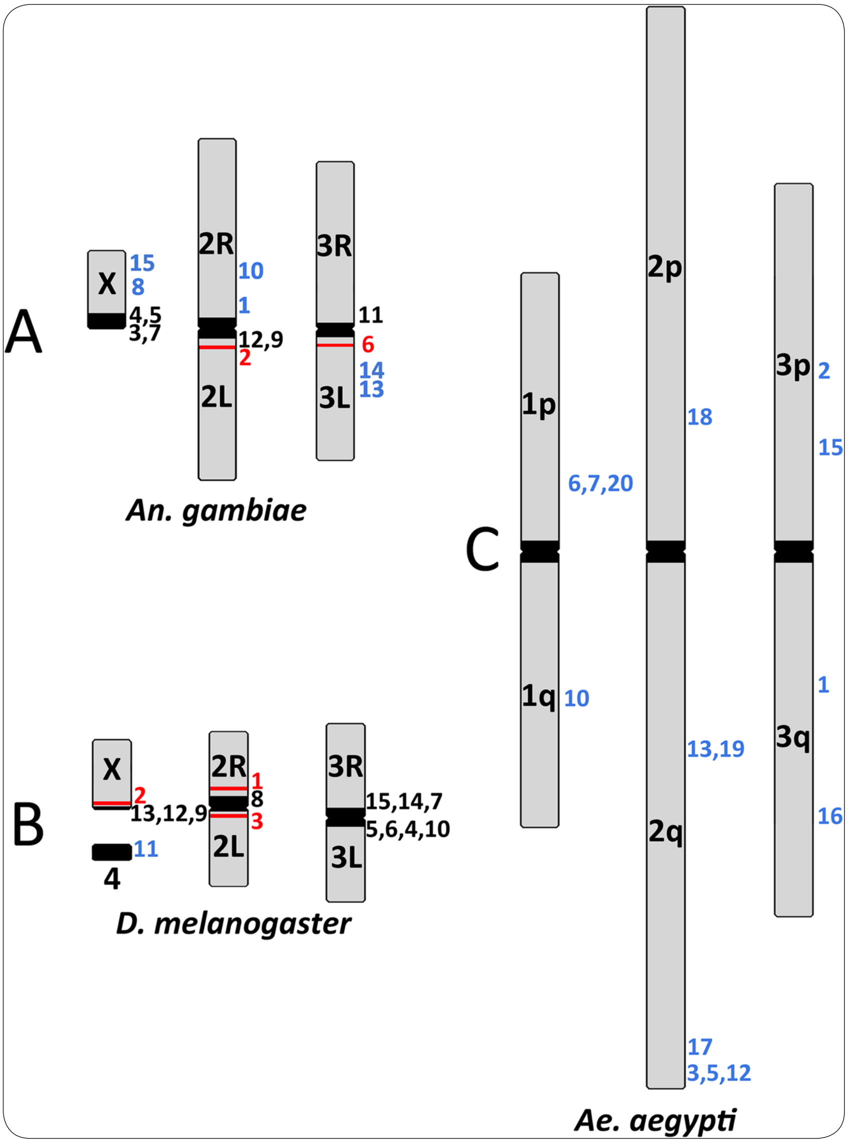
Epigenetics & Chromatin



Increased production of piRNAs from euchromatic clusters and genes in *Anopheles gambiae* compared with *Drosophila melanogaster*

George et al.

BioMed Central

George *et al. Epigenetics & Chromatin (2015) 8:50*
DOI 10.1186/s13072-015-0041-5

Epigenetics & Chromatin

RESEARCH

Open Access

Increased production of piRNAs from euchromatic clusters and genes in *Anopheles gambiae* compared with *Drosophila melanogaster*

Phillip George[1†], Silke Jensen[2,3,4†], Romain Pogorelcnik[2,3,4], Jiyoung Lee[5], Yi Xing[5], Emilie Brasset[2,3,4], Chantal Vaury[2,3,4*] and Igor V. Sharakhov[1,5*]

Abstract

Background: Specific genomic loci, termed Piwi-interacting RNA (piRNA) clusters, manufacture piRNAs that serve as guides for the inactivation of complementary transposable elements (TEs). The piRNA pathway has been accurately detailed in *Drosophila melanogaster*, while it remains poorly examined in other insects. This pathway is increasingly recognized as critical for germline development and reproduction. Understanding of the piRNA functions in mosquitoes could offer an opportunity for disease vector control by the reduction of their reproductive potential.

Results: To analyze the similarities and differences in this pathway between *Drosophila* and mosquito, we performed an in-depth analysis of the genomic loci producing piRNAs and their targets in the African malaria vector *Anopheles gambiae*. We identified 187 piRNA clusters in the *An. gambiae* genome and 155 piRNA clusters in the *D. melanogaster* genome. We demonstrate that many more piRNA clusters in the mosquito compared with the fruit fly are uni-directionally transcribed and are located outside pericentromeric heterochromatin. About 11 % of the *An. gambiae* piRNA population map to gene transcripts. This is a noticeable increase compared with the ~6 % of the piRNA population mapped to genes in *D. melanogaster*. A subset of the piRNA-enriched genes in *An. gambiae* has functions related to reproduction and development. At least 24 and 65 % of the mapped piRNAs correspond to genomic TE sequences in *An. gambiae* and *D. melanogaster*, respectively. DNA transposons and non-LTR retrotransposons are more abundant in *An. gambiae*, while LTR retrotransposons are more abundant in *D. melanogaster*. Yet, piRNAs predominantly target LTR retrotransposons in both species, which may point to a distinct feature of these elements compared to the other classes of TEs concerning their silencing by the piRNA pathway.

Conclusions: Here, we demonstrate that piRNA-producing loci have more ubiquitous distribution in the *An. gambiae* genome than in the genome of *D. melanogaster*. Also, protein-coding genes have an increased role in production of piRNAs in the germline of this mosquito. Genes involved in germline and embryonic development of *An. gambiae* generate a substantial portion of piRNAs, suggesting a role of the piRNA pathway in the epigenetic regulation of the reproductive processes in the African malaria vector.

Keywords: *Anopheles gambiae*, Development, *Drosophila melanogaster*, Euchromatin, Germline, Heterochromatin, piRNA clusters, Reproduction, Small RNAs, Transposable element

*Correspondence: chantal.vaury@udamail.fr; igor@vt.edu
†Phillip George and Silke Jensen contributed equally to this study
[1] Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA
[2] Laboratoire Génétique, Reproduction, et Développement, Clermont Université, Université d'Auvergne, BP 38, 63001 Clermont-Ferrand, France
Full list of author information is available at the end of the article

© 2015 George et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

Background

Piwi-interacting RNAs (piRNAs) are 24–30 nucleotide (nt) small RNAs that play an important role in silencing active transposable elements (TEs) through slicer-mediated cleavage of messenger RNA (mRNA) [1]. piRNAs are by far the most numerous among all types of coding and non-coding RNAs in any animal, mostly acting in the germline. piRNAs of different species share similar features, including a typical motif of predominant Uridine at position one (1U) of antisense TE-derived piRNAs and Adenine at position ten (10A) of sense strand TE-derived piRNAs [2]. Members of the PIWI clade, a subfamily of Argonaute, interact with piRNAs to effectively create an RNA-induced silencing complex (RISC) that can target and silence complementary TE mRNA sequences. A mutation of any of the three key PIWI proteins—Piwi, Aubergine (Aub), and Argonaute 3 (Ago3)—results in de-repression of TEs with mutagenic or disruptive consequences in the *Drosophila melanogaster* germline [3–5], indicating the necessity of these proteins in functional TE silencing. Two mechanisms for piRNA production have been identified in *D. melanogaster* [2]; both mechanisms stem from long single-stranded piRNA precursors that originate from vestigial TEs. In the first mechanism, single-stranded RNA transcripts are processed into primary piRNAs, which are loaded onto the Piwi protein. This process has been referred to as primary piRNA biogenesis [2]. Trimming of the piRNA to the 24–30 nt characteristic size of these small RNAs requires the cytoplasmic endonuclease Zucchini [6, 7]. In another mechanism, secondary piRNAs, responsible for a large portion of the total piRNA pool in the germline, are generated through an amplification loop referred to as the ping-pong cycle [2] and loaded onto Ago3. The Aub protein is posited to work within the ping-pong cycle by binding tertiary piRNAs that are generated through the amplification loop. A ten base-pair overlap can be seen between complementary primary and secondary piRNAs [2, 8]. Many piRNAs associating with the Aubergine and Piwi proteins are antisense to TEs and show a typical 1U feature, while piRNAs associated with Argonaute 3 are sense to the TE transcripts and show a 10A feature.

The piRNA pathway is a major epigenetic programming mechanism in higher eukaryotes and it has been increasingly implicated in germline development of eukaryotes. The Piwi protein is essential to fertility in *D. melanogaster* [9, 10], *Caenorhabditis elegans* [11, 12], *Danio rerio* [13], *and Mus musculus* [14]. Germline stem cell loss has also been documented in multiple organisms as a result of piRNA pathway mutation [9, 10, 14–17]. The PIWI proteins of the Asian malaria vector *Anopheles stephensi* are expressed at high levels in the germline cells of ovaries as expected and, importantly, their expression is further increased after a blood meal [18]. In addition to TE-derived piRNAs, a fraction of piRNAs map in the sense orientation to the 3′ untranslated regions (UTRs) of protein-coding transcripts [19–23]. piRNAs are produced by various protein-coding genes, including *stellate, suppressor of stellate* [24], and *traffic jam* [19, 25], that are important for the germline development in *D. melanogaster*. piRNAs from *suppressor of stellate* functionally silence *stellate* transcripts, and a deletion of *suppressor of stellate* leads to *stellate* overexpression and meiotic abnormalities in *Drosophila* testis [24]. *Wolbachia* can control the maternal transmission of endogenous gypsy retroviruses in *D. melanogaster* [26]. Knowledge of the mechanisms of how the piRNA pathway regulates reproduction in mosquitoes could be useful for both basic and applied science. Our increased understanding of reproductive processes in disease vectors will facilitate the identification of novel targets for vector control [27]. The piRNA pathway has also been linked to other epidemiologically important phenotypes in mosquitoes. For example, a role of piRNAs in antiviral immune responses in both *Aedes aegypti* and *Ae. albopictus* has been demonstrated [28, 29]. A recent study has shown that *Wolbachia* can manipulate the mosquito cell RNAi/miRNA/piRNA machinery by inducing or suppressing specific small RNAs [30].

The majority of piRNAs originate from clusters, genomic regions ranging in size from approximately 1–250 kb [2, 22]. Produced piRNAs serve as guides for targeted inactivation of complementary TEs. piRNA clusters do not have an explicit strand bias; however, in some cases, they do exhibit high percentages of TEs in one orientation or the other [2, 31, 32]. In *D. melanogaster*, the piRNA clusters are almost exclusively located in heterochromatin—the pericentromeric and subtelomeric regions—regions with an abundance of TEs [2]. It is not clear if the predominant location of piRNA clusters in heterochromatin is specific to the fruit fly or is typical to Diptera. The African malaria mosquito, *Anopheles gambiae,* represents an intermediate in terms of the genome assembly size (273.1 Mb) [33] compared to other phylogenetically distant Dipterans with studied piRNA pathways *D. melanogaster* (143.9 Mb) [34] and *Ae. aegypti* (1311 Mb) [35] (Fig. 1). Moreover the genomic distribution of TEs differs among the three species. Over 77 % of pericentromeric heterochromatin and only 7 % of euchromatin are occupied by TEs in *D. melanogaster* [36]. This difference is less dramatic in *An. gambiae*: 33.1 % of pericentromeric heterochromatin and 14.5 % of the rest of the genome are covered by TEs [37]. The large regions of intercalary heterochromatin present in *An. gambiae* are mainly responsible for the high peaks of TE coverage outside the pericentromeric regions. In

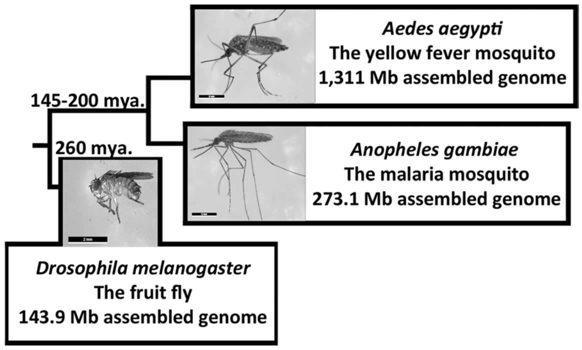

Aedes aegypti
The yellow fever mosquito
1,311 Mb assembled genome
145-200 mya.
260 mya.
Anopheles gambiae
The malaria mosquito
273.1 Mb assembled genome
Drosophila melanogaster
The fruit fly
143.9 Mb assembled genome

Fig. 1 Phylogenetic delineation of three Dipteran species: *D. melanogaster*, *Ae. aegypti*, and *An. gambiae*

an extreme case, *Ae. aegypti* has a homogeneously high coverage of TEs (~52 %) across pericentromeric heterochromatin and other chromosomal regions [38]. Do mosquito species with a more ubiquitous distribution of TEs reflectively have a redistribution of piRNA clusters from heterochromatic to euchromatic regions? Does the piRNA pathway have conserved functions in organisms with different genome sizes, chromatin landscapes, and predominant TE families? These questions can be addressed by mapping piRNAs to annotated features of chromosome-based genome assemblies.

