed P19 and NLS-P19 S2 cells were previously cloned for sequencing using the DGE-Small RNA Sample Prep Kit and the Small RNA

Sample Prep v1.5 Conversion Kit from Illumina, following manufacturer instructions (see Fagegaltier et al, PNAS 2009).

6. Pre-clearing can be done overnight to decrease background due to non-specific protein binding to the beads.

7. DMSO seems to improve end-labelling especially with difficult to label RNAs. However, higher concentration of DMSO considerably inhibits ligase activity.

8. For the pCp reaction, 4°C is the recommended temperature. However, reaction can be done at 37°C during 4 hours if needed.

9. The pCp reaction as described above should give several million *cpm* of labelled RNA.

10. Empty vector control is absolutely recommended to detect unspecific or artifactual signal of pCp reaction (see pMT-DEST48 line in *Figure. A*).

Outline of the methods described in this chapter:

- Transfection (transitory) and induction of VSR constructs (4-5 days).
- Isolation of co-immunoprecipitated RNA (1 day).
- Analysis by [5'-^{32}P]pCp 3' end labelling and/or Northern Blot (2 days).

Acknowledgments

We thank D. Kirschner and B. Berry for VSR expression vectors and D. Fagegaltier and A.L. Bougé for fruitful discussions. This work was supported by post-doctoral fellowships from the *Agence Nationale de la Recherche* to CC (grant number ANR BLAN 1210 01 "Nuclear endosiRNAs" to CA) and PhD fellowships from the French government to MvdB.

References.

1. Siomi, M. C., Saito, K. & Siomi, H. How selfish retrotransposons are silenced in Drosophila germline and somatic cells. *FEBS Letters* **582**, 2473–2478 (2008).
2. Fagegaltier, D. *et al.* The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21258–21263 (2009).
3. Ding, S.-W. & Voinnet, O. Antiviral Immunity Directed by Small RNAs. *Cell* **130**, 413–426 (2007).
4. Hammond, S. M. Dicing and slicing: The core machinery of the RNA interference pathway. *FEBS Letters* **579**, 5822–5829 (2005).
5. Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs Exhibit Strand Bias. *Cell* **115**, 209–216 (2003).
6. Schwarz, D. S. *et al.* Asymmetry in the Assembly of the RNAi Enzyme Complex. *Cell* **115**, 199–208 (2003).
7. Horwich, M. D. *et al.* The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. *Current Biology* **17**, 1265–1272 (2007).
8. Li, F. & Ding, S.-W. Virus Counterdefense: Diverse Strategies for Evading the RNA-Silencing Immunity. *Annu. Rev. Microbiol.* **60**, 503–531 (2006).

9. Nayak, A. *et al.* Cricket paralysis virus antagonizes Argonaute 2 to modulate antiviral defense in Drosophila. *Nat Struct Mol Biol* **17**, 547–554 (2010).
10.Vargason, J. M., Szittya, G., Burgyán, J. & Hall, T. M. T. Size selective recognition of siRNA by an RNA silencing suppressor. *Cell* **115**, 799–811 (2003).
11.Rawlings, R. A., Krishnan, V. & Walter, N. G. Viral RNAi suppressor reversibly binds siRNA to outcompete Dicer and RISC via multiple-turnover. *J Mol Biol* **408**, 262–276 (2011).
12.Silhavy, D. *et al.* A viral protein suppresses RNA silencing and binds silencing-generated, 21- to 25-nucleotide double-stranded RNAs. *The EMBO Journal* **21**, 3070–3080 (2002).
13.Lakatos, L., Szittya, G., Silhavy, D. & Burgyán, J. Molecular mechanism of RNA silencing suppression mediated by p19 protein of tombusviruses. *EMBO J* **23**, 876–884 (2004).
14.Dunoyer, P., Lecellier, C.-H., Parizotto, E. A., Himber, C. & Voinnet, O. Probing the microRNA and small interfering RNA pathways with virus-encoded suppressors of RNA silencing. *Plant Cell* **16**, 1235–1250 (2004).
15.Liu, X., Houzet, L. & Jeang, K.-T. Tombusvirus P19 RNA silencing suppressor (RSS) activity in mammalian cells correlates with charged amino acids that contribute to direct RNA-binding. *Cell Biosci* **2**, 41 (2012).
16.Chao, J. A. *et al.* Dual modes of RNA-silencing suppression by Flock House virus protein B2. *Nature Structural & Molecular Biology* **12**, 952–957 (2005).
17.Singh, G. *et al.* Suppression of RNA silencing by Flock house virus B2 protein is mediated through its interaction with the PAZ domain of Dicer. *FASEB J.* **23**, 1845–1857 (2009).

Introduction II: reproducible computational analysis with Galaxy

1. Development of Galaxy tools linked to data analysis

Before I joined the laboratory, Christophe Antoniewski wrote and published a number of commandline analysis pipelines for miRNA analyses (Vandormael-Pournin et al. 2012; Reinhardt et al. 2012), profiling of viral small RNA reads (Antoniewski 2011), endo-siRNA (Fagegaltier et al. 2009) and the detection of small RNA signatures (Antoniewski 2014). During the first year of my PhD we set up a Galaxy server in the lab, and in the course of my PhD we adapted and improved the aforementioned analysis pipelines within the Galaxy framework. This work served as a basis for a number of collaborations, as the ease of using bioinformatic pipelines within Galaxy allowed for individual biologists to perform complex analyses autonomously.

Early during this process I participated mostly in administrating the system, as I had very little programming background, which changed during my 2nd year of the PhD. I nevertheless presented my work of designing the backup mechanism at the Galaxy Days in December 2013 (http://www.ifb-galaxy.org/4dec2013.html). This also involved the development of scripts for the automatic creation of Virtual Machines from these Backups. While these scripts are still at work in our environment, and I heavily used these Virtual Machine Images to develop tools for small RNA visualization, I never pushed these scripts to publication level, as I was more involved in the analysis of my own data and I felt that these scripts were not flexible enough and that other, new technologies would be more suitable for the task. Nevertheless, I became more familiar with how Galaxy and virtualization works.

Chronologically, we first received small RNA samples from heads. As mentioned in the introduction to the first article manuscript, we were initially interested to see whether we could detect small RNAs near the *white* gene that might explain variegation in the wm^4 background strain. Following the realization that only few small RNAs were present, we nevertheless made an interesting observation. It appeared that siRNAs that aligned to TEs were upregulated. To efficiently analyse which TEs were upregulated, in which library and whether there were any hotspots for all 121 canonical TEs, I modified and combined existing scripts that address this question separately. We managed to generalize these analyses using the R lattice package. The development of these tools allowed me to get a good grasp on Galaxy tool development (detailed in part 4) and the difficulties associated with this. I presented the resulting code in a Lightning Talk during the Galaxy Community Conference 2014 (A poster is hosted here: https://wiki.galaxyproject.org/Documents/Posters/GCC2014?action=AttachFile&do=view&target=P6vandenBeek.pdf, and a video of the lightning talk is available here:

http://jh.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=9fed7061-735f-4c3a-93a3-95b7ff51dd35).

Around the same time I made first steps in understanding and writing python code, while working on a small script that clusters small RNAs. When exploring the RNA-seq data for the first article, I got more familiar with the python language, writing simple, interactive analysis notebooks (an interesting example is in Annex 2) using IPython (Pérez and Granger 2007) almost every day, but I never integrated these analyses in Galaxy, as writing complete, universal tools would have taken too much time for an analysis that would be performed only once or twice and will almost certainly not lead to any publication.

However I learned about containerization using Docker at the Galaxy Community Conference, and later during the summer I realized that containerization could be combined with the Galaxy Tool Factory (Lazarus, Kaspi, and Ziemann 2012) to archive these run-once analyses side-by-side with their input data in Galaxy, thereby maintaining the analysis code directly with the input-data. Beyond this purpose, the DockerToolFactory also allows the generation of simple Galaxy tools, which can be installed from the Galaxy toolshed (Blankenberg, Von Kuster, et al. 2014) (a system similar to commercial app stores, but for Galaxy tools). Finally it is possible to chain these simple tools together in workflows, allowing any arbitrary operation to be executed, without risk of damage through intentional or unintentional malicious code execution. Therefore I have written an article that describes this enhancement to the original Galaxy ToolFactory (Lazarus, Kaspi, and Ziemann 2012) (van den Beek and Antoniewski, in preparation).

2. The cornerstones of reproducible research

Computational methods have become a cornerstone of many research projects in the life sciences, especially in genomics, but also in proteomics, metabolomics and microscopy and personalized medicine, and this true also for my thesis project. In the field of genomics microarrays and next generation sequencing have changed in many ways how research is being done, enabling projects that were not possible before. However, these projects generate datasets that are far too large and complex to be handled in a simple spreadsheet software. Their proper analysis poses multiple challenges at different levels. An overview of these is listed in the introduction to (Goecks et al. 2010). The challenges can be broken down in 3 major categories: Accessibility, Reproducibility and Transparency.

The **Accessibility** challenge manifests itself for researchers that need to use computational approaches, but have little or no informatics or programming expertise. Hurdles here are the use of command line tools, their installation, configuration and maintenance, as well as the selection of parameters suitable to the problem at hand and the access to sufficiently

powerful computing environments. This is further complicated if multiple analysis steps have to be chained together.

Reproducibility is another major challenge, as it takes considerable experience, stringency and organization effort to document all datasets, analysis steps, tools and parameters used. Although most researcher keep careful track of their wet-lab experiments in lab journals, the sheer mass of tools often used during the exploratory data analysis phase makes proper documentation of all details challenging and time consuming.

Transparency is perhaps the most important aspect of any large scale data analysis. While it is beneficial to be able to reproduce an analysis, it is far more important to document analyses in a way that they can be understood, adopted and modified at any level of detail by fellow scientists.

3. The Galaxy framework

A huge number of efforts exists to address these problems, that reach from online communities dedicated to sequencing experiments (http://seqanswers.com/), to more general question sites (http://stackoverflow.com/, https://www.biostars.org/) and specialised mailing lists. There are specialised software repositories that provide a common syntax (Bioconductor, Biopython, Bioperl, scipy, scikit-learn) and interactive web-based integrated development environments (IDE) (Rstudio, IPython) that facilitate sharing analyses. And then there are high level reproducible research systems (RRS) such as Mobyle (Néron et al. 2009), GenePattern (Reich et al. 2006), Illumina BaseSpace, Taverna (Wolstencroft et al. 2013) and Galaxy (Giardine et al. 2005), that aim to “glue” together different parts of an analysis pipeline in a **user-friendly** interface, while automatically tracking input files, output files, used tools and their parameters (to aid reproducibility). Finally, RRSs should provide means to disseminate and annotate analyses, so that the intent and methods of the analysis can be easily understood (aiding transparency).

Of the before-mentioned RRSs, Galaxy appears to be one of the more feature-rich platforms with strong, active development and a large community, hence we decided to provide the tools that we developed for our own research or for collaborative efforts inside the Galaxy environment. I will not go into the details of how Galaxy works, as the authors did a much better job than I could possibly do ((Goecks et al. 2010) and https://usegalaxy.org), and so I will limit myself to pointing out the tool-relevant parts and highlight some current limitations, and how I propose to lift these.

4. Galaxy tool development

Any piece of software that can be operated on the command line can be integrated into galaxy. To do so, a **wrapper** is needed that, using a number of key-words, “describes” to Galaxy what the user-interface should look like, where the input datasets are expected during the construction of the command line, the type of input data that can be used, the number and type of output datasets, and many more things. These wrapper differ in their complexity and are written in the XML standard. A typical, simple wrapper might look like the following:

```
1 <tool id="mismatch_analyzer" name="Mismatch Analyzer" version="0.0.1" hidden="false" >
2        <description>Analyze mismatches in BAM/SAM alignments</description>
3        <command interpreter="python">mismatch_analyzer.py --input
4                #for i in $rep
5                        "$i.input_file"
6                #end for
7                --name
8                #for i in $rep
9                        "$i.input_file.name"
10                #end for
11                  --output_pdf $output_pdf
12                  --output_tab $output_tab
13                  --min $min_length
14                  --max $max_length
15        </command>
16        <inputs>
17            <repeat name="rep" title="alignment files">
18                <param name="input_file" type="data" format="bam,sam" label="Alignment
file"                           help="The input alignment file(s) for which to analyze the
mismatches."/>
19            </repeat>
20          <param name="min_length" type="text" value="21"/>
21          <param name="max_length" type="text" value="21"/>
22        </inputs>
23        <outputs>
24                <data format="pdf" name="output_pdf" />
25                <data format="tabular" name="output_tab" />
26        </outputs>
27 </tool>
```

I wrote and used this wrapper to analyse the mismatches as in Figure 1G of the first article. The purpose of this wrapper is to provide the correct command line options to the “`mismatch_analyzer.py`” script (lines 3-15), based on the options selected by a Galaxy user (lines 16-22).

If the script was operated directly on the command line the user would have to type the following line:

`python mismatch_analyzer.py --input "my input.bam" --name "my input" --output_pdf "my output.pdf" --output_tab "my output.tab" --min "21" --max "21"`. Instead, the wrapper instructs Galaxy to provide a tool form, where the user can add one or more alignment files, and where he can select the length of the reads to investigate (with 21 as a default value). Galaxy will then take care to inject the user-selected values at the right place in the command line.

