Multitasking of the piRNA Silencing Machinery: Targeting Transposable Elements and Foreign Genes in the Bdelloid Rotifer *Adineta vaga*

Fernando Rodriguez and Irina R. Arkhipova[1]
Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, Woods Hole, Massachusetts 02543
ORCID IDs: 0000-0003-4044-8734 (F.R.); 0000-0002-4805-1339 (I.R.A.)

ABSTRACT RNA-mediated silencing processes play a key role in silencing of transposable elements, especially in the germ line, where piwi-interacting RNAs (piRNAs) are responsible for suppressing transposon mobility and maintaining genome integrity. We previously reported that the genome of *Adineta vaga*, the first sequenced representative of the phylum Rotifera (class Bdelloidea), is characterized by massive levels of horizontal gene transfer, by unusually low transposon content, and by highly diversified RNA-mediated silencing machinery. Here, we investigate genome-wide distribution of pi-like small RNAs, which in *A. vaga* are 25–31 nucleotides in length and have a strong 5′-uridine bias, while lacking ping-pong amplification signatures. In agreement with expectations, 71% of mapped reads corresponded to annotated transposons, with 93% of these reads being in the antisense orientation. Unexpectedly, a significant fraction of piRNAs originate from predicted coding regions corresponding to genes of putatively foreign origin. The distribution of piRNAs across foreign genes is not biased toward 3′-UTRs, instead resembling transposons in uniform distribution pattern throughout the gene body, and in predominantly antisense orientation. We also find that genes with small RNA coverage, including a number of genes of metazoan origin, are characterized by higher occurrence of telomeric repeats in the surrounding genomic regions, and by higher density of transposons in the vicinity, which have the potential to promote antisense transcription. Our findings highlight the complex interplay between RNA-based silencing processes and acquisition of genes at the genome periphery, which can result either in their loss or eventual domestication and integration into the host genome.

KEYWORDS piwi-interacting RNAs; horizontal gene transfer; mobile genetic elements; telomeres

BDELLOID rotifers are common invertebrate animals a few tenths of a millimeter long, primarily found in fresh water, temporary pools, and other ephemerally aquatic habitats (Ricci and Melone 2000). Class Bdelloidea is classified in four families and 19 genera, with >460 described species (Segers 2007) in which there is no evidence of sex or meiosis in their life cycle, suggesting completely asexual reproduction (Mark Welch *et al.* 2008; Flot *et al.* 2013). The class, characterized by their ciliated head and bilateral ovaries, is known to have an ancient origin: bdelloid remains have been identified in 35- to 40-million-year-old amber, and the substantial number of synonymous site substitutions indicates that different bdelloid families have diverged tens of millions of years ago (Poinar and Ricci 1992; Mark Welch *et al.* 2008). They are able to survive in ephemerally aquatic environments due to their ability to withstand desiccation and subsequently resume reproduction at any stage of their life cycle. Bdelloids can also resist high doses of ionizing radiation, probably as a result of adaptation to their desiccation-prone life style. A dose as high as 500 Gy can cause DNA double strand breaks throughout the entire genome, and they are still able to survive and to resume reproduction after such exposure (Gladyshev and Meselson 2008).

The 244-Mb draft genome of the bdelloid rotifer *Adineta vaga* (Flot *et al.* 2013) is characterized by degenerate tetraploid structure: intragenomic sequence comparisons show that many genes are located in collinear regions or homologous blocks with conserved gene order, consisting of pairs of

doi: 10.1534/genetics.116.186734
Manuscript received January 3, 2016; accepted for publication March 21, 2016; published Early Online March 25, 2016.
Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.186734/-/DC1.
[1]Corresponding author: Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543. E-mail: iarkhipova@mbl.edu.

highly similar allelic regions which can be further grouped into quartets composed of more divergent anciently duplicated segments. However, the putatively allelic regions are often found on the same genomic scaffold in direct or inverted orientation, a structure incompatible with proper pairing of chromosomes and conventional meiosis. Together with the presence of nearly all meiosis-specific genes (Flot *et al.* 2013), these findings led to a recent hypothesis that unconventional meiosis may exist in bdelloids, which does not require pairing along most of the chromosome length, but invokes pairing at telomeres to explain allele sharing in natural populations (Signorovitch *et al.* 2015). The *A. vaga* genome is also characterized by elevated frequency of gene conversion, which is manifested in excessive identity track lengths for resolved allelic regions, and in double coverage of ~10% of genomic scaffolds, indicating that some nearly identical regions were not resolved during assembly (Flot *et al.* 2013).