Here, we identified piRNAs from the ovarian tissue of *An. gambiae* females and characterized potential relationships between these small RNAs and the mosquito's genomic features. We compared chromosome location of piRNA clusters in three Dipteran species with remarkably different patterns of genomic organization: *D. melanogaster*, *Ae. aegypti*, and *An. gambiae* [39]. We discovered a noticeable shift in piRNA cluster location in *An. gambiae* compared with *D. melanogaster*; the 15 most productive piRNA clusters are less confined to the pericentromeric heterochromatin, but can also be found in intercalary heterochromatin and euchromatin. The *Ae. aegypti* top 15 piRNA clusters [22] are even more pervasive, occupying euchromatic genomic scaffolds that have been previously placed to chromosomes by physical and genetic mapping [38, 40]. There is an increase in gene-derived piRNAs in the malaria mosquito when compared with the fruit fly. Among the genes identified in *An. gambiae* as rich in piRNA mapping, a subset includes genes potentially critical to reproduction and germline development, including *oskar* (AGAP003545).

Results

The piRNA population in the ovarian tissue of *An. gambiae*

To accurately characterize the sequence and genomic location of piRNAs produced by the African malaria vector, we isolated and sequenced small RNAs from blood-fed ovaries of the Mali strain (the M form) of *An. gambiae* using the Illumina Small RNA TruSeq technology. The generated library of small RNAs, which we have named the non-collapsed non-unique (NCNU) library—reads including any duplicates (referred to as non-collapsed or NC) that map to one or more locations in the genome (referred to as non-unique or NU)—showed a bi-modal length distribution with two peaks (Additional file 1: Figure S1). The definition and sizes of each library are shown in Additional file 2: Table S1. The first peak occurred at a size of 22 nt, which we attributed to microRNAs (miRNA). The second broad peak spanning 24–29 nt with an apex at 27 nt represented the potential piRNA pool. The 27 nt peak has also been reported for the G3 strain (the S form) of *An. gambiae* [21, 23]. We noted that it differs from the 26 nt peak of *D. melanogaster* and 28 nt peak of *Ae. aegypti* (Fig. 2a, c, e). However, the total small RNA size range is in close concordance with *Ae. aegypti* (24–31 nt) [22], *D. melanogaster* (23–29 nt) [2], *Bombyx mori* (26–31 nt) [41], and *Danio rerio* (24–30 nt) [13].

A majority of piRNAs target TEs (predominantly LTR retrotransposons) in both *An. gambiae* and *D. melanogaster*

To accurately characterize TE-derived piRNAs, 24–29 nt sequences from the NCNU library were mapped to the 7080 annotated Hexapoda TEs from RepBase [42]. The *An. gambiae* TE library is not nearly as extensive as the *D. melanogaster* TE library; therefore, we used all identified TE sequences from the subphylum Hexapoda to bolster the reference library used for piRNA mapping and to help mitigate potential non-annotated ancestral TEs. It is generally believed that the percentage of reported TE-matching piRNAs is an underestimation of the real percentage because of incomplete annotation of TEs in insect genomes [43]. Two approaches were taken to identify piRNAs derived from TEs: "consensus" mapping and "overlap" mapping. The first approach used traditional read mapping with Bowtie2, aligning the piRNAs to consensus TE sequences with up to three mismatches. This method relies on the assumption that three mismatches would be sufficient to identify TEs degraded from their original consensus sequences. Using this methodology, approximately 23.6 % of the total piRNA library was identified as being derived from TE sequences (Fig. 2b). We applied the same method to other Dipteran species with characterized piRNAs [22, 44] and found that this proportion of TE-mapped piRNAs is similar to that of *Ae. aegypti* (26.3 %) (Fig. 2d) but is much smaller than the 65.4 % identified in *D. melanogaster* (Fig. 2f). Some of this difference may be attributed to the larger proportion

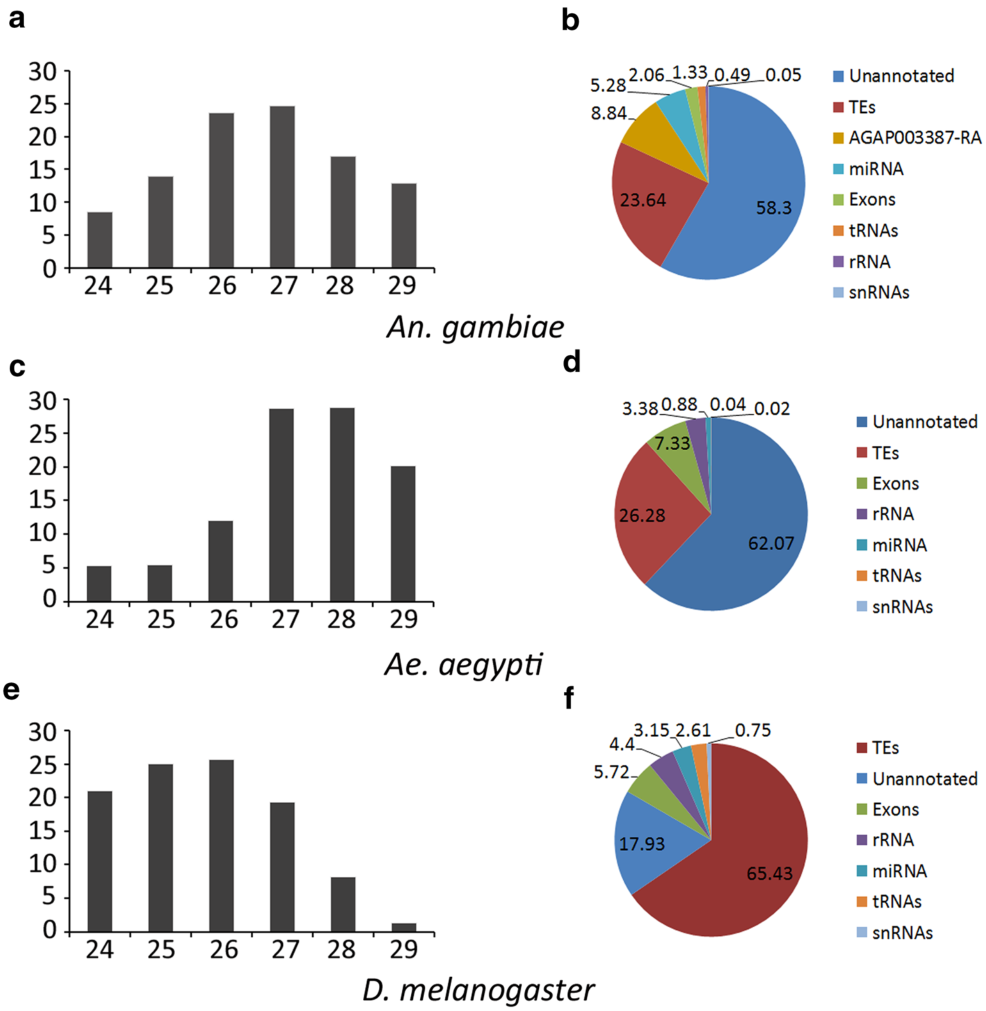


Fig. 2 The size distribution and annotation of small RNAs within the piRNA size range (24–29 nt) of the NCNU libraries in Dipteran species *An. gambiae, Ae. aegypti,* and *D. melanogaster*. **a**, **c**, **e** represent the distribution of small RNA sequences found in the sequenced total library for the respective species (*X* axis: size in nucleotides (nt), *Y* axis: percentage of all 24–29 nt mapping small RNAs). **b**, **d**, **f** depict the annotation of these small RNAs for the three species

of unannotated sequences in *An. gambiae*, where at least 58.3 % of piRNAs are mapped to the unannotated part of the genome. This number is similar to that in *Ae. aegypti,* but is much larger than the 17.9 % identified in *D. melanogaster.*

The second approach of identifying TE-derived piRNAs sought to overcome limitations caused by a possible incomplete annotation of TE libraries and by severe sequence divergence from consensus TE sequences. Many of the TEs residing within a genome, especially within the heterochromatin, have been mutated to the point that they are quite different from their initial consensus sequences. By using up to three mismatches when mapping to consensus TE sequences, a substantial portion of low-identity TE sequences may be missed. We designed an "overlap" method that incorporated repeat masking data generated using the RepeatMasker platform [45] to identify genomic positions of the same 7080 annotated Hexapoda TEs sequences from RepBase [42]. Genomic piRNA positions were concurrently identified by short-read mapping the small RNAs to the organism's reference genome assembly. The two position lists were overlapped to identify common genomic sequences that we considered TE-derived piRNA originators (see "Methods"). This second method resulted in 39.4 % of the piRNA pool that may be derived from TE sequences in the mapped portion of the *An. gambiae* genome (Additional file 3: Table S2). We tested the "overlap" approach for identifying TE-derived piRNAs on *D. melanogaster.* To do this, a w^{1118} ovarian piRNA library [44] was mapped to the *D. melanogaster* genome (version R6.04) with no mismatches. We used the same Hexapoda TEs from RepBase [42] to be consistent with our *An. gambiae* analysis. We found 81.6 % (2,149,899 of 2,634,680 RNAs) of the total sequenced piRNAs mapped to TEs in *D. melanogaster,* which is still much higher than the 39.4 % (8,890,061 of 22,569,568 RNAs) of TE-mapped piRNA identified in *An. gambiae* (Additional file 3: Table S2).

Although the two approaches for identifying TE-derived piRNAs result in considerably different predictions in respect to the percentage of piRNAs derived from TEs in *An. gambiae*, they still agree on the families of TEs with the highest abundance of mapped piRNAs (Fig. 3a–d). TE classes with the highest piRNA representation are LTR retrotransposons (59 % of piRNAs) followed by non-LTR retrotransposons (28 %) and DNA transposons (12 %), as determined by the "overlap" method. However, classes and families most abundant in piRNA mapping do not correspond to the TE classes and families most prevalent in the *An. gambiae* genome (Fig. 3e, f). For example, LTR retrotransposons represent the most abundant class of TEs enriched in piRNAs (Fig. 3b, d). Yet, non-LTR retrotransposons are more abundant in the *An. gambiae* genome than are LTR retrotransposons (34 vs. 29 %, respectively, Fig. 3f). Moreover, 37 % of all TE sequences in the *An. gambiae* genome are represented by DNA transposons, while DNA transposons generate only 12 % of piRNAs (Fig. 3d).

TE mapping data provide evidence of an active ping-pong amplification cycle in *An. gambiae* ovaries

In the ping-pong amplification cycle, antisense piRNAs are considered to be almost exclusively derived from piRNA clusters and are proposed to be primarily responsible for TE silencing through the piRNA pathway [2]. In the *An. gambiae* dataset, we identify a strong bias toward antisense piRNAs. About 69 % of the piRNAs identified as TE-derived were found in the opposite orientation with regard to TEs identified by the "overlap" method. This finding holds consistent across each of the arms and

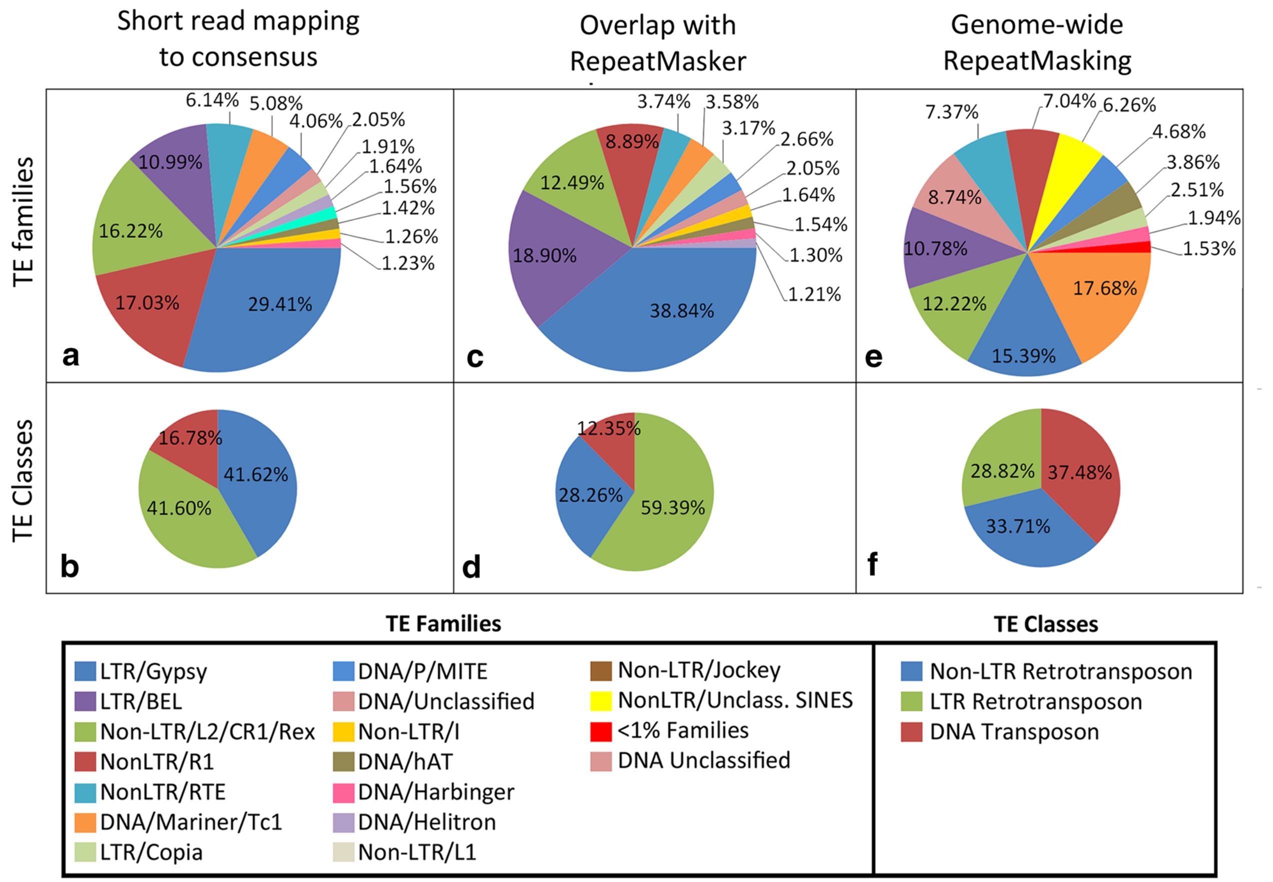


Fig. 3 TE-derived piRNAs in *An. gambiae* identified by two different approaches. The percentage of TE-derived piRNAs mapping to the consensus sequences of annotated TEs with up to three mismatches in TE families (**a**) and TE classes (**b**). The percentage of TE-derived piRNAs with mapping positions that overlap RepeatMasker-identified positions in TE families (**c**) and TE classes (**d**). Genome-wide TE-content, comparison of TE families (**e**), and classes (**f**) by RepeatMasker. The Hexapoda library from RepBase [42] was used as reference

to an extent within the unknown (UNKN) chromosome, which consists of unmapped genomic scaffolds (Additional file 4: Figure S2).