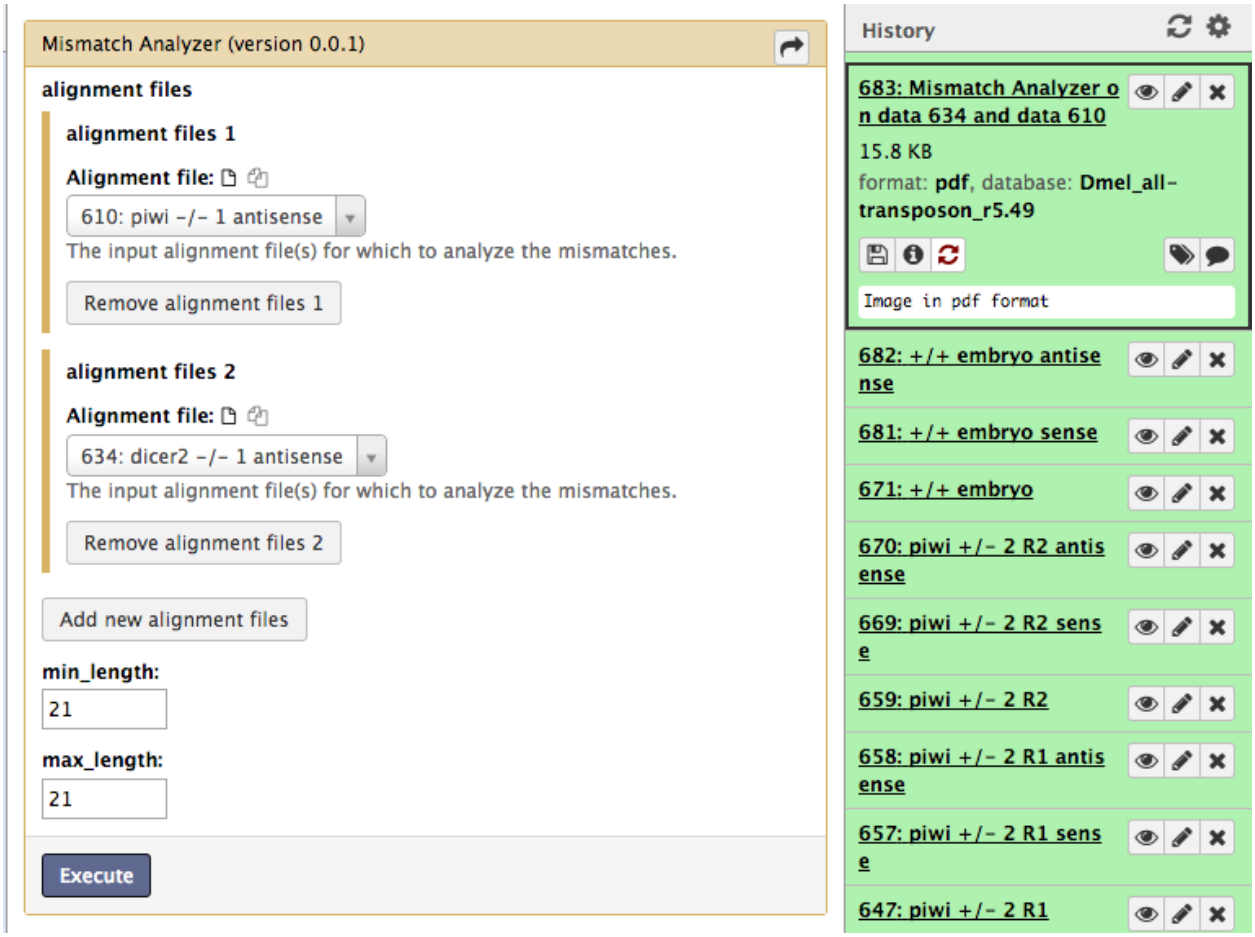


Figure 9. The Mismatch Analyzer tool form, as defined in in the above xml wrapper.

All instructions for Galaxy are contained within the tool tagset. tagsets start with a keyword, e.g. "<command" and end with "/>", or if a tagset spans multiple lines, the keyword is repeated, e.g. </command>. Tagsets can be nested, for example the "input" tagset contains a "repeat" tagset, which contains a "param" tagset.

A list of available keywords and options can be found in the Galaxy wiki (https://wiki.galaxyproject.org/Admin/Tools/ToolConfigSyntax). These xml forms can quickly become very long and unreadable, and they are sensitive to missing or invalid options. This often makes tool development a trial-and-error process. Nevertheless, it also provides

traceability, as the options are “fixed”. When tools are run, the user-selected values are stored in a database, and can be retrieved when necessary. If for example, in the next version of mismatch_analyzer.py the option to specifiy the miminal readlength to analyse would change from “`--min`” to “`--minimum_readlength`”, line 13 would need to be adapted to read “`--minimum_readlength $min_length`”. Users that now wish to redo an old analysis with the same parameters as before can do so, as the value stored in the database is linked to the parameter `"min_length"` defined in the wrapper.
Subjectively speaking, it does take a bit of experience to write a good Galaxy wrapper. Once a wrapper is finished, it has to be either added to a toolshed, or referenced to in a Galaxy configuration file, and finally Galaxy needs to be restarted (in the latter case) or the tool needs to be installed from the toolshed. These processes can only be performed by Galaxy administrators, and so Galaxy users are limited by the choice of tools that are currently installed on the server. This might be a serious drawback for users that wish to take advantage of the Galaxy Platform, and that at the same time know how to write analysis code themselves.

5. DockerToolFactory in the light of current limitations

The problems I addressed with the DockerToolFactory tool are:
[1] Galaxy tool development is slow, error-prone and requires good knowledge of available options.
[2] Only Galaxy administrators can install new tools.
[3] Experienced users cannot run custom analysis scripts.

As mentioned before, DockerToolFactory is based on the Galaxy Tool Factory (Lazarus, Kaspi, and Ziemann 2012). The original Galaxy Tool Factory addressed point [1], as it was possible to generate new galaxy tools, based on a user-supplied script. However the Galaxy Tool Factory could be executed only as a galaxy administrator, as any script could be executed, which raises security concerns. I modified the Galaxy Tool Factory in a way that only data that belong to the current galaxy user are available to the script, and these data can only be read, but not modified. Scripts run completely isolated from the rest of the system. For the isolation I made use of containerization using Docker (https://www.docker.com/). Containerization allows running multiple operating systems (“containers”) side by side, with very little overhead. All that is shared with the host system (= the Galaxy server, or the computing node, if tools are executed on a computing cluster) is the kernel. In simplified terms, the kernel is responsible for communicating operations between the physical hardware and the software.

By isolating (sometimes called “sandboxing”) the execution of scripts in a container, it is now not a security problem anymore to let users run arbitrary code. Therefore, galaxy administrators can now allow users to execute arbitrary scripts, and using the DockerToolFactory, these scripts can be converted into galaxy tools, or used directly in interactive analyses, or as an intermediary step in a workflow.

In addition I extended the Galaxy Tool Factory for multiple input files, and allowed the specification of output types. These modifications have been incorporated back into the original Galaxy Tool Factory.

This approach turned out to be quite powerful, and I am routinely using the DockerToolFactory in situations where I would like to do simple data transformations, or when I need to plot data in a way that is difficult to generalize as a multi-purpose galaxy tool. As an example, all Figures except Figure 3A in my first article have been made using the DockerToolFactory. Annex 3 contains a table with links to the Galaxy histories and workflows used to construct these Figures.

Results II: Running arbitrary user code on Galaxy using DockerToolFactory

Title: Running arbitrary user code on Galaxy using DockerToolFactory

Marius van den Beek, Christophe Antoniewski

Drosophila Genetics and Epigenetics ; Université Pierre et Marie Curie 9, Quai St Bernard Building C – 5th floor – Room 517 ; 75252 Paris cedex 05 ; Phone: +33 1 44 27 34 39

Motivation: Galaxy is a software framework that enables reproducible research for data intensive applications. Drawbacks are that (1) for the integration of new tools or scripts, one needs administrator access, (2) the Galaxy tool generation is slow, error-prone process and is not very well documented.

Results: DockerToolFactory is a regular Galaxy tool installable from the Galaxy toolshed. Once installed by a Galaxy administrator, it allows the execution of arbitrary scripts, and the generation of new Galaxy tools for sharing on the public Galaxy toolshed (GTS). Secure execution of code is achieved by running code in a secure and isolated Docker container. DockerToolFactory is a fork of The Tool Factory.

Availability and implementation: The Galaxy administrative interface supports automated installation from the main GTS. Source code and support are available at the project website, https://bitbucket.org/mvdbeek/dockertoolfactory. The DockerToolFactory is implemented as an installable Galaxy tool.

Contact: mvandenb@snv.jussieu.fr

1. Introduction

Galaxy (Giardine et al. 2005; Blankenberg et al. 2010; Goecks et al. 2010; Nekrutenko and Taylor 2012) is a web framework that is centered around enabling reproducible research. This includes genomics, epi-genomics, proteomics, metabolomics, statistics, and medical research, which tend to be very data-intensive. Galaxy is not limited to any kind of analysis, given that a command line version of the analysis can be produced. It provides users the ability to run tools that would require extensive knowledge of the command line. Through its workflow engine and Galaxy Pages, analyses can be easily repeated and shared, thereby allowing users to obtain results quickly and reproducibly, including almost publication grade Documentation.

A serious limitation to these goals is a recurrent theme among software packages that aim to simplify biological analyses: If a required intermediate step in the analyses is not provided by the software package, the user has to obtain the intermediate file and treat it on his own or with the help of a (bio-) informatician.
In the case of Galaxy this means that the workflow is interrupted, the user has to manually download a file, treat it on his workstation and re-upload it to Galaxy to continue the analysis. In the best case scenario the user documented the data treatment well and provides a script when the resulting paper is published. In the worst case scenario the user introduces an error during his analysis, especially if a large number of files need to be treated, does not provide a script and so the error will be left unnoticed and the analysis cannot be reproduced, or the user decides to abandon Galaxy altogether and relies on classical script development with all its advantages and disadvantages.

To remedy the situation Lazarus et al. developed the Galaxy Tool Factory (Lazarus, Kaspi, and Ziemann 2012), which is a tool that can be used to rapidly and easily transform existing single input/output scripts into Galaxy tools. Because the user can potentially execute malicious code, installation on production instances was strongly advised against, and thus a private instance of a Galaxy server was a requirement.

To improve the situation, we developed DockerToolFactory. It is a fork of The Tool Factory that allows the secure execution of scripts and can treat multiple input/output files. Execution of code is secured using Docker. Generated tools can be uploaded to the Galaxy toolshed (Blankenberg et al. 2014), or scripts can be run interactively or in workflows.

2. Methods

DockerToolFactory is a fork of Tool Factory and is implemented in Python. By using Docker containers, secure and lightweight sandbox are established on-the-fly during script execution.
DockerToolFactory requires the docker daemon to be running and accessible to the system user under which Galaxy is run.
As in the original Galaxy Tool Factory, the user is presented with a standard Galaxy tool screen, where he can paste his executable script. In addition it is now possible to select multiple input files, specify their file format and the file format of the output files.
When the user executes a script within DockerToolFactory, a container is created using the supplied Dockerfile. The Dockerfile specifies which system-wide software is available to the user running his script and can be easily adapted by the Galaxy administrator.

Input files, output files and intermediate scripts are then bind-mounted to the container. The user-supplied script can thus only access the files that are selected as input for the DockerToolFactory tool, limiting potential security issues to the user's own files.
By allowing arbitrary shell, python, R and perl scripts users can install software not available on the Galaxy instance or in the specified Dockerfile.
When the script has terminated or has hit the configurable resource limit, the docker container is stopped and terminated. Outputs appear in Galaxy as regular outputs and can be embedded in workflows and can be re-run. In addition, the user may choose to generate a tool from his script. The generated tool is compatible with the Galaxy toolshed and may be installed by a Galaxy administrator on any Galaxy instance. Scripts may accept multiple inputs and write multiple outputs of different file formats, or, if a script writes many output files, output files can be collected and linked within an html file.

3. Results

Non-administrator users of local Galaxy instances can now execute arbitrary pieces of code and install software packages in their Galaxy environment, containing all of their previously generated datasets. Through an admin-interface panel administrators can adjust the resources available to run docker containers.

4. Discussion

The Galaxy environment is very appealing to novice users, as it allows to run complex multi-step analysis pipelines without taking care of installing all necessary software packages, assembling complex pipelines and maintaining them. In addition it is very easy to develop, share and reproduce analyses with collaborators. However, "intermediate users" of Galaxy are often comfortable in writing simple scripts. To run these scripts in Galaxy, knowledge in the quite complex xml format is a prerequisite, as are administrator rights on a Galaxy instance. The installation and maintenance of a Galaxy instance, physically or in the cloud, is again time- and resource consuming. This might lead "intermediate users", which could be valuable tool developers, to back away from the Galaxy environment. In addition, they might also advise novice users to develop their analysis pipelines on their local machines.
By making it possible to run user-supplied scripts, this target audience might be convinced to follow the Galaxy approach of reproducible science.