Theory predicts that asexual organisms may experience an increase of transposable element (TE) content in their genomes, which could potentially drive the population to extinction, unless their proliferation is brought under control (Hickey 1982; Arkhipova and Meselson 2000, 2005a). Indeed, we found that only 3% of the *A. vaga* genome was composed of TEs (Flot *et al.* 2013). Although the diversity of TE families in *A. vaga* is very high, with 255 reported families, most of them are present in very low copy numbers (normally with one or two full-length copies). Genomic distribution of TEs in *A. vaga*, and apparently in bdelloid rotifers in general, is strongly biased toward telomeric regions, and TEs are nearly excluded from the core gene-rich regions. The relatively low degree of TE divergence, their subtelomeric location, and low abundance of decayed TEs suggest that many of them represent recent arrivals acquired by horizontal gene transfer (HGT). TE limitation and decay may be explained, at least in part, by recombinational excision (LTR–LTR recombination) and by microhomology-mediated deletion, molecular footprints of which are evident in many fragmented TE copies (Gladyshev and Arkhipova 2010; Flot *et al.* 2013). Furthermore, we found that the genome is characterized by a high degree of expansion and diversification of the RNA-mediated silencing machinery, including Dicers, Argonaute/Piwi family proteins, and RNA-dependent RNA polymerases (RdRP) (Flot *et al.* 2013).

Along with other components, these protein families play an important role in genome surveillance through small RNA (sRNA)-guided transcriptional and post-transcriptional mechanisms. In the piRNA pathway, piwi-interacting RNAs (piRNAs) interact with Piwi proteins, members of the Argonaute protein family, which have a structural compartment that accommodates a small RNA molecule with sequence homology sufficient for base pairing with longer complementary RNAs. The Piwi-piRNA system controls TEs in the germ line of diverse metazoans (Vagin *et al.* 2006; Aravin *et al.* 2007a,b; Brennecke *et al.* 2007; Siomi *et al.* 2008, 2011). Typically, piRNAs are 24–31 nucleotides (nt) in length, and are initially derived from long RNA precursor transcripts that are predominantly antisense to TEs (Brennecke *et al.* 2007; Gunawardane *et al.* 2007; Siomi *et al.* 2011). While biogenesis of microRNAs (miRNAs) and small interfering RNAs (siRNAs) from double strand RNA (dsRNA) precursors depends on the type III ribonuclease Dicer, piRNA biogenesis is Dicer-independent. Different mutants in the piRNA pathway can be associated with different types of derepressed TEs (Vagin *et al.* 2006; Chen *et al.* 2007; Chambeyron *et al.* 2008).

A high rate of HGT has been reported in bdelloid rotifers, with multiple genes apparently having originated from bacteria, fungi, and plants (Gladyshev *et al.* 2008). This has been attributed to the desiccation-prone lifestyle, which may result in transient damage to DNA and cell membranes, occasionally facilitating the uptake and incorporation of foreign genetic material. Over 8% of the genes annotated in the *A. vaga* genome are much more similar to their nonmetazoan counterparts than to metazoan ones, with the timing of HGT events ranging from ancient to recent (Flot *et al.* 2013). Approximately 10% of expressed genes in the congeneric *A. ricciae* are of foreign origin (Boschetti *et al.* 2012). In *A. vaga* and other studied bdelloids, foreign genes tend to accumulate in subtelomeric regions of the genome, along with TEs, which likewise could have been gained by HGT (Gladyshev *et al.* 2008). In this study, we sought to investigate whether foreign genes obtained by HGT, independently or in association with TEs, could trigger the host surveillance response from the small RNA silencing pathways used to preserve genome integrity, and, if so, to what extent the piRNA machinery can target such genes *per se*, or possibly represent an extension of TE silencing in subtelomeric regions. Such processes could play a role in adaptation of foreign genes to the host genomic environment, adjustment of their transcriptional activity, and eventual incorporation into diverse metabolic pathways.