In addition, the ping-pong amplification loop promotes the generation of both sense and antisense RNAs that help to guide piRNA complexes toward transposon mRNAs for their subsequent processing. The RNAs generated from this amplification loop typically exhibit two motifs. Both motifs, a precise 10 nucleotide overlap between overlapping piRNAs and positional biases with Uridine at position 1 and Adenine at position 10 [2, 8, 22], stem from the way in which the ping-pong amplification generates piRNAs. In *D. melanogaster*, Piwi and Aubergine proteins have a noted preference toward piRNAs that are antisense to mRNAs encoded by TEs with a Uridine at the 5′ end [2]. As the PIWI-mediated cleavage is now known to occur between positions 10 and 11 of the RNA complementary to the guide strand, the resulting secondary, sense piRNA is expected to contain an Adenine at the tenth nucleotide. Ago3 proteins have a preference toward piRNAs sense to TE mRNAs. This targeted cleavage results in an amplified piRNA pool capable of targeting both sense and antisense transposon transcripts.

From a total of 22,569,568 small RNAs ranging from 24 to 29 nt and constituting the potential piRNA pool in *An. gambiae*, we further examined this population for piRNA signatures and for sequences from which these piRNAs might be derived. We identified 79.1 % of the 24–29 nt bona fide reads (excluding miRNA, rRNA, tRNA, snRNA) as having a U at position 1 and 28.4 % as having an A at position 10 (Fig. 4a). We also analyzed TE-mapping piRNAs separately, allowing 0–3 mismatches. Using the NCNU library, the sense reads for TEs show a 60.9 % bias of 1U and a 54.0 % bias of 10A, while antisense reads show a 84.3 % bias of 1U and a 27.8 % bias of 10A. The values obtained from the collapsed NU library are very similar (Fig. 4a), indicating that 1U and 10A

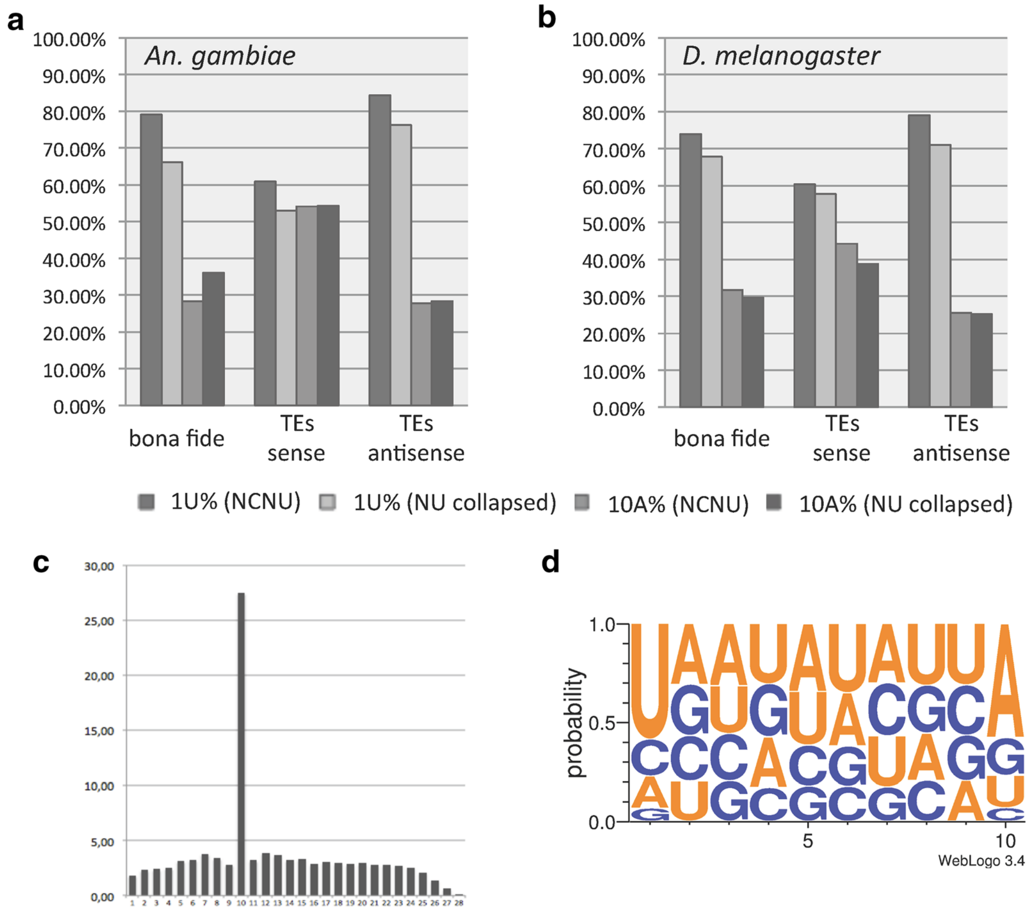


Fig. 4 Characteristics of piRNA sequences (24–29 nt). **a** Percentages of 1U and 10A signatures in bona fide and in TE piRNAs in *An. gambiae*. **b** Percentages of 1U and 10A signatures in bona fide and in TE piRNAs in *D. melanogaster*. **c** 5′ ends of complementary piRNAs of *An. gambiae* frequently exhibited a 10-bp overlap (the ping-pong signature) in all piRNA genome-mappers. **d** Nucleotide abundance in the 10-bp overlap section of ping-pong piRNA partners in *An. gambiae*

biases are not due to multiplicity of many piRNA reads. These percentages are very close to the corresponding percentages in *D. melanogaster* (Fig. 4b). We detected a strong 10 nt overlap bias (Fig. 4c) with 1U and 10A signatures (Fig. 4d) in the 24–29 nt subgroup from *An. gambiae*. In accordance with previous studies [21, 23], the presence of these signatures suggests that a mechanism similar to the established ping-pong amplification loop in *D. melanogaster* exists in the ovarian tissue of *An. gambiae*, functioning with Ago3 and Aub orthologs, which load preferentially sense and antisense piRNAs, respectively.

187 piRNA clusters are identified in the genome of *An. gambiae*

The *An. gambiae* PEST genome is chromosomally assembled [33, 46], allowing for the spatial identification of chromosomal positions associated with piRNA enrichment. The reference PEST strain has haplotypes of both S and M forms of *An. gambiae* segregating in different regions of the genome [33]. Overall, the M form genome is slightly more similar to the PEST genome than is the S form genome [47], which gave us an advantage in mapping the Mali strain (the M form) reads to the reference genome. piRNA clusters in *An. gambiae* were identified using a strategy that incorporated the methods used previously in both *D. melanogaster* [2] and *Ae. aegypti* [22] (see "Methods"). A total of 120 piRNA clusters in *An. gambiae* were found on the assembled chromosomes, with 18 clusters being identified on chromosome X, 35 clusters on 2R, 27 on 2L, 18 on 3R, and 22 on 3L (Additional file 5: Table S3). Chromosome UNKN, being made up of variously sized scaffolds sorted in decreasing order by length, provides approximately 40 % of the mapped piRNA sequences. The abundance of small, artificially concatenated scaffolds in the UNKN chromosome made it difficult to determine the boundaries of piRNA clusters when working with 5-kb windows, so we limited our analyses to the first 10 Mb. Scaffolds at the 10-Mb region of chromosome UNKN were at least ~50 kb in length. We identified 67 extra clusters present in the first 10 Mb of chromosome UNKN. The total 187 clusters span in size from 10 to 1.29 Mb, which is a departure from the smaller clusters (2–242 kb, 6–184 kb, respectively) identified in previous studies in *D. melanogaster* [2] and *Ae. aegypti* [22]. We ran our piRNA cluster analysis pipeline using the w^{1118} piRNA library [44] on the Dmel_R6.04 release of the *D. melanogaster* [34] genome assembly using the same methodology as for *An. gambiae* in an effort to compare genomic location of piRNA clusters in the different species (see "Methods"). We identified 155 clusters in *D. melanogaster* ranging in size from 10 to 1.13 Mb (Additional file 5: Table S3).

A large proportion of *An. gambiae* piRNA clusters is uni-directionally transcribed

A majority of ovarian piRNA clusters are bi-directionally transcribed in *D. melanogaster* [2]. Unidirectional clusters, like *flamenco*, are primarily expressed in the somatic follicular cells, while bidirectional clusters are transcribed in ovarian nurse cells [48]. In *An. gambiae*, 66 of the 120 clusters (55 %) belonging to assembled chromosomes had more than 75 % of piRNAs mapping to a single strand (strong bias) (Fig. 5a). Still, 17.5 % of mapped piRNA clusters had more than 90 % of piRNAs mapping to only a plus or minus strand (near exclusive bias) (Fig. 5b). In contrast, only 18 % of piRNA clusters had a strong plus or minus strand bias in *D. melanogaster* (Fig. 5c). Furthermore, only 4.5 % of piRNA clusters had a near exclusive bias of piRNAs mapping to a single strand in *D. melanogaster* (Fig. 5d). If we consider all 187 piRNA clusters in *An. gambiae*, 38 % had a strong strand bias and 11.2 % had a near exclusive strand bias. This result indicates that the UNKN chromosome clusters, which are likely heterochromatic, are mostly bidirectional. The very top piRNA cluster, which is located in the *An. gambiae* euchromatin, is unidirectional (Additional file 6: Figure S3A). The next two top piRNA clusters are bidirectional and are located in intercalary and pericentromeric heterochromatin of *An. gambiae* (Additional file 6: Figure S3B, C).

Genomic location of piRNA clusters vary among three dipteran species

Of the total 120 chromosomally mapped piRNA clusters in *An. gambiae*, the longest piRNA clusters localized to heterochromatic regions [37]. Twenty-six of the clusters (21.7 %) were located in pericentromeric heterochromatin; two were found in the intercalary heterochromatin of arms 2L and 3L, and 92 piRNA clusters occupied euchromatic regions of all chromosomes (Additional file 5: Table S3). Of the top 15 piRNA clusters (ranked by the number of unique piRNAs), seven (46.7 %) were located in pericentromeric heterochromatin of chromosomes X, 2L, and 3R, two clusters were located in intercalary heterochromatin of arms 2L and 3L, and six piRNA clusters occupied euchromatic regions of chromosomes X, 2R, and 3L (Fig. 6a).

Of the total 155 piRNA clusters identified in *D. melanogaster*, 76 (49 %) were located in pericentromeric heterochromatin. In addition, 26 clusters were found in intercalary heterochromatin [49], and 2 clusters were mapped to Y-chromosome heterochromatin. The remaining 51 piRNA clusters occupied euchromatic regions. All but three of the top 15 clusters identified by Brennecke et al. [2] correspond to the clusters with the highest number of unique piRNAs identified using our methodology. Unlike the previous study [2], we classified the three highest piRNA-producing clusters in

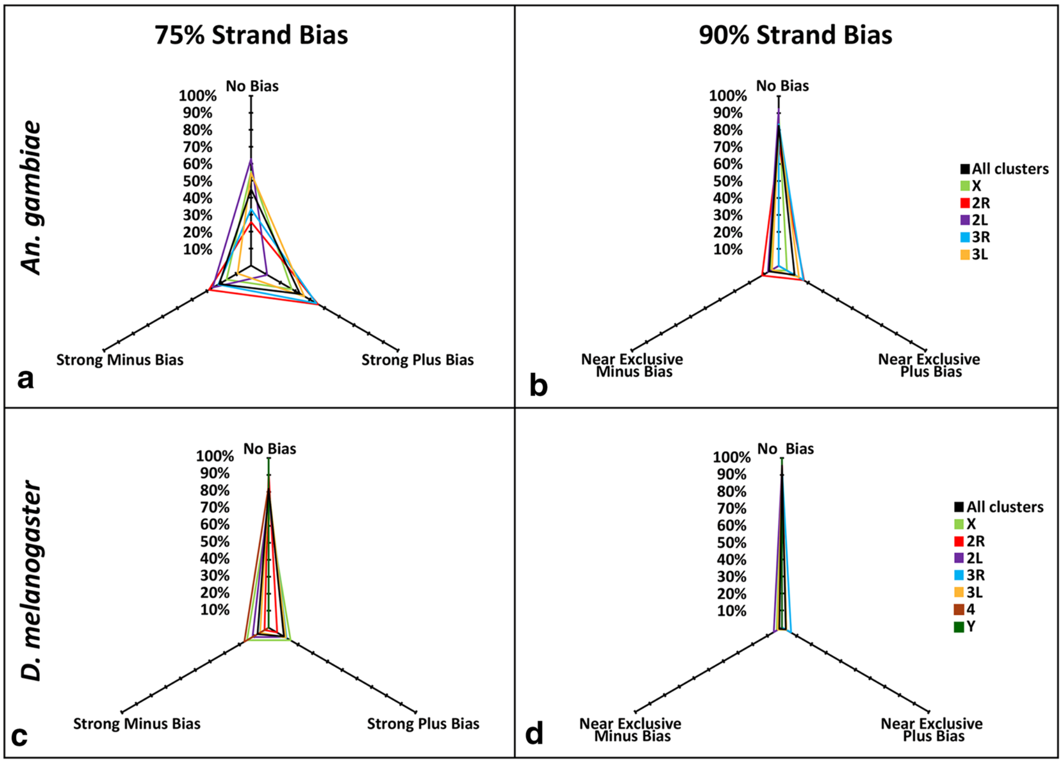


Fig. 5 *Anopheles gambiae* has more unidirectional piRNA clusters than does *D. melanogaster*. Spider chart depicting bias (>75 % of piRNAs map to one genomic strand) (**a**, **c**) or (>90 % of piRNAs map to one genomic strand) (b, d) for clusters per chromosomal arm as well as all clusters combined. The percentage of clusters on each arm (or total) is displayed for either bias or no bias in *An. gambiae* (**a**, **b**) and *D. melanogaster* (**c**, **d**)

subdivisions 2R:42AB, X:20A, and 2L:38C as intercalary heterochromatin instead of pericentromeric heterochromatin based on the presence of late replication sites and location of Suppressor of Under Replication (SuUR) in wild-type flies [49]. The cluster in region X:20A included the *flamenco* locus and the second top cluster from an earlier study [2]. Our approach was designed to determine genomic loci that are responsible for the most abundant production of piRNAs in order to compare piRNA clusters between species, rather than to identify precise boundaries of clusters. We found that 11 of the top 15 chromosomally mapped piRNA clusters are pericentromeric, while the remaining clusters were within intercalary heterochromatin or the euchromatin of chromosome 4 (Fig. 6b).