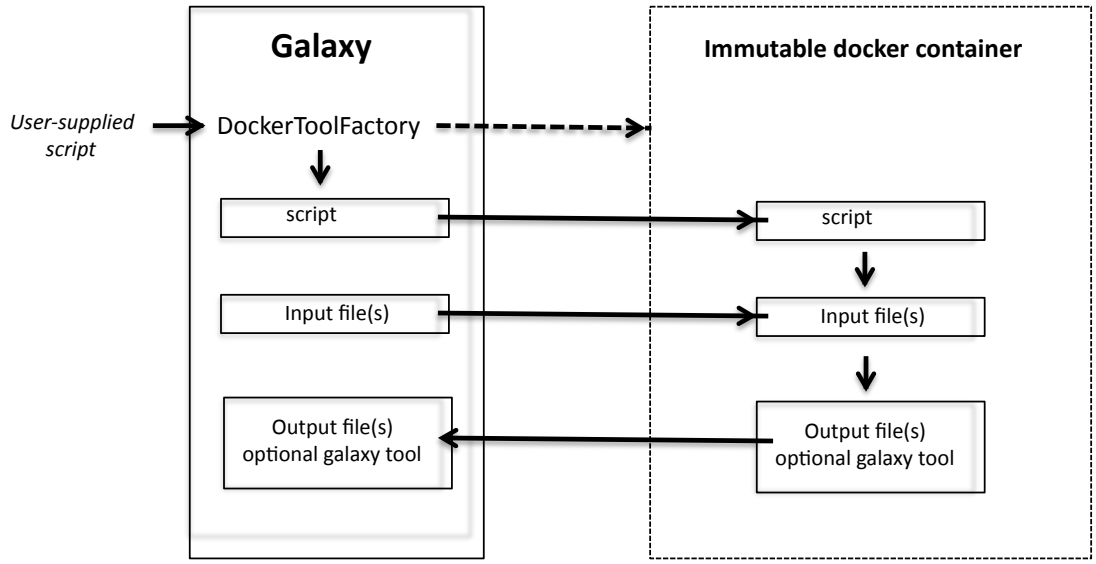


Fig. 1. User-supplied script can be pasted into the DockerToolFactory tool, installed in Galaxy. When the user runs the tool, a new immutable container is created. The supplied script and input files are mounted into the created container and the script is executed. When script execution finishes, the containers are stopped and removed, while output files appear in the users' history

Furthermore, the approach to have users be able to run scripts and generate tools themselves minimizes the need for separate Galaxy instances that only exist because users want to run tools not available in their institute's local Galaxy instance.

The generated tools serve as a good starting point for developers to add more complex xml tool syntax. Together with the recent efforts to integrate Docker-based IPython sessions in Galaxy, Galaxy itself is becoming a serious alternative to local Integrated Development Environments (IDE).

Currently the docker container is running as the system's Galaxy user, and inside the container the user's script is executed as a non-privileged user, so software that requires root access cannot be executed.

The Docker development team is currently working on the possibility to run docker containers in user namespace. When this feature will become available, users might be allowed to upload their own Dockerfiles to Galaxy and become the root user inside these images. This means that users could upload their own Dockerfiles, specifying exactly which software should be available for their scripts.

5. Availablilty

The DockerToolFactory is installable through the Galaxy administrative interface from the GTS. Source code is available at the project homepage https://bitbucket.org/mvdbeek/dockertoolfactory.

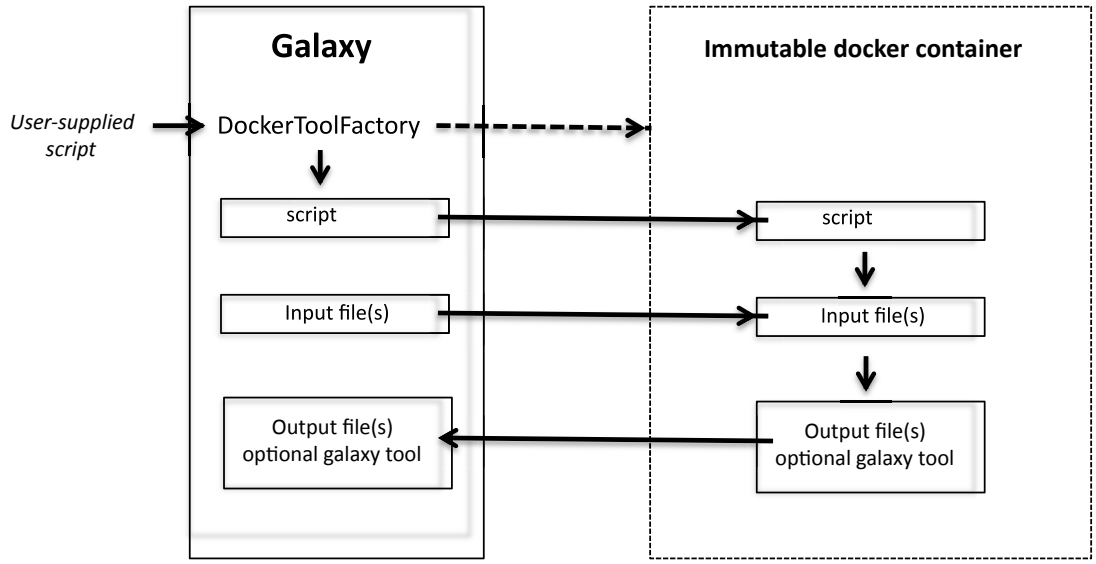


Fig. 1. User-supplied script can be pasted into the DockerToolFactory tool, installed in Galaxy. When the user runs the tool, a new immutable container is created. The supplied script and input files are mounted into the created container and the script is executed. When script execution finishes, the containers are stopped and removed, while output files appear in the users' history

Furthermore, the approach to have users be able to run scripts and generate tools themselves minimizes the need for separate Galaxy instances that only exist because users want to run tools not available in their institute's local Galaxy instance.

The generated tools serve as a good starting point for developers to add more complex xml tool syntax. Together with the recent efforts to integrate Docker-based IPython sessions in Galaxy, Galaxy itself is becoming a serious alternative to local Integrated Development Environments (IDE).

Currently the docker container is running as the system's Galaxy user, and inside the container the user's script is executed as a non-privileged user, so software that requires root access cannot be executed.

The Docker development team is currently working on the possibility to run docker containers in user namespace. When this feature will become available, users might be allowed to upload their own Dockerfiles to Galaxy and become the root user inside these images. This means that users could upload their own Dockerfiles, specifying exactly which software should be available for their scripts.

5. Availablilty

The DockerToolFactory is installable through the Galaxy administrative interface from the GTS. Source code is available at the project homepage https://bitbucket.org/mvdbeek/dockertoolfactory.

5. References

Blankenberg, Daniel, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, Galaxy Team, James Taylor, and Anton Nekrutenko. 2014. "Dissemination of Scientific Software with Galaxy ToolShed." *Genome Biology* 15 (2): 403.

Blankenberg, Daniel, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. 2010. "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 19 (January): Unit 19.10.1–21.

Giardine, Belinda, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. "Galaxy: A Platform for Interactive Large-Scale Genome Analysis." *Genome Research* 15 (10): 1451–55.

Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences." *Genome Biology* 11 (8): R86.

Lazarus, Ross, Antony Kaspi, and Mark Ziemann. 2012. "Creating Reusable Tools from Scripts: The Galaxy Tool Factory." *Bioinformatics* 28 (23): 3139–40.

Nekrutenko, Anton, and James Taylor. 2012. "Next-Generation Sequencing Data Interpretation: Enhancing Reproducibility and Accessibility." *Nature Reviews. Genetics* 13 (9): 667–72.

Perspectives for DockerToolFactory

All aspects I have mentioned in the above article draft are implemented, work and can stand on its own, but I would like to introduce a few additional features that further augment the utility of the DockerToolFactory before submitting the manuscript to a journal.

I am planning to extend the DockerToolFactory with the following features:

1. Multiple output files
2. The possibility to use reference data stored in Galaxy's tool-data tables
3. Split the DockerToolFactory in 4 modules: run-script, wrapper-only, wrapper-modification, tool-dependency generation
4. Automatic upload into galaxy toolshed and code-repositories

1. Multiple output files

Currently, only a single output file is appearing in the history, in addition to an optional HTML file. At present the only way to produce multiple output files is using the HTML-output mode. In that case all files that are written by the users' scripts are collected and placed as downloadable links within the HTML file. These output files are not available to further treatment in Galaxy, unless they are downloaded and re-uploaded. Allowing the generation of multiple output files could be in the form of repeat elements, as is done for multiple input files. However, the user then needs to take care of which variables represent input and which files are output files. Alternatively, all files could be automatically collected and placed in the history, but this would break workflows, as the number and order of outputs are not known to the workflow engine before the tool has finished. At present I would prefer the former solution. To aid script authors one could prefill the text box in which the script is pasted with instructions of how to reference input and output files and other optional parameters. In addition I would like to add textboxes to specify "name" and "help" tagsets to all input and parameter fields.

2. Make Galaxy's reference data available

Galaxy provides the possibility to centrally store reference data, so that all users automatically have access to these data when necessary. Reference data includes genome sequences, reference indexes in different formats for different read aligners, blast databases and so on. These are organized in tool-data tables (Blankenberg, Johnson, et al. 2014). It

would useful to be able to select from these reference data sources in the DockerToolFactory.

3. Split the DockerToolFactory in task-specific modules

Adding the above-mentioned options would add more complexity to the tool form, that would not be required if one was only interested in running simple scripts. Therefore, a simpler **script-only** tool could be generated that is used only for adding input files, the script to be executed and specifying the type of output file.
In addition, it might be possible to first generate a basic wrapper, which is stored in the history. Details, like tool-data table entries, help texts, repeats, conditionals, and others could be added in an **interactive visualization** process. One can then make use of java-script to render previews of the generated tools. This mode could also be used to modify existing wrappers. Finally, it would be possible to have a mode that would be used to install dependencies. Right now it is possible to install R and python packages during script execution, but this process would be repeated every time the script is run. Instead, it would be possible to install the dependencies, and to make a snapshot of the container for later use.

4. Automatic upload into galaxy toolshed and code-repositories

To facilitate the process of script-development and versioning, I plan to implement an automatic upload to the galaxy toolshed and to code repositories like github (https://github.com/) and bitbucket (https://bitbucket.org). All of these support automatic upload through an Application Programming Interface (API). A hurdle here is to store API keys, perhaps in the user-preferences panel of galaxy.

References

Ameres, Stefan L, Michael D Horwich, Jui-Hung Hung, Jia Xu, Megha Ghildiyal, Zhiping Weng, and Phillip D Zamore. 2010. “Target RNA–Directed Trimming and Tailing of Small Silencing RNAs.” *Science* 328 (5985): 1534–39.

Antoniewski, Christophe. 2011. “Visitor, an Informatic Pipeline for Analysis of Viral siRNA Sequencing Datasets.” *Methods in Molecular Biology* 721: 123–42.

———. 2014. “Computing siRNA and piRNA Overlap Signatures.” In *Animal Endo-SiRNAs*, 135–46. Methods in Molecular Biology. Springer New York.

Aravin, Alexei A, Gregory J Hannon, and Julius Brennecke. 2007. “The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race.” *Science* 318 (5851): 761–64.

Aravin, Alexei A, Mariana Lagos-Quintana, Abdullah Yalcin, Mihaela Zavolan, Debora Marks, Ben Snyder, Terry Gaasterland, Jutta Meyer, and Thomas Tuschl. 2003. “The Small RNA Profile during Drosophila Melanogaster Development.” *Developmental Cell* 5 (2): 337–50.

Aravin, Alexei A, Natalia M Naumova, Alexei V Tulin, Vasilii V Vagin, Yakov M Rozovsky, and Vladimir A Gvozdev. 2001. “Double-Stranded RNA-Mediated Silencing of Genomic Tandem Repeats and Transposable Elements in the D. Melanogaster Germline.” *Current Biology: CB* 11 (13): 1017–27.

Bell, Callum J, Darrell L Dinwiddie, Neil A Miller, Shannon L Hateley, Elena E Ganusova, Joann Mudge, Ray J Langley, et al. 2011. “Carrier Testing for Severe Childhood Recessive Diseases by next-Generation Sequencing.” *Science Translational Medicine* 3 (65): 65ra4.

Bernstein, Emily, Amy A Caudy, Scott M Hammond, and Gregory J Hannon. 2001. “Role for a Bidentate Ribonuclease in the Initiation Step of RNA Interference.” *Nature* 409 (6818): 363–66.

Blanchard, Daniel P, Daphne Georlette, Lisa Antoszewski, and Michael R Botchan. 2014. “Chromatin Reader L(3)mbt Requires the Myb–MuvB/DREAM Transcriptional Regulatory Complex for Chromosomal Recruitment.” *Proceedings of the National Academy of Sciences* 111 (40): E4234–E4243.

Blankenberg, Daniel, James E Johnson, Galaxy Team, James Taylor, and Anton Nekrutenko. 2014. “Wrangling Galaxy’s Reference Data.” *Bioinformatics* 30 (13): 1917–19.

Blankenberg, Daniel, Gregory Von Kuster, Emil Bouvier, Dannon Baker, Enis Afgan, Nicholas Stoler, Galaxy Team, James Taylor, and Anton Nekrutenko. 2014. “Dissemination of Scientific Software with Galaxy ToolShed.” *Genome Biology* 15 (2): 403.

Brennecke, Julius, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. 2007. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila.” *Cell* 128 (6): 1089–1103.