Materials and Methods

Rotifer cultures

Clonal cultures of *A. vaga*, once started from a single individual, were maintained continuously in filtered spring water and fed with *Escherichia coli*. Rotifers were grown in 150 × 25 mm Petri dishes and transferred into new ones, until the desired biomass was reached.

Small RNA library preparation and sequencing

Initial small RNA libraries were prepared by total RNA extraction from two biological samples: nontreated or wild type (WT) and exposed to ionizing radiation (IR). Rotifers cultured in four 10-cm Petri dishes were transferred to a clean one by pipetting and transported to the irradiation facility. The dish was placed on a block of ice underneath a ^{137}Cs source delivering a total dose of 500 Gy. After receiving the dose, the biomass was collected by low-speed centrifugation and snap frozen with liquid nitrogen. Total RNA extracted with Trizol

was gel purified in15% urea-PAGE, and the 18- to 35-nt area was selected for cloning. Libraries were barcoded, pooled, and sequenced with 35 short-read (SR) cycles on Illumina GAIIx.

For isolation of protein-bound small RNAs, ~1 g of centrifuged rotifers were snap frozen in liquid nitrogen and homogenized in a tissue grinder with a glass pestle in the binding buffer (20 mM Hepes pH 7.9, 10% glycerol, 100 mM KOAc, 0.2 mM EDTA, 1.5 mM $MgCl_2$) with 1× protease inhibitors (Roche). After centrifugation, the supernatant was saved for subsequent chromatography. Protein lysates were fractionated on small-scale ion-exchange Q columns (HiTrap Q HP 1 ml, GE) following the protocol of Lau *et al.* (2006). Small RNA–protein complexes were eluted in mild salt solution (0.1 to 1 M potassium acetate gradient).

Different chromatography fractions were deproteinized by acid phenol-chloroform extraction, and the resulting RNAs where precipitated with ethanol. To monitor the content of each eluate fraction, RNAs were 3′-end labeled with ^{32}P-cordycepin (Perkin-Elmer) and *E. coli* poly(A)-polymerase I (Thermo Scientific) and checked by denaturing polyacrylamide gel electrophoresis. Expected piRNA fraction was ligated to adaptors: the 3′ ligation was done with T4 RNA ligase 2 truncated (NEB) to minimize the effects of 2′-O-methyl modification of small RNA populations (Munafó and Robb 2010), and the 5′ adaptor ligation was done with T4 RNA ligase 1 (NEB). Final cDNA libraries were constructed with custom oligonucleotides, using the proprietary sequences for small RNA library construction provided by Illumina upon request, and sequenced on the Illumina HiSeq platform (50-bp SR). Two biological replicates per sample were run with indexed barcodes.

Computational analysis

Demultiplexing of reads was done using Casava 1.8.2 (Illumina). The 3′ adaptors were trimmed (cutadapt, Martin 2011), as well as any sequence with low quality score (Q33) and/or 16 nucleotides in length (FASTX Toolkit). Final extracted reads were aligned using Bowtie (Langmead *et al.* 2009) to the *A. vaga* reference genome (Flot *et al.* 2013). Downstream processing of the reads mapped by Bowtie was performed using a variety of available software and custom Linux scripts. Aligned sequence reads were counted by genomic feature with htseq-count (Anders *et al.* 2014), and GBrowse (Donlin 2007) was used for genome and alignment visualization. Genomic coordinates of gene models were downloaded from the *A. vaga* Genoscope genome browser (http://www.genoscope.cns.fr/adineta/cgi-bin/gbrowse/adineta), along with scaffold sequences.