To determine the chromosomal location of the top piRNA clusters in *Ae. aegypti* [22], we identified the cluster containing supercontigs on the physical [38] and linkage [40] maps of this species. We were able to find chromosomal locations for 24 of top 30 piRNA clusters of *Ae. aegypti* reported previously [22] (Additional file 7: Table S5 and Additional file 8: Figure S6). None of these clusters were located in pericentromeric heterochromatin, but all major clusters were detected in euchromatic regions of all three chromosomes of *Ae. aegypti* (Fig. 6c). Analysis of these data provide evidence that the degree of confinement of the top 15 piRNA clusters to the pericentromeric heterochromatin varies among the three dipteran species from 73.3 % in *D. melanogaster* to 46.7 % in *An. gambiae* to 0 % in *Ae. aegypti*. If we consider piRNA clusters from the UNKN chromosome of *An. gambiae*, four of them would be among the top 15 clusters (ranking 4, 12, 13, 14). However, we cannot assign them to any heterochromatic or euchromatic region.

piRNA production shifted from pericentromeric regions to the rest of the genome in the mosquito compared with the fruit fly

We investigated possible differences in piRNA production between *An. gambiae* and *D. melanogaster* when considering the genomic location of piRNA clusters. A previous study indicated that pericentromeric clusters are the primary production sites of piRNAs in *D. melanogaster* [2]. Although we reclassified the three highest piRNA mapping pericentromeric clusters as intercalary heterochromatin, we still found that pericentromeric regions in *D. melanogaster* produce 41.6 % of all genome-unique piRNAs (Additional file 9: Table S4). In contrast, pericentromeric clusters in *An. gambiae* produce only

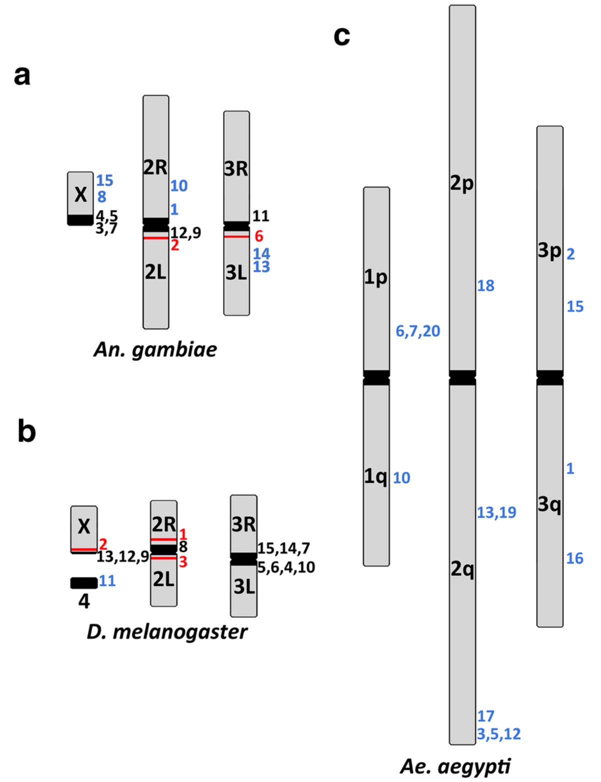


Fig. 6 The chromosome distribution of the top 15 mappable piRNA clusters within three Dipteran species. **a** *An. gambiae*, **b** *D. melanogaster*, **c** *Ae. aegypti*. Clusters are numbered to the right of the chromosome arms, with *colors* (*gray* euchromatic, *black* pericentromeric heterochromatin, *red* intercalary heterochromatin) indicating predicted chromatin type

24.8 % of unique piRNAs, excluding the UNKN chromosome (Additional file 5: Table S3). Even if we include the UNKN chromosome in the analysis and artificially assign all UNKN chromosome clusters to the pericentromeric heterochromatin, still only 34.3 % of unique piRNAs would be produced by the pericentromeric clusters in *An. gambiae.* In addition, 29.3 and 6.6 % of unique piRNAs are derived from euchromatic clusters in *An. gambiae* and *D. melanogaster,* respectively, if we exclude the UNKN chromosome. Still, euchromatic clusters in *An. gambiae* would produce 17.6 % of unique piRNAs, if we include the UNKN chromosome in the analysis. Another interesting difference we found is that as much as 43.2 % of total unique piRNAs is produced outside clusters identified in our study in *An. gambiae* vs. 22 % in *D. melanogaster,* further supporting the shift in piRNA production from pericentromeric regions to the rest of the genome when comparing the mosquito with the fruit fly.

To delve into possible reasons for the difference in piRNA production, we compared TE distribution landscapes between the mosquito and fruit fly. We determined the TE content in the genomes and piRNA clusters using RepeatMasker [45] and a library of 7080 annotated Hexapoda TEs sequences from RepBase [42]. Only 41.8 % of pericentromeric heterochromatin was covered by TEs in *An. gambiae* (Additional file 10: Figure S4), while 77.3 % of pericentromeric heterochromatin was occupied by TEs in *D. melanogaster* (Additional file 11: Figure S5). The rest of the genomes had a similar TE coverage: 6.7 % in *An. gambiae* and 5.5 % in *D. melanogaster.* There was good correspondence between piRNAs and TE sequences in the *An. gambiae* (Additional file 10: Figure S4) and *D. melanogaster* (Additional file 11: Figure S5) genomes. Of the 120 total mapped piRNA clusters in *An. gambiae*, the 26 pericentromeric clusters were enriched in TE sequences (48.1 %) in comparison with the remaining heterochromatic and euchromatic clusters (22.8 % TE content) (Additional file 5: Table S3). Altogether, the 120 mapped clusters had an average TE content of 31 % (Additional file 5: Table S3). The clusters identified on chromosome UNKN had ~40 % TE content, suggesting that they are likely unassembled sequences belonging to the pericentromeric and/or intercalary heterochromatin. However, a much more dramatic difference in TE content could be seen in *D. melanogaster* between the 76 pericentromeric clusters, having an average of 84.3 % TE content and the remaining heterochromatic and euchromatic piRNA clusters having 32.0 % TE content (Additional file 9: Table S4). The very high abundance of TEs in *Drosophila* pericentromeric heterochromatin could be responsible for increased piRNA production by this genomic domain of the fruit fly. On the other hand, the top 24 mapped clusters in *Ae. aegypti*, which are mainly euchromatic (Additional file 8: Figure S6), have a mean TE content of 47 % (Additional file 7: Table S5) [22]. Thus, the lower overall TE content in piRNA clusters of *An. gambiae* compared to that in two other species suggests that other sequences (unidentified TEs or genes) could serve as progenitors of piRNAs.

We documented a marked shift in major piRNA cluster location: from mainly pericentromeric heterochromatin in *D. melanogaster*, to both intercalary heterochromatin and euchromatin in *An. gambiae*, to mainly scattered euchromatic regions in *Ae. aegypti* (Fig. 6). The intercalary heterochromatin regions in *An. gambiae* are, like the pericentromeric heterochromatin, high in TE content and low in genic content [37]. Two regions of diffuse intercalary heterochromatin regions are responsible for a substantial portion of the piRNA population. However, one region of compact intercalary heterochromatin, located in subdivision 35B of 3R arm, does not have enough piRNA enrichment to be considered a cluster. Compact intercalary heterochromatin has a different composition than the diffuse intercalary and

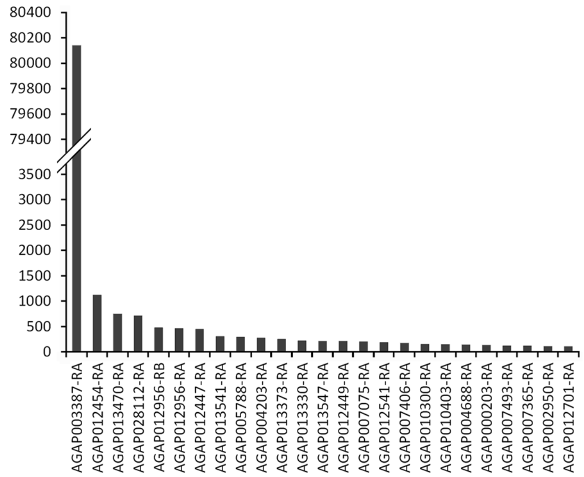


Fig. 7 The *An. gambiae* genes with substantial piRNA enrichment. The top 25 genes with piRNA mapping are shown. The mapping includes AGAP003387, accountable for 8 % of genic piRNAs. The *Y* axis indicates the number of reads per million genome-mappers

pericentromeric heterochromatin. For example, the Ty3/gypsy TEs represent 6.9 % of the diffuse intercalary heterochromatin and only 1.6 % of the compact intercalary heterochromatin in *An. gambiae*. piRNAs mapped to Ty3/gypsy represent the most substantial class of TE-mapped piRNAs. In addition, subdivision 35B of 3R arm has a higher enrichment of solo-LTR retrotransposons than any other region in the genome (euchromatic or heterochromatic) [50].

An. gambiae ovaries have an abundance of gene-derived piRNAs and siRNAs

In *An. gambiae*, a peak of 21-nt siRNAs is observed for gene transcript mapping small RNAs (Additional file 12: Figure S7). One of the most notable roles for siRNAs is within the RNA interference (RNAi) pathway, where the endogenous siRNAs help to regulate gene expression through Dicer-mediated cleavage [8]. Exogenous siRNAs, on the other hand, play crucial roles in the defense against RNA arboviruses in *Aedes* mosquitoes [28, 29]. However, the exogenous siRNA pathway plays no detectable role in antiviral defense in the midgut of *An. gambiae* [51]. We show that gene-derived small RNAs from ovarian tissue comprise both siRNA (21 nt) and piRNA (24–29 nt) populations in different proportions, depending on the genes (Additional file 12: Figure S7). Some transcripts have relatively high levels of siRNA mapping (Additional file 12: Figure S7B, C). As piRNAs are more widely associated with TE sequences, the presence of a large fraction of gene-derived piRNAs, one that is larger than even the siRNA contingent (Additional file 12: Figure S7A), is an interesting observation. By looking at their mapping sites within transcripts, we further analyzed this sub-population of genic piRNAs, as well as the piRNA-enriched gene's functions.

Previous reports have shown that a majority of piRNAs associated with genes are derived from the 3′ UTRs of the transcripts [19–23]. The 3′ UTR has been associated with mRNA localization, translation, and stabilization [52], which all support post-transcriptional regulation of gene expression. MicroRNAs (miRNAs), another class of small RNAs, can bind to the 3′ UTR region of transcripts and affect gene expression through translational inhibition or transcript degradation [53]. To evaluate if piRNAs play a role in regulation of gene expression in *An. gambiae*, the original small RNA library was modified to contain non-collapsed, unique (NCU) piRNAs totaling 6,805,309 sequences ranging from 24 to 29 nt (Additional file 2: Table S1). We localized the NCU piRNA library to annotated *An. gambiae* gene transcripts and identified 5024 genes with more than 0.5 reads per million (RPM) genome-unique piRNAs that mapped within the transcripts. piRNA reads were normalized using RPM in order to compare mapping trends between *An. gambiae* and *D. melanogaster*. However, upon further analysis, many gene transcripts appear to have TE sequences within them. We identified 69 transcripts from 65 genes that gave more than 0.5 RPM of TE-matching piRNAs (allowing 0–3 mismatches) (Additional file 13: Table S6). Five of these genes, AGAP012494-RA, AGAP003870-RA, AGAP001582-RA, AGAP005927-RA, and AGAP000983-RA, had more than half of the piRNA reads mapped to TEs. We analyzed in detail the 65 genes and their piRNAs. The gene sequences that map to TE-matching piRNAs are mostly restricted to short stretches of 30–100 nt. There is one exception with AGAP012494-RA, where TE-matching piRNAs cover 660 nt, corresponding to 90 % of the transcript. The AGAP012494-RA transcript has 90 % sequence identity over 685 nt to AgaP8MITE2450, a DNA transposon. The gene generates 183.8 RPM piRNAs, of which, 162.4 RPM piRNAs map to AgaP8MITE2450 when allowing 0–3 mismatches. These data suggest that AGAP012494 is derived from an ancient AgaP8MITE2450 element. All transcripts' piRNAs, which also map to TEs when allowing 0–3 mismatches, equal 14,350 reads (859.7 RPM). About half of these piRNAs, 49.6 %, map to RepBase reference TEs from *Anopheles* without any mismatches, suggesting a rather recent origin of the corresponding sequences within protein-coding gene transcripts. The piRNAs map to many types of TEs: SINEs, LINEs, Gypsy-like, BEL-like and DNA transposons (Additional file 14: Table S7). Still, a large number of these piRNAs, 5346 reads, map to SINEX-1_AG, a SINE element. We then removed the piRNAs that map to both gene transcripts and TEs, and used the remaining set of piRNAs for further transcript analysis. We identified a single euchromatic gene, AGAP003387, which had 88,391 RPM bona fide piRNAs (24–29 nt) (81 % of the genic piRNAs) localized to its transcript's sequence (Fig. 7). Nearly all of these hits correspond to two short regions within the 3′ UTR of AGAP003387, in agreement with previous studies [21, 23]. AGAP003387 is a putative lipoprotein gene with orthologs restricted to the closely related species of the *An. gambiae* complex. This gene has no significant resemblance to any currently identified TEs, and shows a similar RNA expression profile to a variety of protein-coding genes. Expression of AGAP003387 significantly increases 9.3-fold between 0 and 10-day-old adult female mosquitoes [54]. Exclusion of AGAP003387 from the analysis results in 2.1 % of the total piRNA pool being generated from transcripts of the remaining gene dataset (Fig. 2b). Nearly 58 % of the unique piRNAs that mapped to genes were located within the 3′ UTR of gene transcripts. Approximately 800 genes with any piRNA mapping had piRNA reads entirely derived from the 3′ UTR.