Brennecke, Julius, Alexander Stark, Robert B Russell, and Stephen M Cohen. 2005. “Principles of MicroRNA–Target Recognition.” *PLoS Biology* 3 (3). Public Library of Science: e85.

Bronkhorst, Alfred W, Koen W R van Cleef, Hanka Venselaar, and Ronald P van Rij. 2014. “A dsRNA-Binding Protein of a Complex Invertebrate DNA Virus Suppresses the Drosophila RNAi Response.” *Nucleic Acids Research* 42 (19): 12237–48.

Carré, Clément, Caroline Jacquier, Anne-Laure Bougé, Fabrice de Chaumont, Corinne Besnard-Guerin, Hélène Thomassin, Josette Pidoux, et al. 2013. “AutomiG, a Biosensor to Detect Alterations in miRNA Biogenesis and in Small RNA Silencing Guided by Perfect Target Complementarity.” *PloS One* 8 (9): e74296.

Castañeda, Julio, Pavol Genzor, Godfried W van der Heijden, Ali Sarkeshik, John R Yates 3rd, Nicholas T Ingolia, and Alex Bortvin. 2014. “Reduced Pachytene piRNAs and Translation Underlie Spermiogenic Arrest in Maelstrom Mutant Mice.” *The EMBO*

Journal 33 (18): 1999–2019.
Cenik, Elif Sarinay, Ryuya Fukunaga, Gang Lu, Robert Dutcher, Yeming Wang, Traci M Tanaka Hall, and Phillip D Zamore. 2011. “Phosphate and R2D2 Restrict the Substrate Specificity of Dicer-2, an ATP-Driven Ribonuclease.” *Molecular Cell* 42 (2): 172–84.
Cernilogar, Filippo M, Maria Cristina Onorati, Greg O Kothe, A Maxwell Burroughs, Krishna Mohan Parsi, Achim Breiling, Federica Lo Sardo, et al. 2011. “Chromatin-Associated RNA Interference Components Contribute to Transcriptional Regulation in Drosophila.” *Nature* 480 (7377): 391–95.
Chendrimada, Thimmaiah P, Kenneth J Finn, Xinjun Ji, David Baillat, Richard I Gregory, Stephen A Liebhaber, Amy E Pasquinelli, and Ramin Shiekhattar. 2007. “MicroRNA Silencing through RISC Recruitment of eIF6.” *Nature* 447 (7146): 823–28.
Czech, Benjamin, Colin D Malone, Rui Zhou, Alexander Stark, Catherine Schlingeheyde, Monica Dus, Norbert Perrimon, et al. 2008. “An Endogenous Small Interfering RNA Pathway in Drosophila.” *Nature* 453 (7196): 798–802.
Czech, Benjamin, Jonathan B Preall, Jon McGinn, and Gregory J Hannon. 2013. “A Transcriptome-Wide RNAi Screen in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway.” *Molecular Cell* 50 (5): 749–61.
De Vanssay, Augustin, Anne-Laure Bougé, Antoine Boivin, Catherine Hermant, Laure Teysset, Valérie Delmarre, Christophe Antoniewski, and Stéphane Ronsseray. 2012. “Paramutation in Drosophila Linked to Emergence of a piRNA-Producing Locus.” *Nature* 490 (7418): 112–15.
Denli, Ahmet M, Bastiaan B J Tops, Ronald H A Plasterk, René F Ketting, and Gregory J Hannon. 2004. “Processing of Primary microRNAs by the Microprocessor Complex.” *Nature* 432 (7014): 231–35.
Dennis, Cynthia, Vanessa Zanni, Emilie Brasset, Angeline Eymery, Liang Zhang, Rana Mteirek, Silke Jensen, Yikang S Rong, and Chantal Vaury. 2013. “‘Dot COM’, a Nuclear Transit Center for the Primary piRNA Pathway in Drosophila.” *PloS One* 8 (9): e72752.
Dietzl, Georg, Doris Chen, Frank Schnorrer, Kuan-Chung Su, Yulia Barinova, Michaela Fellner, Beate Gasser, et al. 2007. “A Genome-Wide Transgenic RNAi Library for Conditional Gene Inactivation in Drosophila.” *Nature* 448 (7150): 151–56.
Elbashir, S M, J Martinez, A Patkaniowska, W Lendeckel, and T Tuschl. 2001. “Functional Anatomy of siRNAs for Mediating Efficient RNAi in Drosophila Melanogaster Embryo Lysate.” *The EMBO Journal* 20 (23): 6877–88.
Eulalio, Ana, Eric Huntzinger, and Elisa Izaurralde. 2008. “GW182 Interaction with Argonaute Is Essential for miRNA-Mediated Translational Repression and mRNA Decay.” *Nature Structural & Molecular Biology* 15 (4): 346–53.
Fagegaltier, Delphine, Anne-Laure Bougé, Bassam Berry, Emilie Poisot, Odile Sismeiro, Jean-Yves Coppée, Laurent Théodore, Olivier Voinnet, and Christophe Antoniewski. 2009. “The Endogenous siRNA Pathway Is Involved in Heterochromatin Formation in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (50): 21258–63.
Fire, Andrew, Siqun Xu, Mary K Montgomery, Steven A Kostas, Samuel E Driver, and Craig C Mello. 1998. “Potent and Specific Genetic Interference by Double-Stranded RNA in Caenorhabditis Elegans.” *Nature* 391 (6669): 806–11.
Fukunaga, Ryuya, Bo W Han, Jui-Hung Hung, Jia Xu, Zhiping Weng, and Phillip D Zamore. 2012. “Dicer Partner Proteins Tune the Length of Mature miRNAs in Flies and Mammals.” *Cell* 151 (3): 533–46.
Förstemann, Klaus, Yukihide Tomari, Tingting Du, Vasily V Vagin, Ahmet M Denli, Diana P Bratu, Carla Klattenhoff, William E Theurkauf, and Phillip D Zamore. 2005. “Normal microRNA Maturation and Germ-Line Stem Cell Maintenance Requires Loquacious, a Double-Stranded RNA-Binding Domain Protein.” *PLoS Biology* 3 (7): e236.
Galiana-Arnoux, Delphine, Catherine Dostert, Anette Schneemann, Jules A Hoffmann, and Jean-Luc Imler. 2006. “Essential Function in Vivo for Dicer-2 in Host Defense against RNA Viruses in Drosophila.” *Nature Immunology* 7 (6): 590–97.
Gasciolli, Virginie, Allison C Mallory, David P Bartel, and Hervé Vaucheret. 2005. “Partially

Redundant Functions of Arabidopsis DICER-like Enzymes and a Role for DCL4 in Producing Trans-Acting siRNAs." *Current Biology: CB* 15 (16): 1494–1500.
Ghildiyal, Megha, Herve Seitz, Michael D Horwich, Chengjian Li, Tingting Du, Soohyun Lee, Jia Xu, et al. 2008. "Endogenous siRNAs Derived from Transposons and mRNAs in Drosophila Somatic Cells." *Science* 320 (5879): 1077–81.
Giardine, Belinda, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. "Galaxy: A Platform for Interactive Large-Scale Genome Analysis." *Genome Research* 15 (10): 1451–55.
Goecks, Jeremy, Anton Nekrutenko, James Taylor, and Galaxy Team. 2010. "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences." *Genome Biology* 11 (8): R86.
Goriaux, Coline, Sophie Desset, Yoan Renaud, Chantal Vaury, and Emilie Brasset. 2014. "Transcriptional Properties and Splicing of the Flamenco piRNA Cluster." *EMBO Reports* 15 (4): 411–18.
Gou, Lan-Tao, Peng Dai, Jian-Hua Yang, Yuanchao Xue, Yun-Ping Hu, Yu Zhou, Jun-Yan Kang, et al. 2014. "Pachytene piRNAs Instruct Massive mRNA Elimination during Late Spermiogenesis." *Cell Research* 24 (6): 680–700.
Grimaud, Charlotte, Frédéric Bantignies, Manika Pal-Bhadra, Pallavi Ghana, Utpal Bhadra, and Giacomo Cavalli. 2006. "RNAi Components Are Required for Nuclear Clustering of Polycomb Group Response Elements." *Cell* 124 (5): 957–71.
Gu, Tingting, and Sarah C R Elgin. 2013. "Maternal Depletion of Piwi, a Component of the RNAi System, Impacts Heterochromatin Formation in Drosophila." *PLoS Genetics* 9 (9): e1003780.
Gunawardane, Lalith S, Kuniaki Saito, Kazumichi M Nishida, Keita Miyoshi, Yoshinori Kawamura, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. 2007. "A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5' End Formation in Drosophila." *Science* 315 (5818): 1587–90.
Ha, I, B Wightman, and G Ruvkun. 1996. "A Bulged Lin-4/lin-14 RNA Duplex Is Sufficient for Caenorhabditis Elegans Lin-14 Temporal Gradient Formation." *Genes & Development* 10 (23): 3041–50.
Haley, Benjamin, and Phillip D Zamore. 2004. "Kinetic Analysis of the RNAi Enzyme Complex." *Nature Structural & Molecular Biology* 11 (7): 599–606.
Hammond, Scott M, Sabrina Boettcher, Amy A Caudy, Ryuji Kobayashi, and Gregory J Hannon. 2001. "Argonaute2, a Link Between Genetic and Biochemical Analyses of RNAi." *Science* 293 (5532): 1146–50.
Handler, Dominik, Katharina Meixner, Manfred Pizka, Kathrin Lauss, Christopher Schmied, Franz Sebastian Gruber, and Julius Brennecke. 2013. "The Genetic Makeup of the Drosophila piRNA Pathway." *Molecular Cell* 50 (5): 762–77.
Hausser, Jean, and Mihaela Zavolan. 2014. "Identification and Consequences of miRNA-Target Interactions [mdash] beyond Repression of Gene Expression." *Nature Reviews. Genetics* 15 (9). Nature Publishing Group: 599–612.
Heinrich, Verena, Jens Stange, Thorsten Dickhaus, Peter Imkeller, Ulrike Krüger, Sebastian Bauer, Stefan Mundlos, Peter N Robinson, Jochen Hecht, and Peter M Krawitz. 2012. "The Allele Distribution in next-Generation Sequencing Data Sets Is Accurately Described as the Result of a Stochastic Branching Process." *Nucleic Acids Research* 40 (6): 2426–31.
Horwich, Michael D, Chengjian Li, Christian Matranga, Vasily Vagin, Gwen Farley, Peng Wang, and Phillip D Zamore. 2007. "The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC." *Current Biology: CB* 17 (14): 1265–72.
Huang, Xiao A, Hang Yin, Sarah Sweeney, Debasish Raha, Michael Snyder, and Haifan Lin. 2013. "A Major Epigenetic Programming Mechanism Guided by piRNAs." *Developmental Cell* 24 (5): 502–16.
Iwasaki, Shintaro, Maki Kobayashi, Mayuko Yoda, Yuriko Sakaguchi, Susumu Katsuma, Tsutomu Suzuki, and Yukihide Tomari. 2010. "Hsc70/Hsp90 Chaperone Machinery