Transposable element annotation

The dataset of known *A. vaga* TE families (Flot *et al.* 2013) was used as database/library for searching and annotating them in the genome using RepeatMasker (Smit *et al.* 2013). We used RMBlast (National Center for Biotechnology Information (NCBI) Blast modified for use with RepeatMasker) as search engine. Initial TE content in the RepeatMasker output was treated with the perl script one_code_to_find_them_all.pl (Bailly-Bechet *et al.* 2014), previously modified to incorporate *A. vaga* TE families. The output was converted into gff3 format for subsequent analysis. The resulting annotation was parsed along with previous gene prediction models to eliminate any duplication events spanning both databases.

Gene classification

The gene set was taken from the existing annotation of the *A. vaga* genome in the supplementary data file 5 of Flot *et al.* (2013), in which the "foreignness" of each CDS (coding sequence) is measured by the Alien Index (AI) log metric, based on *e*-values of similarity to best metazoan *vs.* nonmetazoan blastp hits (Gladyshev *et al.* 2008). AI can take values between −460 and +460, where −30 and +30 were used as thresholds in current and previous analyses as follows: genes with AI ≤ -30 can be considered of probable metazoan origin; genes with $-30 < \text{AI} < 30$ of uncertain origin; and genes with AI ≥ 30 of probable nonmetazoan origin. For downstream analysis, the "uncertain origin" gene category was additionally split into two, $0 \leq \text{AI} < 30$ (more likely of nonmetazoan origin) and $-30 < \text{AI} < 0$ (more likely of metazoan origin). While the latter subdivision cannot be considered definitive, since the interval $0 \leq \text{AI} < 30$ includes some host genes and the interval $-30 < \text{AI} < 0$ includes some foreign genes, it nevertheless allows for continuity of comparison across the entire AI range without disregarding a large number of genes in the gene set. Annotated CDS without significant blast hits, which could not be assigned to any category, were not included in the analysis. Supplemental material, File S1 (sheet 1) lists sRNA counts for GeneIDs sorted by AI.

Phylogenetic analysis of foreign gene families

The nonmetazoan origin of selected foreign gene families not displaying modular or repetitive structure was verified by phylogenetic analysis. Protein multiple sequence alignments (MUSCLE, Edgar 2004) were converted into codon-based DNA alignments with PAL2NAL (Suyama *et al.* 2006). Phylogenetic trees from each alignment were built by PhyML (Guindon and Gascuel 2003). Final trees were edited with FigTree (A. Rambaut, http://tree.bio.ed.ac.uk/software/figtree/).

Analysis of metazoan gene families

Genes of likely metazoan origin ($\text{AI} \leq -30$) were analyzed for small RNA profiles over their sequence. The following conditions were imposed to select genes with non-UTR small RNA profiles: to have at least 10 reverse counts in the CDS regions, but not >10 reverse counts spanning the UTRs. From the initial extracted dataset (327 gene IDs), the ortholog analysis was performed to verify their metazoan origin. For this purpose, we compared them against the OrthoMCL database (http://www.orthomcl.org, Fischer *et al.* 2011) considering only the phyletic groups consisting of metazoan species. From the initial 203 phyletic groups, 48 groups were retained (74 gene IDs).

To look for gene family expansions in host-associated genes with small RNA signal, the initial dataset of 327 genes was compared against the *A. vaga* transcriptome using an initial blast "all *vs.* all" search, followed by ortholog clustering analysis with OrthoMCL 1.4 (Li *et al.* 2003). From different orthogroups, those incorporating a significant number of the initial 327 host-related genes and showing a potential family expansion were retrieved. The ortholog group number 16, containing sequences homologous to cytochrome p450 (CYP), was retained for phylogenetic analysis as above. CYP sequences from different organisms were extracted from the Cytochrome P450 Homepage and assigned to families and superfamilies according to the Nelson classification scheme (Nelson 2006, 2009).

Data availability

Sequences obtained in this study have been deposited at the NCBI SRA database under accession no. SRP070765.