piRNAs have been mapped primarily to the sense strand of various genes in *D. melanogaster* [19, 20] and *M. musculus* [55], including *Traffic jam*, *brat*, and *Klp10.* We saw a similar phenomenon in *An. gambiae*, where 81 % of genome-unique piRNAs matching protein-coding transcripts were derived from the sense strand. This, in concordance with previous reports [21–23], likely suggests that in mosquitoes, there is a mechanism that generates piRNAs from various non-TE-related sequences, protein-coding genes in particular. Most of these piRNAs, lacking their complementary antisense partners, are not fed through the ping-pong amplification cycle. The resulting sense bias of these piRNAs indicates that the small RNAs do not serve as complementary guides for targeted transcript degradation as characterized in TE regulation. Rather, it is more likely that these piRNAs interact directly with sequences elsewhere in the genome, or have a yet unknown role in gene regulation.

piRNA-producing genes play a role in regulation of gene expression, reproduction, and development in *An. gambiae*

The piRNAs stemming from non-TE sources are less studied, and the process and function of these piRNAs have yet to be elucidated. As part of this study, we aimed to identify genes enriched in piRNA mapping and to classify them based on predicted Gene Ontologies (GO). Functional annotation of these genes may provide insight as to why some transcripts are at the origin of a large subset of piRNAs and when these piRNAs may be generated. Transcripts for annotated genes were used as reference sequences to identify piRNAs derived from RNA sequences post-splicing. Based on prior piRNA mapping enrichment, genes were assigned to three groups for further exploration: genes with a unique-mapping piRNA RPM of 5–10, genes with an RPM of 10–50, and genes with an RPM higher than 50. Each list of genes was submitted to the DAVID functional annotation tool [56, 57] to determine potential GO terms.

The DAVID annotation tool provides automatic clustering capabilities, resulting in collections of genes with similar functions and processes. We used an EASE score cutoff of greater than 1.3 to recognize relevant annotation clusters. The EASE score represents the mean of p values from a cluster, where a score of 1.3 is equivalent to a p value of 0.05. The 5-10 RPM piRNA group consisted of 342 annotated genes that were separated into 14 significant clusters. The 10–50 RPM piRNA group contained 263 annotated genes, creating six distinct, significant annotation clusters (Additional file 15: Table S8). Forty-six annotated genes had more than a 50 RPM, with no significant biological meanings being shared amongst the group. GO term analysis resulted in many various predicted functional annotation terms (FATs) for the two gene subsets with significant clustering. From the different FATs that were identified, similar broad functions for the most enriched clusters of both subsets could be identified. FATs for the 5–10 RPM piRNA group range from translation initiation, protein binding, DNA and RNA binding, methyltransferase activity, to chromatin binding. Many of these FATs suggest potential roles in regulation of gene expression and chromosome organization.

FATs for the 10–50 RPM piRNA group included chromatin assembly and organization, nucleosome assembly and organization, protein transport and localization, and regulation of translation. These terms suggest that proteins from this subset of genes are also involved in regulation of gene expression. One cluster of 14 genes, comprising the third most enriched cluster of the 10–50 RPM piRNA group, was identified to represent functions related to development, gamete generation, and sexual reproduction (Fig. 8). Four of these genes are found within piRNA clusters (AGAP000145, AGAP001157, AGAP002651, AGAP005134), and the rest are located outside. Because reproduction and development of mosquitoes are important biological functions to target while designing novel vector control strategies, the identification of a number of piRNA-enriched genes associated with these functions warrants further investigation. For example, *oskar* (AGAP003545), one of the 14 "reproduction and development" cluster genes, produces mRNAs that are restricted to female ovaries and the posterior pole of mosquito embryos [58]. In *D. melanogaster, oskar* is associated with pole plasm determination and axis specification within oocytes [59]. The *Drosophila oskar* has been

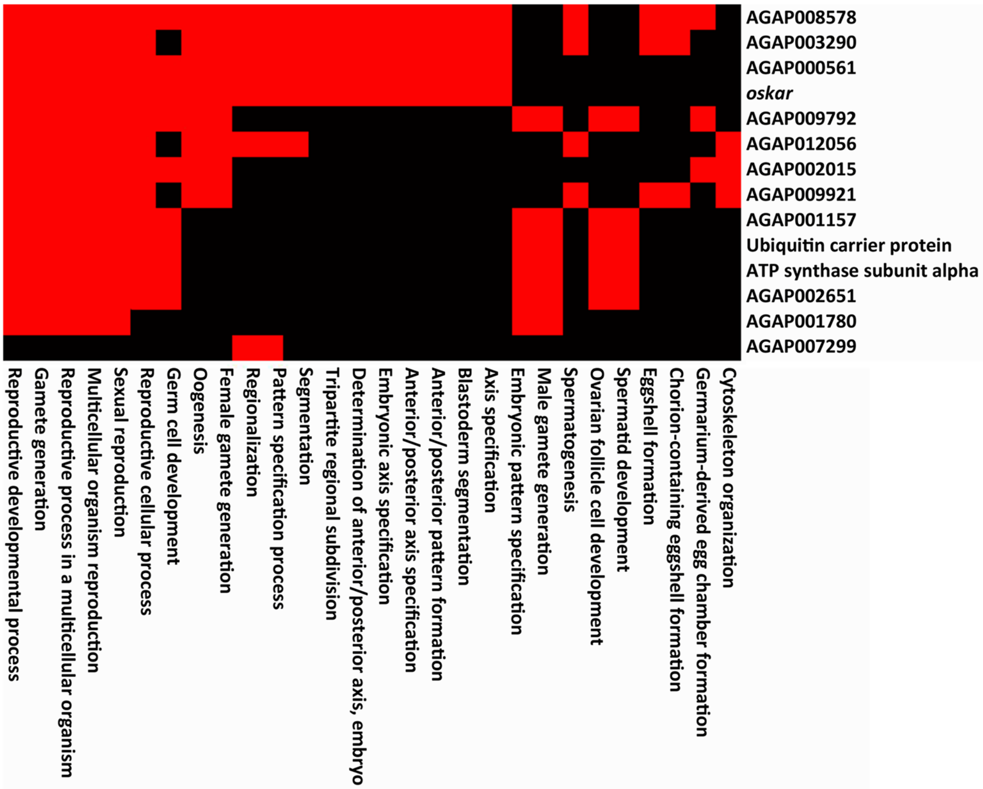


Fig. 8 DAVID functional annotation terms of the "reproduction and development" cluster genes with 10–50 RPM mapped unique piRNAs in *An. gambiae*. Genes within the cluster (*Y* axis) are overrepresented by GO terms (*X* axis), with many of the genes sharing similar predicted functions (*red boxes*)

shown to be repressed in early oocyte development, but mutations in the piRNA pathway proteins result in ectopic expression in early oocytes as well as defects in germline development [6, 59]. Other *An. gambiae* genes in the "reproduction and development" cluster have orthologs in *D. melanogaster* where they have been experimentally linked to germline development and maintenance, spermatid development, and oogenesis (Additional file 16: Table S9). Another gene from this cluster, AGAP000561, is an ortholog of the *D. melanogaster* Kinesin heavy chain (FBgn0001308), which plays a role in *oskar* mRNA localization to the pole plasm [60].

We further analyzed expression profiles for the group of 14 reproduction- and development-related genes. These expression profiles, available in the Expression Browser [54] through VectorBase [61] were analyzed to potentially discover common trends associated with these 14 genes that may provide insight into the production pattern of these piRNAs (Fig. 9). Nine of the 14 genes exhibited a down-regulation at 3 h after blood feeding. The 3-h time point is likely when metabolism functions are shifted toward blood digestion [62], making this part of the gonotrophic cycle a critical stage for proper egg formation. Twelve of 14 genes showed an up-regulation 24 h after blood feeding indicating their possible role in ovarian development (Fig. 9a). This expression pattern was similar to that of many genes involved in the piRNA pathway: 11 of 15 genes were down-regulated at 3 h after blood feeding, and all 15 genes were up-regulated at 24 h after blood feeding (Fig. 9b). Our piRNA library is derived from ovaries dissected at 24 h after blood feeding, and the high numbers of piRNAs mapped to the reproduction- and development-related genes suggest interplay between piRNA production and gene transcription (see "Discussion"). Furthermore, 11 of 14 genes from the "reproduction and development" cluster exhibited a pronounced decline in expression between 2 and 6 h of embryonic development, followed by a relatively constant level of expression afterward (Fig. 9c). Similarly, expression of 10 of 15 genes involved in the piRNA pathway also declined between 2 and 6 h of embryonic development (Fig. 9d). The transcripts of the 15 piRNA pathway genes themselves can produce piRNAs as well, but the amount is approximately 2.5 times less than the amount of piRNAs derived from transcripts of the "reproduction and development" cluster genes (Additional file 17: Table S10). This suggests that

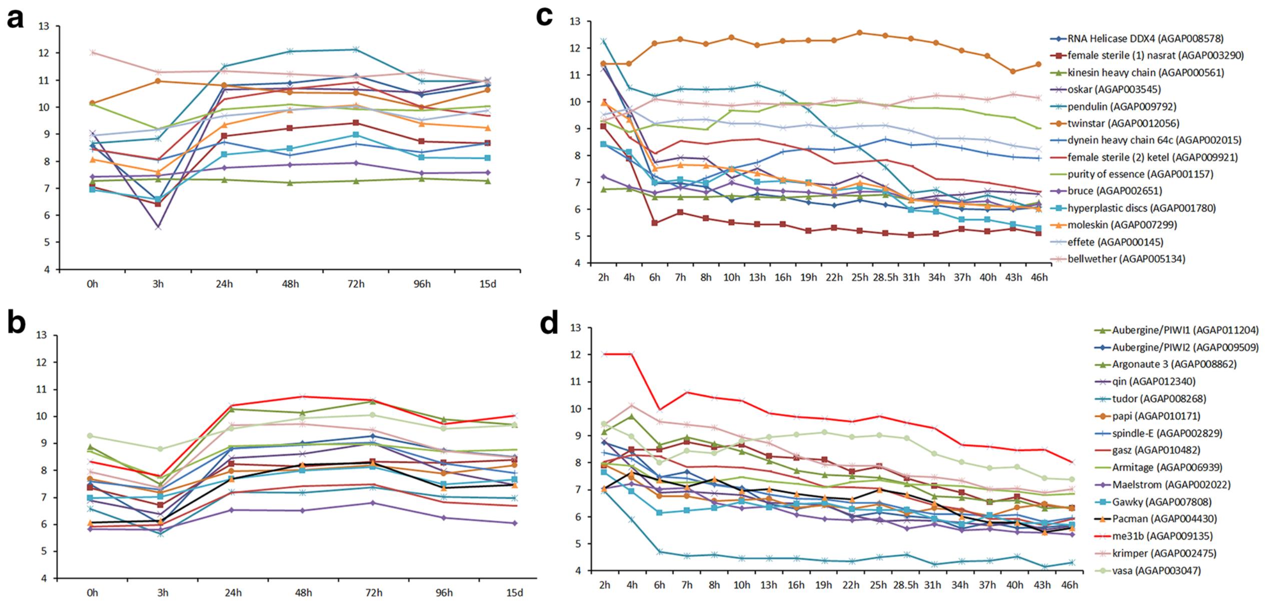


Fig. 9 Trends in gene expression of 14 "reproduction and development" cluster genes and *An. gambiae* genes involved in the piRNA pathway in relation to time post-blood feeding (**a**, **b**) and embryonic development (**c**, **d**)

the piRNA production from the "reproduction and development" cluster genes is rather specific, and might have a functional role in development (see "Discussion").

Discussion

A role of the piRNA pathway in the TE mobilization control and in germline development has been demonstrated for *D. melanogaster* [1]. *Anopheles gambiae* and *D. melanogaster* delineated approximately 260 million years ago (Fig. 1), and the degree of conservation of piRNA functions between these two species is an open question. Moreover, Dipteran species vary greatly in genome size and the pattern of genomic distribution of TEs: from a highly compartmentalized, small genome (144 Mb) in *D. melanogaster* with TEs concentrated in pericentromeric heterochromatin [34], to an intermediate genome size (273 Mb) in *An. gambiae* [33], to a large genome (1310 Mb) in *Ae. aegypti* with homogeneously high TE coverage [35, 38]. How does the organization of the piRNA pathway change with varying genome size and repeat landscape? By studying piRNAs in *An. gambiae*, we uncovered conserved and diverse features of the piRNA pathway across Dipteran species and gained new insights into its role in the regulation of the reproductive processes in the African malaria vector.