Mediates ATP-Dependent RISC Loading of Small RNA Duplexes." *Molecular Cell* 39 (2): 292–99.
Ji, Lijuan, and Xuemei Chen. 2012. "Regulation of Small RNA Stability: Methylation and beyond." *Cell Research* 22 (4): 624–36.
Kataoka, Y, M Takeichi, and T Uemura. 2001. "Developmental Roles and Molecular Characterization of a Drosophila Homologue of Arabidopsis Argonaute1, the Founder of a Novel Gene Superfamily." *Genes to Cells: Devoted to Molecular & Cellular Mechanisms* 6 (4): 313–25.
Kawamata, Tomoko, Herve Seitz, and Yukihide Tomari. 2009. "Structural Determinants of miRNAs for RISC Loading and Slicer-Independent Unwinding." *Nature Structural & Molecular Biology* 16 (9): 953–60.
Kawaoka, Shinpei, Natsuko Izumi, Susumu Katsuma, and Yukihide Tomari. 2011. "3′ End Formation of PIWI-Interacting RNAs In Vitro." *Molecular Cell* 43 (6): 1015–22.
Kemp, Cordula, Stefanie Mueller, Akira Goto, Vincent Barbier, Simona Paro, François Bonnay, Catherine Dostert, et al. 2013. "Broad RNA Interference-Mediated Antiviral Immunity and Virus-Specific Inducible Responses in Drosophila." *Journal of Immunology* 190 (2): 650–58.
Kennerdell, J R, and R W Carthew. 1998. "Use of dsRNA-Mediated Genetic Interference to Demonstrate That Frizzled and Frizzled 2 Act in the Wingless Pathway." *Cell* 95 (7): 1017–26.
Kiuchi, Takashi, Hikaru Koga, Munetaka Kawamoto, Keisuke Shoji, Hiroki Sakai, Yuji Arai, Genki Ishihara, et al. 2014. "A Single Female-Specific piRNA Is the Primary Determiner of Sex in the Silkworm." *Nature* 509 (7502): 633–36.
Klenov, Mikhail S, Sergey A Lavrov, Alina P Korbut, Anastasia D Stolyarenko, Evgeny Y Yakushev, Michael Reuter, Ramesh S Pillai, and Vladimir A Gvozdev. 2014. "Impact of Nuclear Piwi Elimination on Chromatin State in Drosophila Melanogaster Ovaries." *Nucleic Acids Research* 42 (10): 6208–18.
Kudo, N, B Wolff, T Sekimoto, E P Schreiner, Y Yoneda, M Yanagida, S Horinouchi, and M Yoshida. 1998. "Leptomycin B Inhibition of Signal-Mediated Nuclear Export by Direct Binding to CRM1." *Experimental Cell Research* 242 (2): 540–47.
Lau, Nelson C, Nicolas Robine, Raquel Martin, Wei-Jen Chung, Yuzo Niki, Eugene Berezikov, and Eric C Lai. 2009. "Abundant Primary piRNAs, Endo-siRNAs, and microRNAs in a Drosophila Ovary Cell Line." *Genome Research* 19 (10): 1776–85.
Lazarus, Ross, Antony Kaspi, and Mark Ziemann. 2012. "Creating Reusable Tools from Scripts: The Galaxy Tool Factory." *Bioinformatics* 28 (23): 3139–40.
Lee, Eun Joo, Sourav Banerjee, Hongjun Zhou, Aruna Jammalamadaka, Mary Arcila, B S Manjunath, and Kenneth S Kosik. 2011. "Identification of piRNAs in the Central Nervous System." *RNA* 17 (6): 1090–99.
Lee, Rosalind C, Rhonda L Feinbaum, and Victor Ambros. 1993. "The C. Elegans Heterochronic Gene Lin-4 Encodes Small RNAs with Antisense Complementarity to Lin-14." *Cell* 75 (5): 843–54.
Lee, Yoontae, Kipyoung Jeon, Jun-Tae Lee, Sunyoung Kim, and V Narry Kim. 2002. "MicroRNA Maturation: Stepwise Processing and Subcellular Localization." *The EMBO Journal* 21 (17): 4663–70.
Lee, Yoontae, Minju Kim, Jinju Han, Kyu-Hyun Yeom, Sanghyuk Lee, Sung Hee Baek, and V Narry Kim. 2004. "MicroRNA Genes Are Transcribed by RNA Polymerase II." *The EMBO Journal* 23 (20): 4051–60.
Lee, Young Sik, Kenji Nakahara, John W Pham, Kevin Kim, Zhengying He, Erik J Sontheimer, and Richard W Carthew. 2004. "Distinct Roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA Silencing Pathways." *Cell* 117 (1): 69–81.
Lemieux, C, R Jorgensen, and Napoli. 1990. "Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in Trans." *The Plant Cell* 2 (4): 279–89.
Li, Wanhe, Lisa Prazak, Nabanita Chatterjee, Servan Grüninger, Lisa Krug, Delphine Theodorou, and Josh Dubnau. 2013. "Activation of Transposable Elements during

Aging and Neuronal Decline in Drosophila." *Nature Neuroscience* 16 (5): 529–31.
Lipardi, Concetta, and Bruce M Paterson. 2009. "Identification of an RNA-Dependent RNA Polymerase in Drosophila Involved in RNAi and Transposon Suppression." *Proceedings of the National Academy of Sciences* 106 (37): 15645–50.
Liu, Jidong, Michelle A Carmell, Fabiola V Rivas, Carolyn G Marsden, J Michael Thomson, Ji-Joon Song, Scott M Hammond, Leemor Joshua-Tor, and Gregory J Hannon. 2004. "Argonaute2 Is the Catalytic Engine of Mammalian RNAi." *Science* 305 (5689): 1437–41.
Liu, Li, Hongying Qi, Jianquan Wang, and Haifan Lin. 2011. "PAPI, a Novel TUDOR-Domain Protein, Complexes with AGO3, ME31B and TRAL in the Nuage to Silence Transposition." *Development* 138 (9): 1863–73.
Lucchetta, Elena M, Richard W Carthew, and Rustem F Ismagilov. 2009. "The Endo-siRNA Pathway Is Essential for Robust Development of the Drosophila Embryo." *PloS One* 4 (10): e7576.
Léger, P, E Lara, B Jagla, O Sismeiro, Z Mansuroglu, J Y Coppée, E Bonnefoy, and M Bouloy. 2013. "Dicer-2- and Piwi-Mediated RNA Interference in Rift Valley Fever Virus-Infected Mosquito Cells." *Journal of Virology* 87 (3): 1631–48.
MacRae, Ian J, Kaihong Zhou, and Jennifer A Doudna. 2007. "Structural Determinants of RNA Recognition and Cleavage by Dicer." *Nature Structural & Molecular Biology* 14 (10): 934–40.
Malone, Colin D, Julius Brennecke, Monica Dus, Alexander Stark, W Richard McCombie, Ravi Sachidanandam, and Gregory J Hannon. 2009. "Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary." *Cell* 137 (3): 522–35.
Mani, Sneha Ramesh, Heather Megosh, and Haifan Lin. 2014. "PIWI Proteins Are Essential for Early Drosophila Embryogenesis." *Developmental Biology* 385 (2): 340–49.
Marques, João Trindade, Kevin Kim, Pei-Hsuan Wu, Trevis M Alleyne, Nadereh Jafari, and Richard W Carthew. 2010. "Loqs and R2D2 Act Sequentially in the siRNA Pathway in Drosophila." *Nature Structural & Molecular Biology* 17 (1): 24–30.
Martinez, Javier, Agnieszka Patkaniowska, Henning Urlaub, Reinhard Lührmann, and Thomas Tuschl. 2002. "Single-Stranded Antisense siRNAs Guide Target RNA Cleavage in RNAi." *Cell* 110 (5): 563–74.
McCLINTOCK, B. 1950. "The Origin and Behavior of Mutable Loci in Maize." *Proceedings of the National Academy of Sciences of the United States of America* 36 (6): 344–55.
McClintock, B. 1953. "Induction of Instability at Selected Loci in Maize." *Genetics* 38 (6): 579–99.
Mirkovic-Hösle, Milijana, and Klaus Förstemann. 2014. "Transposon Defense by Endo-siRNAs, piRNAs and Somatic pilRNAs in Drosophila: Contributions of Loqs-PD and R2D2." *PloS One* 9 (1): e84994.
Misquitta, L, and B M Paterson. 1999. "Targeted Disruption of Gene Function in Drosophila by RNA Interference (RNA-I): A Role for Nautilus in Embryonic Somatic Muscle Formation." *Proceedings of the National Academy of Sciences of the United States of America* 96 (4): 1451–56.
Miyoshi, Tomohiro, Akiko Takeuchi, Haruhiko Siomi, and Mikiko C Siomi. 2010. "A Direct Role for Hsp90 in Pre-RISC Formation in Drosophila." *Nature Structural & Molecular Biology* 17 (8): 1024–26.
Mohn, Fabio, Grzegorz Sienski, Dominik Handler, and Julius Brennecke. 2014. "The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila." *Cell* 157 (6): 1364–79.
Moshkovich, Nellie, and Elissa P Lei. 2010. "HP1 Recruitment in the Absence of Argonaute Proteins in Drosophila." *PLoS Genetics* 6 (3): e1000880.
Moshkovich, Nellie, Parul Nisha, Patrick J Boyle, Brandi A Thompson, Ryan K Dale, and Elissa P Lei. 2011. "RNAi-Independent Role for Argonaute2 in CTCF/CP190 Chromatin Insulator Function." *Genes & Development* 25 (16): 1686–1701.
Muerdter, Felix, Paloma M Guzzardo, Jesse Gillis, Yicheng Luo, Yang Yu, Caifu Chen, Richard Fekete, and Gregory J Hannon. 2013. "A Genome-Wide RNAi Screen Draws a

Genetic Framework for Transposon Control and Primary piRNA Biogenesis in Drosophila." *Molecular Cell* 50 (5): 736–48.
Muller, H J. 1930. "Types of Visible Variations Induced by X-Rays inDrosophila." *Journal of Genetics* 22 (3). Springer India: 299–334.
Murota, Yukiko, Hirotsugu Ishizu, Shinichi Nakagawa, Yuka W Iwasaki, Shinsuke Shibata, Miharu K Kamatani, Kuniaki Saito, Hideyuki Okano, Haruhiko Siomi, and Mikiko C Siomi. 2014. "Yb Integrates piRNA Intermediates and Processing Factors into Perinuclear Bodies to Enhance piRISC Assembly." *Cell Reports* 8 (1): 103–13.
Mével-Ninio, Maryvonne, Alain Pelisson, Jennifer Kinder, Ana Regina Campos, and Alain Bucheton. 2007. "The Flamenco Locus Controls the Gypsy and ZAM Retroviruses and Is Required for Drosophila Oogenesis." *Genetics* 175 (4): 1615–24.
Ni, Jian-Quan, Michele Markstein, Richard Binari, Barret Pfeiffer, Lu-Ping Liu, Christians Villalta, Matthew Booker, Lizabeth Perkins, and Norbert Perrimon. 2008. "Vector and Parameters for Targeted Transgenic RNA Interference in Drosophila Melanogaster." *Nature Methods* 5 (1): 49–51.
Nishida, Kazumichi M, Keita Miyoshi, Akiyo Ogino, Tomohiro Miyoshi, Haruhiko Siomi, and Mikiko C Siomi. 2013. "Roles of R2D2, a Cytoplasmic D2 Body Component, in the Endogenous siRNA Pathway in Drosophila." *Molecular Cell* 49 (4): 680–91.
Nishimasu, Hiroshi, Hirotsugu Ishizu, Kuniaki Saito, Satoshi Fukuhara, Miharu K Kamatani, Luc Bonnefond, Naoki Matsumoto, et al. 2012. "Structure and Function of Zucchini Endoribonuclease in piRNA Biogenesis." *Nature* 491 (7423): 284–87.
Nykänen, A, B Haley, and P D Zamore. 2001. "ATP Requirements and Small Interfering RNA Structure in the RNA Interference Pathway." *Cell* 107 (3): 309–21.
Néron, Bertrand, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. 2009. "Mobyle: A New Full Web Bioinformatics Framework." *Bioinformatics* 25 (22): 3005–11.
Okamura, Katsutomo, Nicolas Robine, Ying Liu, Qinghua Liu, and Eric C Lai. 2011. "R2D2 Organizes Small Regulatory RNA Pathways in Drosophila." *Molecular and Cellular Biology* 31 (4): 884–96.
Olivieri, Daniel, Martina M Sykora, Ravi Sachidanandam, Karl Mechtler, and Julius Brennecke. 2010. "An in Vivo RNAi Assay Identifies Major Genetic and Cellular Requirements for Primary piRNA Biogenesis in Drosophila." *The EMBO Journal* 29 (19): 3301–17.
Olsen, P H, and V Ambros. 1999. "The Lin-4 Regulatory RNA Controls Developmental Timing in Caenorhabditis Elegans by Blocking LIN-14 Protein Synthesis after the Initiation of Translation." *Developmental Biology* 216 (2): 671–80.
Pal-Bhadra, M, U Bhadra, and J A Birchler. 1997. "Cosuppression in Drosophila: Gene Silencing of Alcohol Dehydrogenase by White-Adh Transgenes Is Polycomb Dependent." *Cell* 90 (3): 479–90.
Pal-Bhadra, Manika, Boris A Leibovitch, Sumit G Gandhi, Madhusudana Rao, Utpal Bhadra, James A Birchler, and Sarah C R Elgin. 2004. "Heterochromatic Silencing and HP1 Localization in Drosophila Are Dependent on the RNAi Machinery." *Science* 303 (5658): 669–72.
Pandey, Radha Raman, and Ramesh S Pillai. 2014. "Primary piRNA Biogenesis: Caught up in a Maelstrom." *The EMBO Journal* 33 (18): 1979–80.
Pasquinelli, A E, B J Reinhart, F Slack, M Q Martindale, M I Kuroda, B Maller, D C Hayward, et al. 2000. "Conservation of the Sequence and Temporal Expression of Let-7 Heterochronic Regulatory RNA." *Nature* 408 (6808): 86–89.
Pelisson, A, S U Song, N Prud'Homme, and P A Smith. 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*. ncbi.nlm.nih.gov. http://www.ncbi.nlm.nih.gov/pmc/articles/pmc395367/.
Perrat, Paola N, Shamik DasGupta, Jie Wang, William Theurkauf, Zhiping Weng, Michael Rosbash, and Scott Waddell. 2013. "Transposition-Driven Genomic Heterogeneity in the Drosophila Brain." *Science* 340 (6128): 91–95.