Results

Small RNA library construction and sequencing

We initially constructed small RNA libraries using size selection (gel excision) of total RNA extracted from fresh *A. vaga* biomass. One of the samples was irradiated to check whether the small RNA populations would be affected under conditions causing significant DNA damage. Sequencing on the Illumina GA IIx platform, after filtering to remove adaptors and low-quality sequences, yielded ∼9.6 M and ∼10.3 M sequences (a total of ∼20-M reads) for the wild-type (WT) and IR samples, respectively.

We observed a high level of *A. vaga* ribosomal RNA representation in our libraries, especially in the IR sample [15% ribosomal RNA (rRNA) in WT and 62% rRNA in IR], which is likely due to overabundance of rRNA breakdown products after irradiation, as was previously observed by others (Lee *et al.* 2009). In addition, size-selected libraries also contained transfer RNA (tRNA) breakdown products. After filtering of rRNA and tRNA sequences, the combined dataset was reduced to ∼6.6-M reads, of which ∼3 M could be mapped to the reference genome.

The patterns of length distribution in small RNA populations were found to be different in comparisons between WT and IR samples, especially in the 22- to 24-nt fraction, which was significantly increased in the IR sample, as was previously observed in other irradiated species, where siRNAs were shown to be involved in DNA repair (Wei *et al.* 2012). Of particular interest, however, was the pronounced shoulder formed by longer RNA species, which we inferred to correspond to piRNAs (Figure S1A). Typically, piRNAs are 25–30 nt in length and display a strong 5′-uridine bias (Vagin *et al.* 2006; Aravin *et al.* 2007a,b; Brennecke *et al.* 2007; Gunawardane *et al.* 2007; Siomi *et al.* 2008, 2011; Iwasaki *et al.* 2015). In agreement with expectations, the initial annotations met the correlation between TEs and potential piRNAs, in accordance with the role of piRNAs as central players in transposon silencing. Indeed, the 25- to 31-nt shoulder (∼1.3-M reads) exhibited a strong 5′-end uridine bias, especially in the 29-nt area of the IR sample (Figure S1B). However, the overwhelming majority of reads was represented by a large peak composed of siRNAs and miRNAs (Figure S1A), and it was difficult to achieve full separation of the presumed piRNA fraction in the shoulder to perform meaningful analyses. In addition, TE-mapped reads displayed highly nonuniform distribution across TE length (Figure S1D), which was likely due to low representation of piRNAs in the samples dominated by rRNAs and other types of small RNAs and the resulting amplification bias.

While our initial experiments agreed with the involvement of small RNAs in repair of DNA damage (Francia *et al.* 2012; Wei *et al.* 2012), we were particularly interested in examining the population of piRNAs, which suppresses TE activity in the germ line and may act to establish *trans*-generational silencing. To investigate piRNAs in *A. vaga*, we had to select a different method that would result in better representation of the piRNA population, since the commercially available ribo-depletion kits proved ineffective in this species. In model organisms, piRNAs can be purified through their association with Piwi proteins following immunoprecipitation with the corresponding antibodies. However, the sheer number of Piwi variants in *A. vaga* (eight different proteins) does not make it feasible to employ anti-Piwi antibodies for piRNA purification. We therefore decided to use Hitrap-Q chromatography columns for isolation of protein-bound small RNAs and simultaneous elimination of rRNA/tRNA, which results in significant enrichment with Argonaute complex proteins and small RNAs bound to them (Lau *et al.* 2006). The Illumina HiSeq library construction and sequencing of protein-bound sRNA molecules resulted in 21.7 million reads (two wild-type replicas). By using a mapping procedure to report only the alignments in the best alignment stratum (*i.e.*, those having the least number of mismatches) and choosing the unique mapping option (one read — one mapped location), we matched a total of ∼19.3-M sequencing reads (91.13%) to the *A. vaga* reference genome. Since this method did not yield any differences between WT and IR samples, presumably due to near saturation of protein complexes by preexisting small RNAs, we proceeded with the analysis of WT samples to obtain a comprehensive picture of piRNA distribution in the *A. vaga* genome.