The *An. gambiae* piRNA pathway chiefly defends against mobilization of transposable elements

Employing the "overlap" method (see above), we found that almost 40 % of the piRNAs mapped to TEs in *An. gambiae*, which is almost twice higher than 23.6 % determined by the traditional "consensus" method, but still twice less than the 81 % of piRNAs mapped to TEs by the "overlap" approach in *D. melanogaster.* This difference may also suggest that the mosquito has a larger proportion of non-transposon-derived piRNAs than the fruit fly. Alternatively, due to the different relative abundance of TE classes, the secondary quantity of LTR retrotransposons in *An. gambiae* may, at least partially, explain the lower percentage of TE-mapping piRNAs.

We show that TE families with the largest number of derived piRNAs do not correspond to the most prevalent TE families in the mosquito genome. DNA transposons and non-LTR retrotransposons are the most abundant subclasses of TEs in *An. gambiae* (Fig. 3b, d), yet LTR retrotransposons are responsible for the origin of 60 % of the TE-derived piRNAs identified by the "overlap" method (Fig. 3f). A study of *An. gambiae* TEs concluded that elements from the major families of LTR elements (BEL/Pao, copia, and gypsy) correspond to putatively active elements [63]. Many of the DNA transposons are highly degraded, with only Mariner, P elements, and MITE*s* showing full-length sequences that can be associated with recent transposition [64]. Although some elements may be prevalent within the genome, increased activity within the ping-pong amplification loop, TE mRNA transcript abundance, and presence within clusters have all been found to be important in determining piRNA abundance [65]. Moreover, as in *Drosophila* [65], the most transpositionally active TE families may not be necessarily the TEs that give the most abundant piRNAs. The preference toward specific classes of TEs may help to explain how other classes of TEs (DNA transposons) have become more pervasive in the mosquito genome.

Peculiar aspects of genomic distribution and organization of piRNA clusters in *An. gambiae*

In *D. melanogaster,* the piRNA clusters are predominantly heterochromatic, and found within pericentromeric regions [2]. Prior to this study, it was unclear if piRNA clusters are mainly located in pericentromeric heterochromatin in other Dipteran species. Using an approach that was standardized between *D. melanogaster* and *An. gambiae*, we identified clear differences in the genomic location and composition of the piRNA clusters among Dipteran species. piRNA clusters were concentrated in the pericentromeric heterochromatin of *D. melanogaster*, while an increasing number of clusters were found in the intercalary heterochromatin and euchromatin of *An. gambiae. Aedes aegypti* serves as the most extreme example of piRNA cluster shifting, having major clusters mainly located in euchromatic loci. The fact that we find no identified piRNA clusters in the compact intercalary heterochromatin of *An. gambiae* influences the overall findings from this study. Our data suggest that a genomic locus being a region of heterochromatin, *i.e.,* low abundance of genes and high repeat content, is not sufficient for containing piRNA clusters, at least in *An. gambiae*.

We hypothesize that as the genome size decreases in insects, the heterochromatin location becomes more restricted to the pericentromeric regions, and TEs, concordantly, become concentrated within the heterochromatin. Similarly, as the genome size increases, heterochromatin and TEs spread into new chromosomal regions. The expansion of piRNA cluster location from pericentromeric regions follows the shift in TE location from pericentromeric heterochromatin to intercalary heterochromatin and to euchromatin. This shifting genomic distribution results in piRNA clusters with attributes more closely resembling euchromatin than heterochromatin. Change in cluster composition may change the pattern of TE and gene regulation [66]. In the *D. melanogaster* genome with well-compartmentalized TE locations, loci in pericentromeric heterochromatin are responsible for production of 41.6 % of piRNAs (Additional file 9: Table S4). piRNA clusters require the H3K9me3 repressive histone mark to be transcribed and to silence TEs in fruit flies [67]. It has been proposed that Piwi binds to heterochromatin protein 1 (HP1), which in turn recruits Su(var)3-9 to add the H3K9me3 marks [68]. Such a positive feedback loop creates conditions for piRNA cluster transcription and for stabilizing the heterochromatin. Whereas in the *An. gambiae* genome with a high TE density not confined to pericentromeric regions, only 25 % of piRNAs are produced by pericentromeric heterochromatin. We propose that the organization of the majority of piRNA clusters in the malaria mosquito is more similar to that in silkworm than in fruit fly. The total estimated genome size of *Bombyx mori* is 428.7 Mb [69], which is 3.6 and 1.54 times larger than that of fruit fly and malaria mosquito, respectively. Silkworm telomeres have telomere-specific transposons and large piRNA clusters marked with heterochromatin histone marks H3K9me2 and H3K9me3. However, 965 piRNA clusters, which are located outside the telomeres, lack the heterochromatin marks, but have euchromatin marks H3K4me2, H3K4me3, and H3K9ac [70].

Heterochromatic piRNA clusters in *D. melanogaster* are enriched in the HP1 homolog Rhino that co-localizes with UAP56, which binds to piRNA precursors [48, 71]. *An. gambiae* lacks the Rhino ortholog (according to FlyBase [72], http://flybase.org/reports/FBgn0004400.html) and has a higher abundance of unidirectional euchromatic piRNA clusters (~55 %) than does *D. melanogaster*, where a majority of piRNA clusters (~78 %) are bidirectional (Fig. 5). Future studies are needed to understand the nature of the majority of piRNA clusters in the mosquito. The dispersal of piRNA clusters in euchromatin may lead to more potential cases of *cis*-regulation, in which a TE or gene is regulated by a cluster located nearby, instead of *trans*-regulation, where the regulation is provided by a pericentromeric heterochromatic cluster. Finally, manipulating the piRNA function in species with spread out clusters is likely more challenging because of the greater decentralization of the piRNA pathway.

Novel insights from *An. gambiae* genes enriched in sense piRNAs

We identified a large proportion of sequenced piRNAs (11 %) that is associated with protein-coding genes in *An. gambiae.* In addition, 58.3 % of piRNAs are mapped to unannotated sequences, a portion of which may represent novel protein-coding genes, TEs, or long noncoding RNA genes (Fig. 2b). A majority of the piRNAs that mapped to protein-coding genes in *An. gambiae* was found in the sense orientation with respect to the transcript. Recent studies provide insights into how these genic piRNAs might be produced. Artificial insertion of a sequence within the 3′ UTR of a piRNA-producing gene resulted in the formation of piRNAs derived from the novel sequence [73]. Genes with TE sequences, considered "trigger piRNAs" within the transcript, can initiate secondary 3′-directed piRNA biogenesis that result in "responder" piRNAs that are in the same orientation as the genic mRNA [74–76]. These piRNAs lack partner piRNAs, and they are most likely generated outside of the ping-pong amplification loop, unless they target other, yet undiscovered genes. For example, piRNAs produced by *traffic jam* are loaded onto Piwi to silence specific target gene *fasciclin 3* in *D. melanogaster* [19].

Can the sense strand bias give us a clue about a possible mechanism of gene regulation via piRNAs? For example, the *An. gambiae oskar* had 398 genome-unique sense reads and no antisense reads (24–29 nt, zero mismatches), meaning that all these piRNAs derive from the *oskar* primary transcript or mRNA. Could a high level of piRNA production lead to a depletion of mRNAs and, thus, to reduced levels of the protein? This would be possible if the piRNA and mRNA production pathways compete with each other by using the same primary transcripts as a source for either piRNAs or mRNAs. This also would be possible if piRNAs are produced from mRNAs. Because the levels of piRNA production are likely regulated by the PIWI proteins, we investigated expression profiles of genes involved in the piRNA pathway in *An. gambiae* blood feeding and embryonic development experiments. Indeed, we found a good correspondence between the patterns of expression of genes from the "reproduction and development" cluster and genes from the piRNA pathway (Fig. 9). A possible reason for these parallel trends of expression is that the piRNA machinery processes primary gene transcripts or mRNAs into piRNAs. For example, when up-regulation of "reproduction and development" genes creates a high abundance of transcripts, the parallel up-regulation of the PIWI genes would process some of these gene transcripts into piRNAs, thus reducing the amount of mRNA. Otherwise, these reproductive genes and piRNA pathway genes could be regulated by the same factors. The idea that piRNAs can be produced from primary transcripts or from mRNAs is supported by a correlation analysis, which found that more piRNAs are produced by more highly expressed genic transcripts [77]. Nevertheless, many highly expressed transcripts do not produce piRNAs, suggesting restricted access of substrates to the piRNA biogenesis machinery. Furthermore, it has been shown that the protein level of the *traffic jam*, whose 3′ UTR generates abundant sense piRNAs, is upregulated in *Drosophila piwi* mutants [20]. Studying the functional role of these piRNAs would require knocking down the piRNA production from all these specific genes without changing their protein expression.

If piRNAs do indeed play a role in regulation of mRNA abundance, it may be to aid in the transition from maternally deposited mRNA to zygotically transcribed mRNA levels. Many of the genes from the "reproduction and development" cluster and from the piRNA pathway exhibited a pronounced decline in expression between 2 and 6 h of embryonic development (Fig. 9). As described earlier, zygotic transcription occurs as early as 2–3 h into embryogenesis [78]. Prior to this, large quantities of mRNAs are maternally deposited into the embryo. Since we see such a marked decrease in the abundance of mRNAs within this cluster of reproductive development genes, it may be possible that the processing of mRNAs into piRNAs is an underlying cause of the decline of maternally deposited mRNAs.

Another gene regulation mechanism is transcript suppression, which occurs in many *D. melanogaster* piRNA clusters that exhibit heterochromatic characteristics, including the epigenetic marks H3K9me3 and HP1 [67, 79]. In *D. melanogaster*, maternally deposited embryonic piRNAs are replaced by TE-derived siRNAs [80], which are posited to aid in the spread of heterochromatin formation through a shift in recruitment of HP1 and related proteins, contributing to gene silencing. If similar epigenetic modifications occur in regions typically devoid of heterochromatin marks, it is possible that the recruitment of heterochromatic proteins in or near these euchromatic regions containing protein-coding genes results in repressed transcription through heterochromatinization. It is unclear if the chromatin state itself is what allows piRNA production [48]. Given the nuclear localization of Piwi, it is conceivable that the Piwi-piRNA complex could associate with the piRNA-producing genes, providing an epigenetic transcriptional regulation through histone modification [43].

Conclusion

Our study demonstrates that the distribution and organization of piRNA clusters observed in fruit flies may not be conserved in other Dipteran species. The more decentralized genomic location of piRNA clusters in *An. gambiae* and *Ae. aegypti* compared with *D. melanogaster* could potentially cause the generation of a larger proportion of non-TE-derived piRNAs. The large number of piRNAs that originate from regions other than pericentromeric heterochromatin suggests that their roles may be more diverse in mosquitoes than in *D. melanogaster*. Moreover, *An. gambiae* has a higher abundance of unidirectional euchromatic piRNA clusters than *D. melanogaster* does, which points to important differences between the piRNA machineries of the two species. Identification of the large pool of piRNAs produced by genes involved in reproduction and development indicates that the piRNA pathway may play a role in reproductive processes in the malaria vector. Future research will lead to understanding the epigenetic mechanisms of how these piRNAs regulate gene expression and affect germline and embryonic development. Our study also suggests that LTR retrotransposons have a distinct capacity to produce piRNAs that may be linked to higher transcription rate, higher frequency of insertion into piRNA clusters and/or higher capacity to enter the ping-pong amplification cycle. The study of piRNA production from various classes of TEs may offer understanding of the potential ability of some TEs to escape piRNA repression.

Methods

Total RNA isolation and small RNA library construction

Christopher's Stage III ovaries were dissected from 25-h gravid females of the Mali strain (M form) of *An. gambiae* obtained from the Malaria Research and Reference Reagent Resource Center (MR4). Isolated ovaries were preserved in Trizol to prevent RNA degradation. Total RNA was extracted from ovaries of approximately 40–50 individual mosquitoes. RNA was precipitated using a standard phenol–chloroform extraction, solubilized in RNase-free water and stored at −80 °C. Total RNA was isolated, and Illumina sequencing was performed by Fasteris, Inc. on the RNAs ranging from 18 to 32 nucleotides in length. A library containing RNAs ranging in size from 18 to 32 nt was created from a single RNA sample. The library had a sharp peak at 22 nt, with a secondary broad peak ranging from 24 to 30 nt that apexed at 26–27 nt.

Small RNA library modification

A total of three libraries were generated for subsequent piRNA analysis. Libraries corresponded to two specific factors: read count within the library (collapsed vs. non-collapsed) and instances of a read found within the *An. gambiae* genome (unique vs. non-unique). The short-read mapper NucBase was run to determine the number of times each RNA sequence mapped within the *An. gambiae* genome [81]. The initial sequenced library, consisting of 22,569,568 total non-unique, non-collapsed reads from 24 to 29 nt, was used for mapping piRNAs to TEs. A second library of 6,805,309 RNAs containing non-collapsed, unique reads (i.e., redundant reads that map to a single genomic locus) was used for gene analysis. A final library consisting of 568,080 collapsed, unique reads (i.e., non-redundant, single genome-mappers) was used for piRNA cluster identification.

Mapping small RNA reads to reference genome sequences

For initial inter-specific comparative assessment of NCNU piRNA libraries, the small RNA reads, 24–29 nt long, were mapped to the AgamP4, AaegL3, and Dmel Release 6.04 genome assemblies for *An. gambiae*, *Ae. aegypti*, and *D. melanogaster,* respectively, allowing zero mismatches (Fig. 2). For the *An. gambiae* analysis, we used the small RNA library from ovaries that contains 16,691,820 genome-mapping piRNAs. For the *D. melanogaster* analysis, we used the small RNA library from ovaries (sample GSM872307) that contains 7,465,629 genome-mapping piRNAs. For the *Ae. aegypti* analysis, we used the small RNA library from whole adults [22] that contains 388,136 genome-mapping piRNAs.