Post, Christina, Josef P Clark, Yuliya A Sytnikova, Gung-Wei Chirn, and Nelson C Lau. 2014. “The Capacity of Target Silencing by Drosophila PIWI and piRNAs.” *RNA* 20 (12): 1977–86.
Prud’homme, N, M Gans, M Masson, C Terzian, and A Bucheton. 1995. “Flamenco, a Gene Controlling the Gypsy Retrovirus of Drosophila Melanogaster.” *Genetics* 139 (2): 697–711.
Pérez, Fernando, and Brian E Granger. 2007. “IPython: A System for Interactive Scientific Computing.” *Computing in Science & Engineering* 9 (3). AIP Publishing: 21–29.
Rajasethupathy, Priyamvada, Igor Antonov, Robert Sheridan, Sebastian Frey, Chris Sander, Thomas Tuschl, and Eric R Kandel. 2012. “A Role for Neuronal piRNAs in the Epigenetic Control of Memory-Related Synaptic Plasticity.” *Cell* 149 (3): 693–707.
Rangan, Prashanth, Colin D Malone, Caryn Navarro, Sam P Newbold, Patrick S Hayes, Ravi Sachidanandam, Gregory J Hannon, and Ruth Lehmann. 2011. “piRNA Production Requires Heterochromatin Formation in Drosophila.” *Current Biology: CB* 21 (16): 1373–79.
Reich, Michael, Ted Liefeld, Joshua Gould, Jim Lerner, Pablo Tamayo, and Jill P Mesirov. 2006. “GenePattern 2.0.” *Nature Genetics* 38 (5): 500–501.
Reinhardt, Anita, Sébastien Feuillette, Marlène Cassar, Céline Callens, Hélène Thomassin, Serge Birman, Magalie Lecourtois, Christophe Antoniewski, and Hervé Tricoire. 2012. “Lack of miRNA Misregulation at Early Pathological Stages in Drosophila Neurodegenerative Disease Models.” *Frontiers in Genetics* 3 (October): 226.
Rivas, Fabiola V, Niraj H Tolia, Ji-Joon Song, Juan P Aragon, Jidong Liu, Gregory J Hannon, and Leemor Joshua-Tor. 2005. “Purified Argonaute2 and an siRNA Form Recombinant Human RISC.” *Nature Structural & Molecular Biology* 12 (4): 340–49.
Robert, Valérie, Nicole Prud’homme, Alexander Kim, Alain Bucheton, and Alain Pélisson. 2001. “Characterization of the Flamenco Region of the Drosophila Melanogaster Genome.” *Genetics* 158 (2): 701–13.
Robine, Nicolas, Nelson C Lau, Sudha Balla, Zhigang Jin, Katsutomo Okamura, Satomi Kuramochi-Miyagawa, Michael D Blower, and Eric C Lai. 2009. “A Broadly Conserved Pathway Generates 3′UTR-Directed Primary piRNAs.” *Current Biology: CB* 19 (24): 2066–76.
Roignant, Jean-Yves, Clément Carré, Bruno Mugat, Dimitri Szymczak, Jean-Antoine Lepesant, and Christophe Antoniewski. 2003. “Absence of Transitive and Systemic Pathways Allows Cell-Specific and Isoform-Specific RNAi in Drosophila.” *RNA* 9 (3): 299–308.
Rouget, Christel, Catherine Papin, Anthony Boureux, Anne-Cécile Meunier, Bénédicte Franco, Nicolas Robine, Eric C Lai, Alain Pelisson, and Martine Simonelig. 2010. “Maternal mRNA Deadenylation and Decay by the piRNA Pathway in the Early Drosophila Embryo.” *Nature* 467 (7319): 1128–32.
Rozhkov, Nikolay V, Molly Hammell, and Gregory J Hannon. 2013. “Multiple Roles for Piwi in Silencing Drosophila Transposons.” *Genes & Development* 27 (4): 400–412.
Ruby, J Graham, Calvin H Jan, and David P Bartel. 2007. “Intronic microRNA Precursors That Bypass Drosha Processing.” *Nature* 448 (7149): 83–86.
Saito, Kuniaki, Sachi Inagaki, Toutai Mituyama, Yoshinori Kawamura, Yukiteru Ono, Eri Sakota, Hazuki Kotani, Kiyoshi Asai, Haruhiko Siomi, and Mikiko C Siomi. 2009. “A Regulatory Circuit for Piwi by the Large Maf Gene Traffic Jam in Drosophila.” *Nature* 461 (7268): 1296–99.
Saito, Kuniaki, Hirotsugu Ishizu, Miharu Komai, Hazuki Kotani, Yoshinori Kawamura, Kazumichi M Nishida, Haruhiko Siomi, and Mikiko C Siomi. 2010. “Roles for the Yb Body Components Armitage and Yb in Primary piRNA Biogenesis in Drosophila.” *Genes & Development* 24 (22): 2493–98.
Saito, Kuniaki, Kazumichi M Nishida, Tomoko Mori, Yoshinori Kawamura, Keita Miyoshi, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. 2006. “Specific Association of Piwi with rasiRNAs Derived from Retrotransposon and Heterochromatic Regions in the Drosophila Genome.” *Genes & Development* 20 (16): 2214–22.

Saito, Kuniaki, Yuriko Sakaguchi, Takeo Suzuki, Tsutomu Suzuki, Haruhiko Siomi, and Mikiko C Siomi. 2007. “Pimet, the Drosophila Homolog of HEN1, Mediates 2′-O-Methylation of Piwi- Interacting RNAs at Their 3′ Ends.” *Genes & Development* 21 (13): 1603–8.

Saleh, Maria-Carla, Michel Tassetto, Ronald P van Rij, Bertsy Goic, Valérie Gausson, Bassam Berry, Caroline Jacquier, Christophe Antoniewski, and Raul Andino. 2009. “Antiviral Immunity in Drosophila Requires Systemic RNA Interference Spread.” *Nature* 458 (7236): 346–50.

Sarot, Emeline, Genevieve Payen-Groschene, Alain Bucheton, and Alain Pelisson. 2004. “Evidence for a Piwi-Dependent RNA Silencing of the Gypsy Endogenous Retrovirus by the Drosophila Melanogaster Flamenco Gene.” *Genetics* 166 (3): 1313–21.

Sement, François Michaël, Emilie Ferrier, Hélène Zuber, Rémy Merret, Malek Alioua, Jean-Marc Deragon, Cécile Bousquet-Antonelli, Heike Lange, and Dominique Gagliardi. 2013. “Uridylation Prevents 3’ Trimming of Oligoadenylated mRNAs.” *Nucleic Acids Research* 41 (14): 7115–27.

Shpiz, Sergey, Sergei Ryazansky, Ivan Olovnikov, Yuri Abramov, and Alla Kalmykova. 2014. “Euchromatic Transposon Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the Drosophila Germline.” *PLoS Genetics* 10 (2): e1004138.

Sienski, Grzegorz, Derya Dönertas, and Julius Brennecke. 2012. “Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression.” *Cell* 151 (5): 964–80.

Szakmary, Akos, Mary Reedy, Hongying Qi, and Haifan Lin. 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): 613–27.

Taliaferro, J Matthew, Julie L Aspden, Todd Bradley, Dhruv Marwha, Marco Blanchette, and Donald C Rio. 2013. “Two New and Distinct Roles for Drosophila Argonaute-2 in the Nucleus: Alternative Pre-mRNA Splicing and Transcriptional Repression.” *Genes & Development* 27 (4): 378–89.

Thomas, Adrien Le, Alicia K Rogers, Alexandre Webster, Georgi K Marinov, Susan E Liao, Edward M Perkins, Junho K Hur, Alexei A Aravin, and Katalin Fejes Tóth. 2013. “Piwi Induces piRNA-Guided Transcriptional Silencing and Establishment of a Repressive Chromatin State.” *Genes & Development* 27 (4): 390–99.

Thomas, Adrien Le, Evelyn Stuwe, Sisi Li, Jiamu Du, Georgi Marinov, Nikolay Rozhkov, Yung-Chia Ariel Chen, et al. 2014. “Transgenerationally Inherited piRNAs Trigger piRNA Biogenesis by Changing the Chromatin of piRNA Clusters and Inducing Precursor Processing.” *Genes & Development* 28 (15): 1667–80.

Tian, Yuan, Dhirendra K Simanshu, Manuel Ascano, Ruben Diaz-Avalos, Ah Young Park, Stefan A Juranek, William J Rice, et al. 2011. “Multimeric Assembly and Biochemical Characterization of the Trax–translin Endonuclease Complex.” *Nature Structural & Molecular Biology* 18 (6): 658–64.

Tomari, Yukihide, Tingting Du, and Phillip D Zamore. 2007. “Sorting of Drosophila Small Silencing RNAs.” *Cell* 130 (2): 299–308.

Vagin, Vasily V, Alla Sigova, Chengjian Li, Hervé Seitz, Vladimir Gvozdev, and Phillip D Zamore. 2006. “A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline.” *Science* 313 (5785): 320–24.

Van Rij, R P, and R Andino. 2008. “The Complex Interactions of Viruses and the RNAi Machinery: A Driving Force in Viral Evolution.” *Origin and Evolution of Viruses*. books.google.com. http://books.google.com/books?hl=en&lr=&id=6JzPVg_QF2cC&oi=fnd&pg=PA161&dq=complex+interactions+viruses+RNAi+machinery+driving+force+viral+evolution+Van+Rij+Andino&ots=fGP9VOY2gc&sig=vCeXX2w6wboWSevvTi79G3rZDGo.

Van Rij, R P, M-C Saleh, B Berry, C Foo, A Houk, C Antoniewski, and R Andino. 2006. “The RNA Silencing Endonuclease Argonaute 2 Mediates Specific Antiviral Immunity in Drosophila Melanogaster.” *Genes & Development* 20 (21): 2985–95.

Vandormael-Pournin, S, J Y Coppée, Q Zhou, and E Heard. 2012. “Naive and Primed

Murine Pluripotent Stem Cells Have Distinct miRNA Expression Profiles." *RNA* rnajournal.cshlp.org. http://rnajournal.cshlp.org/content/18/2/253.short.
Vodovar, Nicolas, Alfred W Bronkhorst, Koen W R van Cleef, Pascal Miesen, Hervé Blanc, Ronald P van Rij, and Maria-Carla Saleh. 2012. "Arbovirus-Derived piRNAs Exhibit a Ping-Pong Signature in Mosquito Cells." *PloS One* 7 (1): e30861.
Wang, Sidney H, and Sarah C R Elgin. 2011. "Drosophila Piwi Functions Downstream of piRNA Production Mediating a Chromatin-Based Transposon Silencing Mechanism in Female Germ Line." *Proceedings of the National Academy of Sciences* 108 (52): 21164–69.
Wang, Xiao-Hong, Roghiyh Aliyari, Wan-Xiang Li, Hong-Wei Li, Kevin Kim, Richard Carthew, Peter Atkinson, and Wei Ding. 2006. "RNA Interference Directs Innate Immunity Against Viruses in Adult Drosophila." *Science*, March, 1125694.
Welker, Noah C, Tuhin S Maity, Xuecheng Ye, P Joseph Aruscavage, Ammie A Krauchuk, Qinghua Liu, and Brenda L Bass. 2011. "Dicer's Helicase Domain Discriminates dsRNA Termini to Promote an Altered Reaction Mode." *Molecular Cell* 41 (5): 589–99.
Wen, Jiayu, Hong Duan, Fernando Bejarano, Katsutomo Okamura, Lacramioara Fabian, Julie A Brill, Diane Bortolamiol-Becet, Raquel Martin, J Graham Ruby, and Eric C Lai. 2014. "Adaptive Regulation of Testis Gene Expression and Control of Male Fertility by the Drosophila Harpin RNA Pathway." *Molecular Cell*, December. doi:10.1016/j.molcel.2014.11.025.
Wolstencroft, Katherine, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, et al. 2013. "The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web or in the Cloud." *Nucleic Acids Research* 41 (Web Server issue): W557–61.
Wu, Ligang, Jihua Fan, and Joel G Belasco. 2006. "MicroRNAs Direct Rapid Deadenylation of mRNA." *Proceedings of the National Academy of Sciences of the United States of America* 103 (11): 4034–39.
Xie, Zhixin, Edwards Allen, April Wilken, and James C Carrington. 2005. "DICER-LIKE 4 Functions in Trans-Acting Small Interfering RNA Biogenesis and Vegetative Phase Change in Arabidopsis Thaliana." *Proceedings of the National Academy of Sciences of the United States of America* 102 (36): 12984–89.
Xiol, Jordi, Pietro Spinelli, Maike A Laussmann, David Homolka, Zhaolin Yang, Elisa Cora, Yohann Couté, et al. 2014. "RNA Clamping by Vasa Assembles a piRNA Amplifier Complex on Transposon Transcripts." *Cell* 157 (7): 1698–1711.
Yan, Zheng, Hai Yang Hu, Xi Jiang, Vera Maierhofer, Elena Neb, Liu He, Yuhui Hu, et al. 2011. "Widespread Expression of piRNA-like Molecules in Somatic Tissues." *Nucleic Acids Research* 39 (15): 6596–6607.
Yi, Rui, Yi Qin, Ian G Macara, and Bryan R Cullen. 2003. "Exportin-5 Mediates the Nuclear Export of Pre-microRNAs and Short Hairpin RNAs." *Genes & Development* 17 (24): 3011–16.
Yoshikawa, Manabu, Angela Peragine, Mee Yeon Park, and R Scott Poethig. 2005. "A Pathway for the Biogenesis of Trans-Acting siRNAs in Arabidopsis." *Genes & Development* 19 (18): 2164–75.
Zambon, Robert A, Vikram N Vakharia, and Louisa P Wu. 2006. "RNAi Is an Antiviral Immune Response against a dsRNA Virus in Drosophila Melanogaster." *Cellular Microbiology* 8 (5): 880–89.
Zamore, Phillip D, Thomas Tuschl, Phillip A Sharp, and David P Bartel. 2000. "RNAi: Double-Stranded RNA Directs the ATP-Dependent Cleavage of mRNA at 21 to 23 Nucleotide Intervals." *Cell* 101 (1): 25–33.
Zhang, Fan, Jie Wang, Jia Xu, Zhao Zhang, Birgit S Koppetsch, Nadine Schultz, Thom Vreven, et al. 2012. "UAP56 Couples piRNA Clusters to the Perinuclear Transposon Silencing Machinery." *Cell* 151 (4): 871–84.
Zhang, Zhao, Jie Wang, Nadine Schultz, Fan Zhang, Swapnil S Parhad, Shikui Tu, Thom Vreven, Phillip D Zamore, Zhiping Weng, and William E Theurkauf. 2014. "The HP1 Homolog Rhino Anchors a Nuclear Complex That Suppresses piRNA Precursor

Splicing." *Cell* 157 (6): 1353–63.
Zhou, R, I Hotta, A Denli, P Hong, N Perrimon, and G Hannon. 2008. "Comparative Analysis of Argonaute-Dependent Small RNA Pathways in Drosophila." *Molecular Cell* 32 (4): 592–99.
Zhuang, Jiali, Jie Wang, William Theurkauf, and Zhiping Weng. 2014. "TEMP: A Computational Method for Analyzing Transposable Element Polymorphism in Populations." *Nucleic Acids Research* 42 (11): 6826–38.