Transposable elements and small RNAs

The length distribution plot of the reads mapped to the *A. vaga* reference genome shows that they are represented mostly by 25- to 32-nt-long sRNAs, peaking at 29 nt (Figure 1A). Sequences longer than 24 nt exhibit a strong 5′-U bias (*e.g.*, 96% of 29-nt reads have uridine at the 5′-end; Figure 1B). After filtering reads not assigned to any annotated features such as TEs or gene models, small RNAs showed a strong preference toward annotated transposons, and the majority of reads were mapped to TEs in reverse (antisense) orientation: while 71% of all reads were derived from TEs (sense and

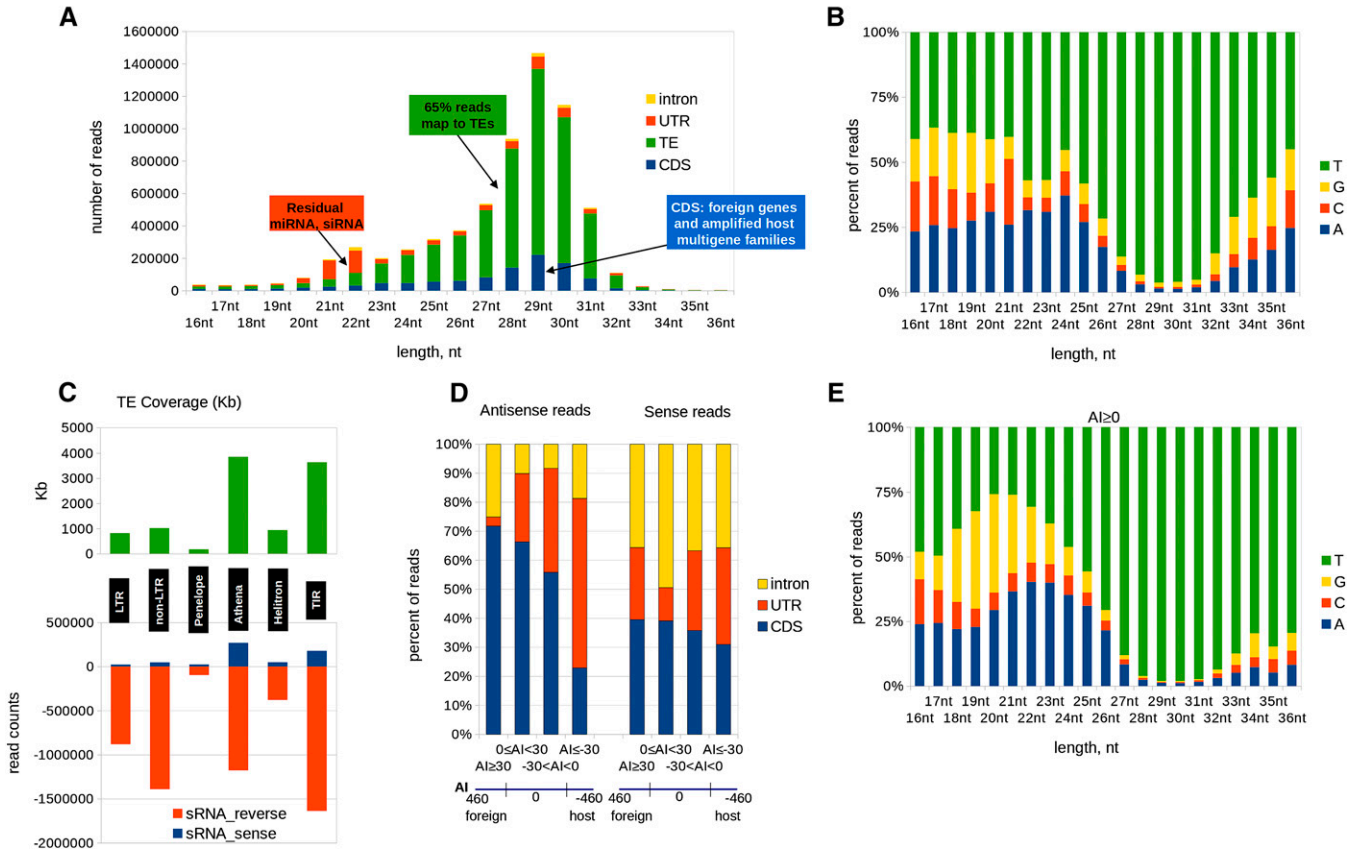


Figure 1 Characterization of protein-bound small RNAs in *A. vaga*. (A) Length distribution of sRNAs mapped to annotated genomic features (TEs, CDS, UTRs, and introns). The *y*-axis shows combined read counts for both replicates. (B) Percentage of 5′-terminal nucleotides (A, T, G, and C) in each sRNA size class. (C) The share of the assembly (in kilobases) occupied by major TE groups (top) and distribution of sRNA reads between TE groups (bottom) in the sense (blue) or antisense (red) orientation. Crypton TEs are not shown due to insignificant contribution to both plots. (D) Distribution of small RNAs across the AI gene categories: annotated CDS (blue), UTR (orange), and intron (yellow) features in the antisense (left) and sense (right) orientation. (E) The 5′-terminal nucleotide composition of small RNAs mapped to *A. vaga* genes of putatively foreign origin (AI ≥ 0); nucleotides are colored as in B.

antisense strands combined), 66% of all reads were in the antisense direction. Since these pi-like RNAs meet the typical piRNA characteristics (25- to 32-nt size range, peaking at 29 nt; strong 5′-U bias; most sequences are TE derived), we further consider these sequences as piRNAs, although technically they were not isolated using anti-Piwi antibodies. However, to keep in mind that a small number of the studied sRNA population may represent noncoding small RNAs other than piRNAs, we collectively designate them as sRNA, even though the vast majority fits the piRNA category.