The resulting genome-mappers were run through size analysis. Annotation of small RNAs was done by mapping the reads on data downloaded from VectorBase [61], FlyBase [72], and MirBase [82] allowing zero mismatches with the exception of TEs being downloaded from RepBase [42] and mapped allowing 0–3 mismatches. The AgamP4 assembly of the *An. gambiae* PEST genome was used as the reference genome for localization by the short-read mapper, NucBase [81]. NucBase counts the number of times a specific RNA sequence maps to the reference genome, allowing the user to filter out all repetitive sequences from the library. We ran NucBase using zero mismatches. As mentioned previously, unique mapping piRNAs were parsed into a secondary, final library that was used to discover the location of piRNA clusters.

BowTie2 was used to map the small RNAs to both TEs and gene transcripts. Default settings were used within the program, with parameters including end-to-end mode, a minimum seed length of 22 nt, and a mismatch penalty of 6. Up to three mismatches were allowed to account for sequence degradation. TEs for the subphylum Hexapoda, available on RepBase [42], were used for all repeat masking and piRNA localization. The AgamP4.1 transcript file was used for identifying piRNAs mapping to genes in *An. gambiae*.

Identification of piRNA clusters

To identify the genomic loci responsible for piRNA generation in *An. gambiae*, data were restricted to sequences that mapped uniquely across the reference genome. All 24–29 nt reads were initially mapped to the *An. gambiae* PEST AGamP4 genome assembly (without repeat masking) using NucBase [81], a short-read mapper designed to align short sequence reads from large nucleic acid databases to genomes or input sequences. After the initial mapping, genome-unique reads, i.e., mapping only once in the *An. gambiae* genome, were retained, and reads present more than once in the library were reduced to one occurrence, resulting in a non-duplicate (collapsed) library of unique mapping reads that was then re-plotted with NucBase. The final mapping run resulted in 568,080 small RNA sequences that mapped to the five chromosomal arms (2R, 2L, 3R, 3L, and X), as well as to the non-assembled chromosome UNKN.

Initial short-read mapping was completed on 5-kb windows for each chromosome arm of the *An. gambiae* AgamP4 assembly. A Perl script was written to identify consecutive windows containing a minimum of ten unique piRNAs that held a single constraint. This constraint allowed the inclusion of gaps of up to four contiguous windows (or 20 kb) containing less than ten piRNAs each that could continue the cluster. The ten piRNA cutoff value used in the *Ae. aegypti* study [22] was necessary to improve cluster detection when working with such a large dataset. To be considered a cluster, 0.05 % of the subset of 568,080 uniquely mapping piRNAs, or 284 piRNAs, was the minimum total piRNAs within a given region.

We ran our piRNA cluster analysis pipeline using the w^{1118} piRNA library (sample GSM327620) [44] on the Dmel_R6.04 release of the *D. melanogaster* [34] genome assembly (without repeat masking), using the same methodology as for *An. gambiae*. We used the unique 24–29 nt sequences as the input piRNA library and excluded auxiliary sequence scaffolds from cluster identification because heterochromatin is better assembled in *D. melanogaster* than in *An. gambiae*. A threshold of 101 piRNAs, 0.05 % of the 202,533 unique piRNA mappers used in the mapping analysis, was set for the minimum number of mappers within consecutive windows that constituted a cluster. Cytological positions of piRNA clusters were determined using the "estimated cytological band" track in the FlyBase genome browser [72].

TE, gene ontology, and expression data analysis

The TE identification was performed by using the RepeatMasker tool to search against all available annotated elements. We elected to use subphylum Hexapoda TEs when performing RepeatMasker searches, as the complete annotation of TEs in *An. gambiae* is still lacking. For "overlap" identification, the genome was masked using the Hexapoda subphylum from RepBase [42] in RepeatMasker [45]. Masked regions (correspondent to TEs) were then mapped by piRNA sequences to identify "overlapping" piRNA-TE sequences and positions, while allowing no mismatches. These positions were then marked to identify the associated TE family/class. We used the −k = 1 option in BowTie2, which reports a single valid alignment for each read in the library. This setting allows the report to identify mapping to TE sequences without over-reporting hits from identical, repetitive sequences. We tested using seed length (allowing for the piRNA to extend past the repeat-masked sequence), but saw minor differences using a seed length of 16 and not using a specific seed. The aligner was run in end-to-end alignment mode. Repeat-masked regions representing TEs were all concatenated into a single file, and piRNA reads were aligned to these sequences allowing no mismatches.

piRNAs were localized to gene exon DNA sequences extracted using BioMART [83–85]. Gene ontology terms for piRNA-mapped genes were identified using DAVID v6.7 [56, 57]. The default settings in DAVID were used with the addition of Bonferroni correction of P values for identification of significant GO terms. Using DAVID's categorization of gene functions, subsets of genes were identified that may be important within the piRNA pathway. Specifically, the subset of genes implicated in reproduction and development was used when looking at expression assays. Differential expression values associated with those 14 genes were further examined to identify potential trends between the genes that could provide insight into the relationship between the gene and the piRNA pathway.

Additional files

Additional file 1: Figure S1. Size distribution of all genome-mapping small RNAs sequenced from the *An. gambiae* ovaries.

Additional file 2: Table S1. piRNA library definitions and sizes for *An. gambiae* and *D. melanogaster.*

Additional file 3: Table S2. The piRNA pool derived from TE sequences in *D. melanogaster* and *An. gambiae.*

Additional file 4: Figure S2. TE-derived piRNA direction bias suggests a ping-pong-like mechanism. % of TE-derived piRNAs in relation to TE orientation is identified by the "overlap" method. % sense indicates the piRNA is in the same orientation as the coding strand of the consensus TE, while % antisense specifies piRNAs complementary to this strand. w/o UN, without chromosome UNKN. w/ UN, with chromosome UNKN.

Additional file 5: Table S3. Distribution and composition of piRNA clusters in the *An. gambiae* genome.

Additional file 6: Figure S3. Structure of the top three piRNA clusters in *An. gambiae*. **A)** Unidirectional euchromatin cluster. **B)** Bidirectional intercalary heterochromatin cluster. **C)** Bidirectional pericentromeric heterochromatin cluster. piRNA mapping (blue for sense reads and red for antisense reads) across the span of an individual cluster. Y axis indicates number of unique piRNAs at a given position (X axis). Repeat-masked TEs are identified by black boxes at the horizontal (X) axis.

Additional file 7: Table S4. Distribution and composition of piRNA clusters in the *D. melanogaster* genome.

Additional file 8: Figure S4. Distribution of TEs and clusters of collapsed, unique piRNAs in the chromosome-based genome assembly of *An. gambiae*. TE percentage across each chromosome arm is broken down by 25-kb windows. Heterochromatin is shown by dark red. Number of mapped 24-29-nt unique piRNAs is shown along chromosomes in 25-kb windows. Clusters and their respective ranks based on piRNA density are shown in light red.

Additional file 9: Table S5. Distribution and composition of top 30 piRNA clusters in the *Ae. aegypti* genome.

Additional file 10: Figure S5. Distribution of TEs and clusters of collapsed, unique piRNAs in the chromosome-based genome assembly of *D. melanogaster*. TE percentage across each chromosome arm is broken down by 25-kb windows. Heterochromatin is shown by dark red. Number of mapped 24-29-nt unique piRNAs is shown along chromosomes in 25-kb windows. Clusters and their respective ranks based on piRNA density are shown in light red.

Additional file 11: Figure S6. Chromosomal distribution of 24 top piRNA clusters in *Ae. aegypti*. Clusters in their respective ranks based on piRNA density are shown in blue.

Additional file 12: Figure S7. Peaks of siRNAs and piRNAs mapped to gene transcript in *An. gambiae*. **A)** Size distribution of small RNAs mapped to all transcripts without AGAP003387. **B)** A high peak of siRNAs mapped to the AGAP001627 transcript. **C)** Peaks of siRNAs and piRNAs mapped to the AGAP001754 transcript.

Additional file 13: Table S6. piRNAs mapping to TEs within genes (0-3 mismatches) in *An. gambiae*.

Additional file 14: Table S7. TE family distribution of genic TE-mapping piRNAs in *An. gambiae*.

Additional file 15: Table S8. GO annotation clusters of genes with high piRNA mapping in *An. gambiae*.

Additional file 16: Table S9. GO annotation of genes in the "reproduction and development" cluster in *An. gambiae*.

Additional file 17: Table S10. piRNA mapping to the piRNA pathway genes and "reproduction and development" cluster genes in *An. gambiae*.

Availability of supporting data

The BioProject PRJNA278159 files are available from the NCBI SRA database (accession number: SRX966734).

Abbreviations

Ago3: Argonaute 3; Aub: Aubergine; FAT: functional annotation terms; GO: gene ontology; H3K9me2: dimethylated histone H3 lysine 9; H3K9me3: trimethylated histone H3 lysine 9; H3K4me2: dimethylated histone H3 lysine 4; H3K4me3: trimethylated histone H3 lysine 4; H3K9ac: acetylated histone H3 lysine 9; HP1: heterochromatin protein 1; LINE: long interspersed nuclear element; LTR: long terminal repeat; miRNA: microRNAs; MITE: miniature inverted-repeat transposable element; MR4: Malaria Research and Reference Reagent Resource Center; mRNA: messenger RNA; NC: non-collapsed; NU: non-unique; NCNU: non-collapsed non-unique; nt: nucleotide; piRNA: Piwi-interacting RNA; rRNA: ribosomal RNA; RISC: RNA-induced silencing complex; RPM: reads per million; SINE: short interspersed nuclear element; snRNA: small nuclear RNA; SuUR: Suppressor of Under Replication; Su(var): Suppressor of variegation; TEs: transposable elements; tRNA: transport RNA; UNKN: unknown; UTRs: untranslated regions.

Authors' contributions

Performed the experiments: PG, SJ, IVS; analyzed the data: PG, SJ, RP, JL, YX, EB, CV, IVS; contributed reagents/materials/analysis tools: RP, JL, YX, CV, IVS; contributed to the writing of the manuscript: PG, SJ, CV, IVS. All authors read and approved the final manuscript.

Author details

[1] Department of Entomology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. [2] Laboratoire Génétique, Reproduction, et Développement, Clermont Université, Université d'Auvergne, BP 38, 63001 Clermont-Ferrand, France. [3] Institut National de la Santé et de la Recherche Médicale, U 1103, BP 38, 63001 Clermont-Ferrand, France. [4] Centre National de Recherche Scientifique, UMR 6293, BP 38, 63001 Clermont-Ferrand, France. [5] The PhD Program in Genomics Bioinformatics and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA.

Acknowledgements

This work is supported by the Fralin Life Science Institute, the USDA National Institute of Food and Agriculture Hatch project 223822, by NIH grant R21AI094289 to IVS and Ligue Nationale contre le cancer to CV. PG received the Chateaubriand Fellowship from the French Embassy in the USA. We thank Melissa Wade for editing the manuscript.

Competing interests

The authors declares that they have no competing interest.

Received: 17 August 2015 Accepted: 4 November 2015
Published online: 27 November 2015

References

1. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet. 2009;10(2):94–108. doi:10.1038/nrg2504.
2. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 2007;128(6):1089–103. doi:10.1016/j.cell.2007.01.043.
3. Reiss D, Josse T, Anxolabehere D, Ronsseray S. Aubergine mutations in *Drosophila melanogaster* impair P cytotype determination by telomeric P elements inserted in heterochromatin. Mol Genet Genomics MGG. 2004;272(3):336–43. doi:10.1007/s00438-004-1061-1.
4. Sarot E, Payen-Groschene G, Bucheton A, Pelisson A. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the *Drosophila melanogaster* flamenco gene. Genetics. 2004;166(3):1313–21.
5. Savitsky M, Kwon D, Georgiev P, Kalmykova A, Gvozdev V. Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. Genes Dev. 2006;20(3):345–54. doi:10.1101/gad.370206.
6. Pane A, Wehr K, Schupbach T. zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. Dev Cell. 2007;12(6):851–62. doi:10.1016/j.devcel.2007.03.022.
7. Nishimasu H, Ishizu H, Saito K, Fukuhara S, Kamatani MK, Bonnefond L, et al. Structure and function of Zucchini endoribonuclease in piRNA biogenesis. Nature. 2012;491(7423):284–7. doi:10.1038/nature11509.
8. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, et al. A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. Science. 2007;315(5818):1587–90. doi:10.1126/science.1140494.
9. Gonzalez-Reyes A, Elliott H, St Johnston D. Oocyte determination and the origin of polarity in Drosophila: the role of the spindle genes. Development. 1997;124(24):4927–37.
10. Lin H, Spradling AC. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. Development. 1997;124(12):2463–76.
11. Wang G, Reinke VAC. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. Curr Biol CB. 2008;18(12):861–7. doi:10.1016/j.cub.2008.05.009.
12. Batista PJ, Ruby JG, Claycomb JM, Chiang R, Fahlgren N, Kasschau KD, et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. Mol Cell. 2008;31(1):67–78. doi:10.1016/j.molcel.2008.06.002.
13. Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell. 2007;129(1):69–82. doi:10.1016/j.cell.2007.03.026.
14. Deng W, Lin H. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev Cell. 2002;2(6):819–30.
15. Cox DN, Chao A, Lin H. Piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. Development. 2000;127(3):503–14.
16. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, et al. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. Dev Cell. 2007;12(4):503–14. doi:10.1016/j.devcel.2007.03.001.
17. Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature. 2014;509(7502):633–6. doi:10.1038/nature13315.
18. Macias V, Coleman J, Bonizzoni M, James AA. piRNA pathway gene expression in the malaria vector mosquito *Anopheles stephensi*. Insect Mol Biol. 2014;23(5):579–86. doi:10.1111/imb.12106.
19. Saito K, Inagaki S, Mituyama T, Kawamura Y, Ono Y, Sakota E, et al. A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature. 2009;461(7268):1296–9. doi:10.1038/nature08501.
20. Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, et al. A broadly conserved pathway generates 3′UTR-directed primary piRNAs. Curr Biol. 2009;19(24):2066–76. doi:10.1016/j.cub.2009.11.064.
21. Castellano L, Rizzi E, Krell J, Di Cristina M, Galizi R, Mori A, et al. The germline of the malaria mosquito produces abundant miRNAs, endo-siRNAs, piRNAs and 29-nt small RNAs. BMC Genom. 2015;16:100. doi:10.1186/s12864-015-1257-2.
22. Arensburger P, Hice RH, Wright JA, Craig NL, Atkinson PW. The mosquito *Aedes aegypti* has a large genome size and high transposable element

load but contains a low proportion of transposon-specific piRNAs. BMC Genom. 2011;12:606. doi:10.1186/1471-2164-12-606.