Annexes

1. Annex 1: Supplemental Tables to article 1
2. Annex 2: Example IPython Notebook showing influence of TE insertions on gene expression

Annex 1

Supplemental table S1

Original data: http://lbcd41.snv.jussieu.fr/galaxy/u/marius-ged/h/compare-simulation-with-real-differences

Differential expression testing using EdgeR between ping-pong negative and ping-pong positive libraries

The top 10 differentially detected miRNA between ping-pong negative libraries

and ping-pong negative libraries supplemented with 2% testicular reads are highlighted in red (Sheet 2)

Name	logFC	logCPM	LR	PValue	adj.p.value	Dispersion	totreads
dme-mir-31b	2.57E+00	6.10E+00	6.44E+01	1.03E-15	2.25E-13	1.34E-01	3.86E+03
dme-mir-959	3.76E+00	3.61E+00	4.95E+01	1.98E-12	2.16E-10	3.19E-01	6.71E+02
dme-mir-991	2.51E+00	3.22E+00	3.81E+01	6.59E-10	4.79E-08	1.85E-01	4.96E+02
dme-mir-983-2	2.78E+00	2.65E+00	3.62E+01	1.75E-09	9.56E-08	2.18E-01	3.22E+02
dme-mir-961	2.99E+00	2.69E+00	3.44E+01	4.40E-09	1.92E-07	2.85E-01	3.38E+02
dme-mir-310	3.81E+00	5.06E-01	3.18E+01	1.68E-08	5.36E-07	1.80E-01	5.40E+01
dme-mir-963	3.28E+00	2.01E+00	3.18E+01	1.72E-08	5.36E-07	3.20E-01	1.95E+02
dme-mir-960	2.72E+00	4.18E+00	3.15E+01	2.04E-08	5.56E-07	2.97E-01	1.03E+03
dme-mir-985	3.86E+00	3.66E+00	2.99E+01	4.65E-08	1.13E-06	5.82E-01	6.96E+02
dme-mir-2494	4.26E+00	6.81E-02	2.79E+01	1.28E-07	2.65E-06	1.59E-01	3.20E+01
dme-mir-983-1	2.48E+00	2.54E+00	2.78E+01	1.34E-07	2.65E-06	2.32E-01	2.92E+02
dme-mir-982	1.88E+00	3.77E+00	2.63E+01	2.85E-07	5.00E-06	1.61E-01	7.42E+02
dme-mir-312	2.18E+00	3.18E+00	2.63E+01	2.98E-07	5.00E-06	2.09E-01	4.84E+02
dme-mir-iab-8	3.52E+00	5.98E-01	2.37E+01	1.12E-06	1.74E-05	3.05E-01	5.80E+01
dme-mir-375	1.21E+00	9.09E+00	2.27E+01	1.94E-06	2.57E-05	9.16E-02	3.04E+04
dme-mir-977	2.70E+00	3.42E+00	2.26E+01	1.97E-06	2.57E-05	4.01E-01	5.84E+02
dme-mir-976	2.92E+00	5.33E-01	2.26E+01	2.00E-06	2.57E-05	1.86E-01	5.40E+01
dme-mir-984	1.94E+00	2.56E+00	2.20E+01	2.73E-06	3.31E-05	1.86E-01	3.05E+02
dme-mir-956	3.35E+00	8.13E+00	2.06E+01	5.75E-06	6.59E-05	7.09E-01	1.58E+04
dme-mir-314	3.66E+00	7.04E+00	1.94E+01	1.04E-05	1.13E-04	8.74E-01	7.46E+03
dme-mir-962	2.67E+00	5.17E-01	1.84E+01	1.79E-05	1.86E-04	2.14E-01	5.20E+01
dme-mir-4966	2.82E+00	9.68E-02	1.74E+01	3.01E-05	2.92E-04	1.58E-01	3.30E+01
dme-mir-974	1.84E+00	1.61E+00	1.74E+01	3.08E-05	2.92E-04	1.70E-01	1.42E+02
dme-mir-1015	2.91E+00	3.39E-01	1.50E+01	1.06E-04	9.62E-04	3.74E-01	4.50E+01
dme-mir-311	2.16E+00	3.77E+00	1.48E+01	1.19E-04	1.04E-03	4.11E-01	7.50E+02
dme-mir-4914	2.90E+00	-2.59E-01	1.32E+01	2.84E-04	2.38E-03	1.62E-01	2.10E+01

dme-mir-973	2.03E+00	8.04E-01	1.09E+01	9.60E-04	7.75E-03	3.15E-01	6.90E+01

http://lbcd41.snv.jussieu.fr/galaxy/u/marius-ged/h/compare-simulation-with-real-differences

Differential expression analysis between ping-pong negative libraries

and ping-pong negative libraries supplemented with 2% testes reads

The 19 differentially detcted miRNA between ping-pong negative libraries

and ping-pong positive libraries (Sheet 1) are highlighted in red

Name	logFC	logCPM	LR	PValue	adj.p.value	Dispersion	totreads
dme-mir-964	6.30E+00	5.74E+00	1.18E+03	3.54E-258	7.29E-256	4.31E-04	9.87E+02
dme-mir-959	5.61E+00	5.57E+00	9.99E+02	3.01E-219	3.10E-217	3.89E-04	8.76E+02
dme-mir-962	8.52E+00	5.28E+00	9.31E+02	1.79E-204	1.23E-202	2.54E-04	7.11E+02
dme-mir-984	5.42E+00	5.46E+00	9.18E+02	1.08E-201	5.54E-200	3.54E-04	8.15E+02
dme-mir-985	4.96E+00	5.10E+00	6.78E+02	1.44E-149	5.94E-148	2.41E-04	6.28E+02
dme-mir-974	5.43E+00	4.66E+00	5.17E+02	1.63E-114	5.60E-113	1.42E-04	4.52E+02
dme-mir-991	4.29E+00	4.83E+00	4.97E+02	4.35E-110	1.28E-108	1.32E-04	5.10E+02
dme-mir-960	3.11E+00	4.85E+00	3.69E+02	2.75E-82	7.09E-81	1.73E-04	5.24E+02
dme-mir-961	4.48E+00	4.16E+00	3.17E+02	6.53E-71	1.49E-69	6.58E-05	3.11E+02
dme-mir-312	3.19E+00	4.22E+00	2.37E+02	1.41E-53	2.90E-52	1.02E-04	3.24E+02
dme-mir-997	5.58E+00	3.55E+00	2.31E+02	4.06E-52	7.61E-51	3.34E-07	1.95E+02
dme-mir-976	7.32E+00	3.15E+00	1.90E+02	2.95E-43	5.06E-42	7.77E-07	1.41E+02
dme-mir-977	2.84E+00	3.81E+00	1.54E+02	2.45E-35	3.88E-34	7.28E-06	2.41E+02
dme-mir-31b	1.35E+00	5.36E+00	1.46E+02	1.37E-33	2.02E-32	3.28E-04	7.61E+02
dme-mir-978	6.69E+00	2.67E+00	1.23E+02	1.26E-28	1.73E-27	1.12E-07	9.20E+01
dme-mir-989	2.63E+00	3.57E+00	1.15E+02	9.46E-27	1.22E-25	3.54E-07	1.98E+02
dme-mir-963	3.46E+00	2.97E+00	9.96E+01	1.85E-23	2.24E-22	5.76E-09	1.21E+02
dme-mir-983-2	2.46E+00	2.87E+00	5.94E+01	1.26E-14	1.45E-13	2.68E-08	1.11E+02
dme-mir-992	5.59E+00	1.89E+00	5.64E+01	5.79E-14	6.28E-13	6.19E-04	4.20E+01
dme-mir-313	2.68E+00	2.52E+00	4.96E+01	1.92E-12	1.92E-11	1.40E-04	8.10E+01
dme-mir-982	1.39E+00	3.81E+00	4.95E+01	1.96E-12	1.92E-11	2.40E-05	2.37E+02
dme-mir-983-1	2.06E+00	2.85E+00	4.48E+01	2.14E-11	2.00E-10	3.17E-08	1.09E+02
dme-mir-303	5.06E+00	1.59E+00	3.88E+01	4.65E-10	4.16E-09	1.54E-06	2.90E+01
dme-mir-318	1.18E+00	3.60E+00	3.15E+01	1.98E-08	1.70E-07	3.94E-07	2.03E+02
dme-mir-4966	3.91E+00	1.59E+00	3.04E+01	3.43E-08	2.83E-07	9.22E-04	2.90E+01
dme-mir-310	3.80E+00	1.54E+00	2.79E+01	1.29E-07	1.03E-06	1.65E-06	2.70E+01
dme-mir-311	1.19E+00	3.23E+00	2.40E+01	9.81E-07	7.49E-06	7.38E-08	1.51E+02
dme-mir-973	2.29E+00	1.96E+00	2.32E+01	1.44E-06	1.06E-05	5.31E-04	4.60E+01
dme-mir-979	4.25E+00	1.22E+00	2.17E+01	3.23E-06	2.29E-05	2.42E-06	1.60E+01
dme-mir-375	2.10E-01	8.56E+00	1.82E+01	2.04E-05	1.40E-04	2.05E-03	7.13E+03
dme-mir-316	3.18E-01	6.43E+00	1.81E+01	2.12E-05	1.41E-04	6.00E-04	1.61E+03
dme-mir-79	1.89E-01	8.04E+00	1.67E+01	4.42E-05	2.84E-04	7.49E-04	5.02E+03
dme-mir-2498	3.32E+00	9.24E-01	1.08E+01	1.03E-03	6.42E-03	3.09E-04	8.00E+00
dme-mir-999	-9.88E-02	1.46E+01	1.07E+01	1.07E-03	6.51E-03	1.50E-03	4.71E+05
dme-mir-278	-9.33E-02	1.26E+01	9.90E+00	1.65E-03	9.72E-03	1.36E-03	1.21E+05

How important is it to know the transposon insertion sites?

I will focus on the OSC data from the Sienski et al (2012, Cell) paper. I will look at the trend for overexpressed genes to have transposon insertions. On the one hand i will just take the group1 transposon insertions from the Dmel-5.49 genome ('group1_flybase.bed'), and compare it with the group1 insertions as sequenced in the paper (group1_wo_het.bed).

In [1]:
```
cd /home/marius/ipython_coding/DESeq_results/
```

```
/home/marius/ipython_coding/DESeq_results
```

In [2]:
```
from metaseq.results_table import DESeqResults
from gffutils.helpers import FeatureNotFoundError
import matplotlib
import gffutils
import pandas
import os
import pybedtools

def convert_ids(list):
    name_list=[]
    for id in list:
        try:
            gene=d.db[id]
            name=str(gene['Name'][0])
            name_list.append(name)
        except FeatureNotFoundError:
            name_list.append(id)
    return name_list
```

fbgn.db is a gffutils database constructed from fbgn.gff (Dmel-5.49-all-no-analysis.gff), which contains all gene annotations, positions etc. OSC.txt is the result of a DESeq differential expression for the RNAseq in piwi and GFP knockdown OSC cells.

In [3]:
```
if not os.path.exists('fbgn.db'):
    gffutils.create_db(data='fbgn.gff', dbfn='fbgn.db')
d = DESeqResults('OSC.txt', db='fbgn.db')
TEs_flybase=pybedtools.BedTool('group1_flybase.bed')
TEs_brennecke=pybedtools.BedTool('group1_wo_het.bed')
```

Now I'll filter out extreme values from the DESeq results

In [4]:
```
pandas.options.mode.chained_assignment = None
d=d[d.data['foldChange'] > 0]
d=d[d.data['foldChange'] != -inf]
d=d[d.data['foldChange'] != inf]
d=d[d.data['baseMean'] > 20]
```

Let's implement a filter to look at genes with fold2 Change higher than 4 (what brennecke used in the paper ...) and group1 transposon insertions inside the gene (according to the flybase transposon gff).