23. Biryukova I, Ye T. Endogenous siRNAs and piRNAs derived from transposable elements and genes in the malaria vector mosquito *Anopheles gambiae*. BMC Genom. 2015;16(1):278. doi:10.1186/s12864-015-1436-1.
24. Aravin AA, Naumova NM, Tulin AV, Vagin VV, Rozovsky YM, Gvozdev VA. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. Curr Biol CB. 2001;11(13):1017–27.
25. Li MA, Alls JD, Avancini RM, Koo K, Godt D. The large Maf factor Traffic Jam controls gonad morphogenesis in Drosophila. Nat Cell Biol. 2003;5(11):994–1000. doi:10.1038/ncb1058.
26. Touret F, Guiguen F, Terzian C. Wolbachia influences the maternal transmission of the gypsy endogenous retrovirus in *Drosophila melanogaster*. MBio. 2014;5(5):e01529-14. doi:10.1128/mBio.01529-14.
27. Shaw WR, Attardo GM, Aksoy S, Catteruccia F. A comparative analysis of reproductive biology of insect vectors of human disease. Curr Opin Insect Sci. 2015;10:142–8.
28. Hess AM, Prasad AN, Ptitsyn A, Ebel GD, Olson KE, Barbacioru C, et al. Small RNA profiling of Dengue virus-mosquito interactions implicates the PIWI RNA pathway in anti-viral defense. BMC Microbiol. 2011;11:45. doi:10.1186/1471-2180-11-45.
29. Morazzani EM, Wiley MR, Murreddu MG, Adelman ZN, Myles KM. Production of virus-derived ping-pong-dependent piRNA-like small RNAs in the mosquito soma. PLoS Pathog. 2012;8(1):e1002470. doi:10.1371/journal.ppat.1002470.
30. Mayoral JG, Etebari K, Hussain M, Khromykh AA, Asgari S. Wolbachia infection modifies the profile, shuttling and structure of microRNAs in a mosquito cell line. PLoS One. 2014;9(4):e96107. doi:10.1371/journal.pone.0096107.
31. Goriaux C, Theron E, Brasset E, Vaury C. History of the discovery of a master locus producing piRNAs: the flamenco/COM locus in *Drosophila melanogaster*. Front Genet. 2014;5:257. doi:10.3389/fgene.2014.00257.
32. Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science. 2007;318(5851):761–4. doi:10.1126/science.1146484.
33. Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, et al. Update of the *Anopheles gambiae* PEST genome assembly. Genome Biol. 2007;8(1):R5. doi:10.1186/gb-2007-8-1-r5.
34. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 2015;. doi:10.1101/gr.185579.114.
35. Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, et al. Genome sequence of Aedes aegypti, a major arbovirus vector. Science. 2007;316(5832):1718–23.
36. Smith CD, Shu S, Mungall CJ, Karpen GH. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. Science. 2007;316(5831):1586–91. doi:10.1126/science.1139815.
37. Sharakhova MV, George P, Brusentsova IV, Leman SC, Bailey JA, Smith CD, et al. Genome mapping and characterization of the *Anopheles gambiae* heterochromatin. BMC Genom. 2010;11:459. doi:10.1186/1471-2164-11-459.
38. Timoshevskiy VA, Kinney NA, deBruyn BS, Mao CH, Tu ZJ, Severson DW, et al. Genomic composition and evolution of Aedes aegypti chromosomes revealed by the analysis of physically mapped supercontigs. BMC Biol. 2014;12:27. doi:10.1186/1741-7007-12-27.
39. Sharakhov IV, Sharakhova MV. Heterochromatin, histone modifications, and nuclear architecture in disease vectors. Curr Opin Insect Sci. 2015;10:110–7. doi:10.1016/j.cois.2015.05.003.
40. Juneja P, Osei-Poku J, Ho YS, Ariani CV, Palmer WJ, Pain A, et al. Assembly of the genome of the disease vector *Aedes aegypti* onto a genetic linkage map allows mapping of genes affecting disease transmission. PLoS Negl Trop Dis. 2014;8(1):e2652. doi:10.1371/journal.pntd.0002652.
41. Kawaoka S, Hayashi N, Katsuma S, Kishino H, Kohara Y, Mita K, et al. Bombyx small RNAs: genomic defense system against transposons in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2008;38(12):1058–65. doi:10.1016/j.ibmb.2008.03.007.
42. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.
43. Chambeyron S, Seitz H. Insect small non-coding RNA involved in epigenetic regulations. Curr Opin Insect Sci. 2014;1:1–9.
44. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An epigenetic role for maternally inherited piRNAs in transposon silencing. Science. 2008;322(5906):1387–92. doi:10.1126/science.1165171.
45. Smit A, Hubley R, Green, P. RepeatMasker Open-4.0. 2013–2015. http://www.repeatmasker.org.
46. George P, Sharakhova MV, Sharakhov IV. High-resolution cytogenetic map for the African malaria vector *Anopheles gambiae*. Insect Mol Biol. 2010;19(5):675–82. doi:10.1111/j.1365-2583.2010.01025.x.
47. Lawniczak MK, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, et al. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. Science. 2010;330(6003):512–4. doi:10.1126/science.1195755.
48. Le Thomas A, Tóth KF, Aravin AA. To be or not to be a piRNA: genomic origin and processing of piRNAs. Genome Biol. 2014;15(1):204. doi:10.1186/gb4154.
49. Zhimulev IF, Belyaeva ES, Makunin IV, Pirrotta V, Volkova EI, Alekseyenko AA, et al. Influence of the SuUR gene on intercalary heterochromatin in *Drosophila melanogaster* polytene chromosomes. Chromosoma. 2003;111(6):377–98. doi:10.1007/s00412-002-0218-0.
50. Tubio JM, Tojo M, Bassaganyas L, Escaramis G, Sharakhov IV, Sharakhova MV, et al. Evolutionary dynamics of the Ty3/gypsy LTR retrotransposons in the genome of *Anopheles gambiae*. PLoS One. 2011;6(1):e16328. doi:10.1371/journal.pone.0016328.
51. Carissimo G, Pondeville E, McFarlane M, Dietrich I, Mitri C, Bischoff E, et al. Antiviral immunity of *Anopheles gambiae* is highly compartmentalized, with distinct roles for RNA interference and gut microbiota. Proc Natl Acad Sci USA. 2015;112(2):E176–85. doi:10.1073/pnas.1412984112/-/DCSupplemental.
52. Surdej P, Riedl A, Jacobs-Lorena M. Regulation of mRNA stability in development. Annu Rev Genet. 1994;28:263–82. doi:10.1146/annurev.ge.28.120194.001403.
53. Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila MicroRNA targets. PLoS Biol. 2003;1(3):E60. doi:10.1371/journal.pbio.0000060.
54. Maccallum RM, Redmond SN, Christophides GK. An expression map for *Anopheles gambiae*. BMC Genom. 2011;12:620. doi:10.1186/1471-2164-12-620.
55. Gan H, Lin X, Zhang Z, Zhang W, Liao S, Wang L, et al. piRNA profiling during specific stages of mouse spermatogenesis. RNA. 2011;17(7):1191–203. doi:10.1261/rna.2648411.
56. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. doi:10.1038/nprot.2008.211.
57. da Huang W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13. doi:10.1093/nar/gkn923.
58. Juhn J, James AA. oskar gene expression in the vector mosquitoes, *Anopheles gambiae* and *Aedes aegypti*. Insect Mol Biol. 2006;15(3):363–72. doi:10.1111/j.1365-2583.2006.00655.x.
59. Cook HA, Koppetsch BS, Wu J, Theurkauf WE. The Drosophila SDE3 homolog armitage is required for oskar mRNA silencing and embryonic axis specification. Cell. 2004;116(6):817–29. doi:10.1016/S0092-8674(04)00250-8.
60. Brendza RP, Serbus LR, Duffy JB, Saxton WM. A function for kinesin I in the posterior transport of oskar mRNA and Staufen protein. Science. 2000;289(5487):2120–2.
61. Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, et al. VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. Nucleic Acids Res. 2015;43(Database issue):D707–13. doi:10.1093/nar/gku1117.
62. Marinotti O, Calvo E, Nguyen QK, Dissanayake S, Ribeiro JM, James AA. Genome-wide analysis of gene expression in adult *Anopheles gambiae*. Insect Mol Biol. 2006;15(1):1–12. doi:10.1111/j.1365-2583.2006.00610.x.
63. Fernandez-Medina RD, Struchiner CJ, Ribeiro JM. Novel transposable elements from *Anopheles gambiae*. BMC Genom. 2011;12:260. doi:10.1186/1471-2164-12-260.

64. Quesneville H, Nouaud D, Anxolabehere D. P elements and MITE relatives in the whole genome sequence of *Anopheles gambiae*. BMC Genom. 2006;7:214. doi:10.1186/1471-2164-7-214.
65. Kelleher ES, Barbash DA. Analysis of piRNA-mediated silencing of active TEs in *Drosophila melanogaster* suggests limits on the evolution of host genome defense. Mol Biol Evol. 2013;30(8):1816–29. doi:10.1093/molbev/mst081.
66. Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, et al. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. Proc Natl Acad Sci USA. 2013;110(49):19842–7. doi:10.1073/pnas.1313677110.
67. Rangan P, Malone CD, Navarro C, Newbold SP, Hayes PS, Sachidanandam R, et al. piRNA production requires heterochromatin formation in Drosophila. Curr Biol. 2011;21(16):1373–9. doi:10.1016/j.cub.2011.06.057.
68. Huang XA, Yin H, Sweeney S, Raha D, Snyder M, Lin H. A major epigenetic programming mechanism guided by piRNAs. Dev Cell. 2013;24(5):502–16. doi:10.1016/j.devcel.2013.01.023.
69. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, et al. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science. 2004;306(5703):1937–40. doi:10.1126/science.1102210.
70. Kawaoka S, Hara K, Shoji K, Kobayashi M, Shimada T, Sugano S, et al. The comprehensive epigenome map of piRNA clusters. Nucleic Acids Res. 2013;41(3):1581–90. doi:10.1093/nar/gks1275.
71. Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, et al. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. Cell. 2009;138(6):1137–49. doi:10.1016/j.cell.2009.07.014.
72. dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, et al. FlyBase: introduction of the *Drosophila melanogaster* Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015;43(Database issue):D690–7. doi:10.1093/nar/gku1099.
73. Muerdter F, Olovnikov I, Molaro A, Rozhkov NV, Czech B, Gordon A, et al. Production of artificial piRNAs in flies and mice. RNA. 2012;18(1):42–52. doi:10.1261/rna.029769.111.
74. Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline. PLoS Genet. 2014;10(2):e1004138. doi:10.1371/journal.pgen.1004138.
75. Mohn F, Handler D, Brennecke J. Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. Science. 2015;348(6236):812–7. doi:10.1126/science.aaa1039.
76. Han BW, Wang W, Li C, Weng Z, Zamore PD. Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. Science. 2015;348(6236):817–21. doi:10.1126/science.aaa1264.
77. Wen J, Mohammed J, Bortolamiol-Becet D, Tsai H, Robine N, Westholm JO, et al. Diversity of miRNAs, siRNAs, and piRNAs across 25 Drosophila cell lines. Genome Res. 2014;24(7):1236–50. doi:10.1101/gr.161554.113.
78. Biedler JK, Hu W, Tae H, Tu Z. Identification of early zygotic genes in the yellow fever mosquito *Aedes aegypti* and discovery of a motif involved in early zygotic genome activation. PLoS One. 2012;7(3):e33933. doi:10.1371/journal.pone.0033933.
79. Moshkovich N, Lei EP. HP1 recruitment in the absence of argonaute proteins in Drosophila. PLoS Genet. 2010;6(3):e1000880. doi:10.1371/journal.pgen.1000880.
80. Fagegaltier D, Bouge AL, Berry B, Poisot E, Sismeiro O, Coppee JY, et al. The endogenous siRNA pathway is involved in heterochromatin formation in Drosophila. Proc Natl Acad Sci USA. 2009;106(50):21258–63. doi:10.1073/pnas.0809208105.
81. Dufourt J, Pouchin P, Peyret P, Brasset E, Vaury C. NucBase, an easy to use read mapper for small RNAs. Mobile DNA. 2013;4(1):1. doi:10.1186/1759-8753-4-1.
82. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42(Database issue):D68–73. doi:10.1093/nar/gkt1181.
83. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40. doi:10.1093/bioinformatics/bti525.
84. Guberman JM, Ai J, Arnaiz O, Baran J, Blake A, Baldock R, et al. BioMart Central Portal: an open database network for the biological community. Database J Biol Databases Curation. 2011;2011:bar041. doi:10.1093/database/bar041.
85. Haider S, Ballester B, Smedley D, Zhang J, Rice P, Kasprzyk A. BioMart Central Portal–unified access to biological data. Nucleic Acids Res. 2009;37:W23–7.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central