In [5]:
```
up=(d.data['foldChange'] > 4)*1
with_insertion=d.genes_with_peak(TEs_flybase)*1
up_with_insertion=up+with_insertion == 2
```

In [6]:
```
up_names=d[up_with_insertion].data.index.tolist()
up_names=convert_ids(up_names)
#down_names=d[d.downregulated(0.001)].data.index.tolist()
#down_names=convert_ids(down_names)
genes_to_highlight=[(d.downregulated(0.05), dict(color='0.75', alpha=0.5,
 names=[], marker="s")),
        (d.upregulated(0.1), dict(color='0.75', alpha=0.5, names=[], mark
er="s")),
        (d.genes_with_peak(TEs_flybase), dict(color='r', names=up_names,
alpha=1),)];
```

In [7]:
```
fig=figure(figsize=(16,10))
```

```
<matplotlib.figure.Figure at 0x54de1d0>
```

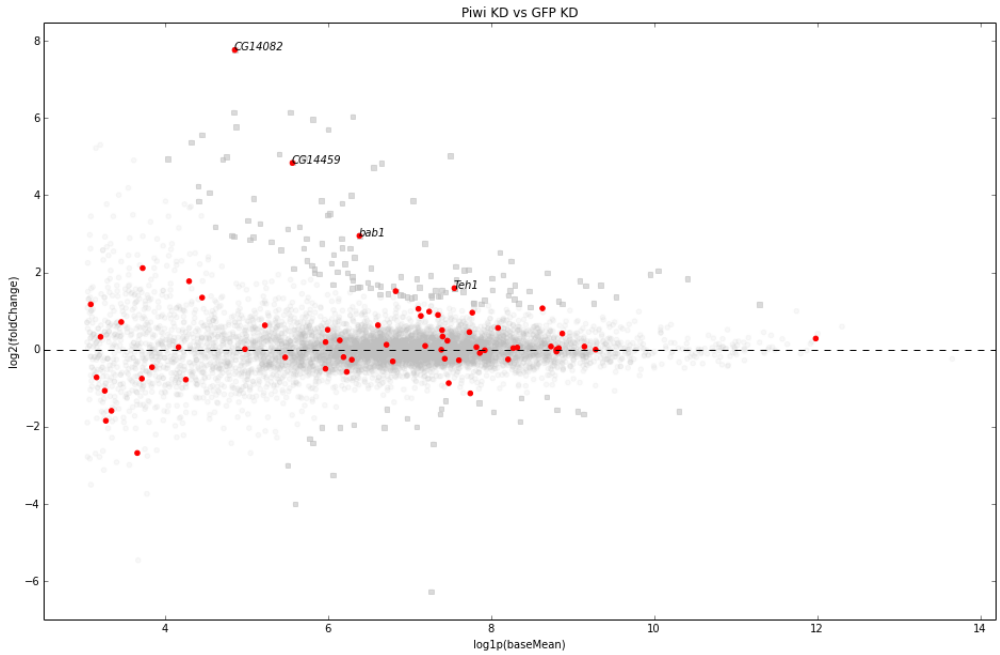
In [8]:
```
ax=fig.add_subplot(111)
```

In [9]:
```
fig=d.scatter(
    ax=ax,
    x='baseMean',
    y='foldChange',
    xfunc=log1p, yfunc=log2,
    general_kwargs=dict(color='0.75', alpha=0.1, ),
    #-----------------------------------------------------
    genes_to_highlight=genes_to_highlight,
    label_kwargs=dict(style='italic', fontsize=10));
fig.set_title(label='Piwi KD vs GFP KD');
fig.axhline(0, color='k', linestyle='--')
```

Out[9]:
```
<matplotlib.lines.Line2D at 0x563e990>
```

In [10]:
```
fig.get_figure()
```

Out[10]:



In [11]:
```
upregulated=d[d.data['foldChange'] > 4]
len(upregulated.data)
```

Out[11]:
```
119
```

In [12]:
```
upregulated_with_insertion_table=d[up_with_insertion]
upregulated_with_insertion_table.data[['baseMean', 'foldChange', 'padj']]
```

Out[12]:

	baseMean	foldChange	padj
id			
FBgn0037171	259.730780	28.611308	7.153039e-08
FBgn0036851	127.560828	217.143743	1.341933e-06
FBgn0004870	591.982064	7.727221	2.500602e-06
FBgn0037766	40.521548	4.330502	1.000000e+00

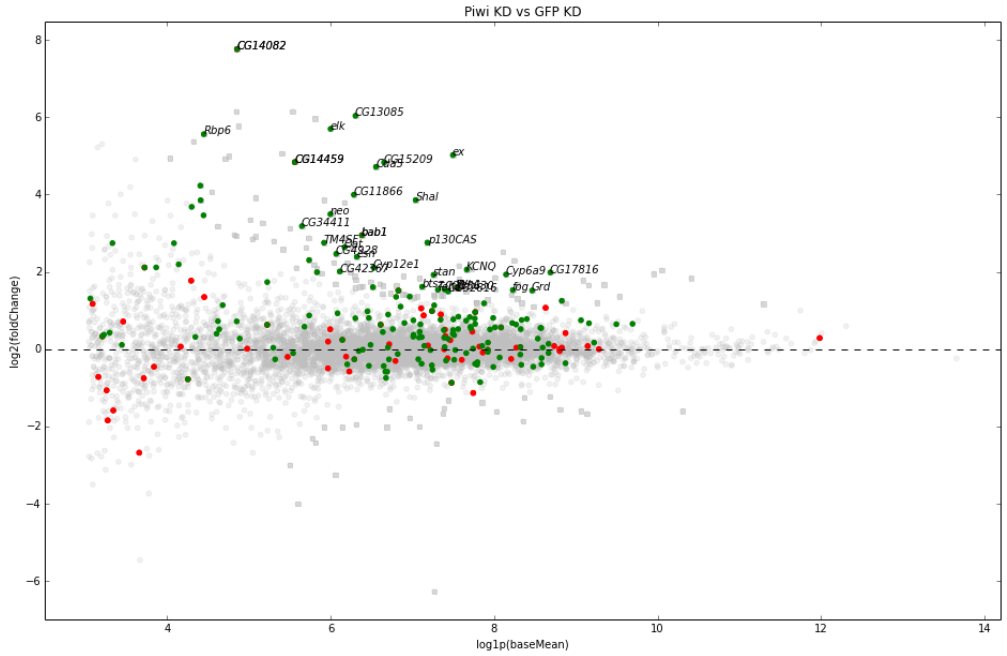
4 rows × 3 columns

OK, of the 119 genes that are more than 4fold upregulated in PIWI knockdown only 4 contain a group 1 TE. Let's repeat the analysis, but use the mapped insertion sites for group1 transposons that brennecke has generated by DNA resequencing for his paper (genes containing group1 TEs are green now).

In [13]:
```
plt.close('all')
up=(d.data['foldChange'] > 4)*1

with_insertion=d.genes_with_peak(TEs_brennecke)*1
up_with_insertion=up+with_insertion == 2
up_names=d[up_with_insertion].data.index.tolist()
up_names=convert_ids(up_names)
genes_to_highlight=[(d.genes_with_peak(TEs_flybase), dict(color='r', name
s=[], alpha=1)),
                    (d.genes_with_peak(TEs_brennecke), dict(color='g', na
mes=up_names, alpha=1)),
         ];
fig=d.scatter(
    ax=ax,
    x='baseMean',
    y='foldChange',
    xfunc=log1p, yfunc=log2,
    general_kwargs=dict(color='0.75', alpha=0.1, ),
    genes_to_highlight=genes_to_highlight,
    label_kwargs=dict(style='italic', fontsize=10));
fig.get_figure()
```

Out[13]:



So indeed, it seems to be very important to know exactly where the transposon insertions are located, because it can change the interpretation of the results a lot. In particular, without having the transposon data available, I would have said that transposon insertions in genes are not important to determine whether a genes is repressed by a piwi-based mechanism (3% of overexpressed genes contain group 1 transposons). However, when we intersect the overexpressed genes with TE insertions, a nice percentage (27%) does have TE insertions, meaning that genes with TE insertion are likely to be repressed by a piwi-based mechanism. In conclusion: Yes it is important to know the exact integration site, as it can change the interpretation.

dme-mir-973	2.03E+00	8.04E-01	1.09E+01	9.60E-04	7.75E-03	3.15E-01	6.90E+01

http://lbcd41.snv.jussieu.fr/galaxy/u/marius-ged/h/compare-simulation-with-real-differences

Differential expression analysis between ping-pong negative libraries
and ping-pong negative libraries supplemented with 2% testes reads
The 19 differentially detcted miRNA between ping-pong negative libraries
and ping-pong positive libraries (Sheet 1) are highlighted in red

Name	logFC	logCPM	LR	PValue	adj.p.value	Dispersion	totreads
dme-mir-964	6.30E+00	5.74E+00	1.18E+03	3.54E-258	7.29E-256	4.31E-04	9.87E+02
dme-mir-959	5.61E+00	5.57E+00	9.99E+02	3.01E-219	3.10E-217	3.89E-04	8.76E+02
dme-mir-962	8.52E+00	5.28E+00	9.31E+02	1.79E-204	1.23E-202	2.54E-04	7.11E+02
dme-mir-984	5.42E+00	5.46E+00	9.18E+02	1.08E-201	5.54E-200	3.54E-04	8.15E+02
dme-mir-985	4.96E+00	5.10E+00	6.78E+02	1.44E-149	5.94E-148	2.41E-04	6.28E+02
dme-mir-974	5.43E+00	4.66E+00	5.17E+02	1.63E-114	5.60E-113	1.42E-04	4.52E+02
dme-mir-991	4.29E+00	4.83E+00	4.97E+02	4.35E-110	1.28E-108	1.32E-04	5.10E+02
dme-mir-960	3.11E+00	4.85E+00	3.69E+02	2.75E-82	7.09E-81	1.73E-04	5.24E+02
dme-mir-961	4.48E+00	4.16E+00	3.17E+02	6.53E-71	1.49E-69	6.58E-05	3.11E+02
dme-mir-312	3.19E+00	4.22E+00	2.37E+02	1.41E-53	2.90E-52	1.02E-04	3.24E+02
dme-mir-997	5.58E+00	3.55E+00	2.31E+02	4.06E-52	7.61E-51	3.34E-07	1.95E+02
dme-mir-976	7.32E+00	3.15E+00	1.90E+02	2.95E-43	5.06E-42	7.77E-07	1.41E+02
dme-mir-977	2.84E+00	3.81E+00	1.54E+02	2.45E-35	3.88E-34	7.28E-06	2.41E+02
dme-mir-31b	1.35E+00	5.36E+00	1.46E+02	1.37E-33	2.02E-32	3.28E-04	7.61E+02
dme-mir-978	6.69E+00	2.67E+00	1.23E+02	1.26E-28	1.73E-27	1.12E-07	9.20E+01
dme-mir-989	2.63E+00	3.57E+00	1.15E+02	9.46E-27	1.22E-25	3.54E-07	1.98E+02
dme-mir-963	3.46E+00	2.97E+00	9.96E+01	1.85E-23	2.24E-22	5.76E-09	1.21E+02
dme-mir-983-2	2.46E+00	2.87E+00	5.94E+01	1.26E-14	1.45E-13	2.68E-08	1.11E+02
dme-mir-992	5.59E+00	1.89E+00	5.64E+01	5.79E-14	6.28E-13	6.19E-04	4.20E+01
dme-mir-313	2.68E+00	2.52E+00	4.96E+01	1.92E-12	1.92E-11	1.40E-04	8.10E+01
dme-mir-982	1.39E+00	3.81E+00	4.95E+01	1.96E-12	1.92E-11	2.40E-05	2.37E+02
dme-mir-983-1	2.06E+00	2.85E+00	4.48E+01	2.14E-11	2.00E-10	3.17E-08	1.09E+02
dme-mir-303	5.06E+00	1.59E+00	3.88E+01	4.65E-10	4.16E-09	1.54E-06	2.90E+01
dme-mir-318	1.18E+00	3.60E+00	3.15E+01	1.98E-08	1.70E-07	3.94E-07	2.03E+02
dme-mir-4966	3.91E+00	1.59E+00	3.04E+01	3.43E-08	2.83E-07	9.22E-04	2.90E+01
dme-mir-310	3.80E+00	1.54E+00	2.79E+01	1.29E-07	1.03E-06	1.65E-06	2.70E+01
dme-mir-311	1.19E+00	3.23E+00	2.40E+01	9.81E-07	7.49E-06	7.38E-08	1.51E+02
dme-mir-973	2.29E+00	1.96E+00	2.32E+01	1.44E-06	1.06E-05	5.31E-04	4.60E+01
dme-mir-979	4.25E+00	1.22E+00	2.17E+01	3.23E-06	2.29E-05	2.42E-06	1.60E+01
dme-mir-375	2.10E-01	8.56E+00	1.82E+01	2.04E-05	1.40E-04	2.05E-03	7.13E+03
dme-mir-316	3.18E-01	6.43E+00	1.81E+01	2.12E-05	1.41E-04	6.00E-04	1.61E+03
dme-mir-79	1.89E-01	8.04E+00	1.67E+01	4.42E-05	2.84E-04	7.49E-04	5.02E+03
dme-mir-2498	3.32E+00	9.24E-01	1.08E+01	1.03E-03	6.42E-03	3.09E-04	8.00E+00
dme-mir-999	-9.88E-02	1.46E+01	1.07E+01	1.07E-03	6.51E-03	1.50E-03	4.71E+05
dme-mir-278	-9.33E-02	1.26E+01	9.90E+00	1.65E-03	9.72E-03	1.36E-03	1.21E+05