It was of interest to investigate the piRNA population for signatures of possible amplification mechanisms. One of the models of piRNA biogenesis, which applies to *Drosophila* and many other species, proposes that secondary piRNAs are generated through iterative Piwi-mediated cleavage of transcripts with complementary sequence, known as the ping-pong amplification cycle (Brennecke *et al.* 2007; Gunawardane *et al.* 2007). We investigated the overlap (in nucleotides) between stranded and reverse reads corresponding to TE annotations. According to the ping-pong model, piRNAs that map to opposite genomic strands tend to have a constant overlap (*e.g.*, 10 nt), in which Argonaute proteins bound to sense-strand piRNAs catalyze antisense-strand cleavage at the tenth-nucleotide base pair that generates the 5′-end of antisense piRNAs. However, in a plot of overlap values for TEs, there is no obvious ping-pong amplification signature at any overlap size (Figure S2A). This result, together with the lack of internal nucleotide bias in sense reads (Figure S2C), leaves open the question of how secondary piRNAs are generated and amplified in *A. vaga*, with numerous RdRPs (20 copies; Flot *et al.* 2013) presumably playing a prominent role.

The initial TE annotation in *A. vaga* covered ∼3% of the genome (Flot *et al.* 2013). Even after TE reannotation, which included previously unknown TE types (I. R. Arkhipova, F. Rodriguez, and I. A. Yushenova, unpublished results) and used the highly sensitive RepeatMasker tool yielding a number of false positives, TE annotations still cover only 4.4% of the genome, which is much less than TE content reported for other metazoans (Arkhipova and Rodriguez 2013). The top panel in Figure 1C depicts the share of the genome (in kilobases) occupied by each of the major TE subclasses. We also reported the high diversity of TE families, which are represented by only one or two full-length copies (209 of 255 families) and about 10 times as many

fragmented copies (Flot *et al.* 2013). Figure S3A illustrates that the low ratio of full-length copies to partial fragments is more or less maintained in each TE subclass.

Small RNAs cover 70.2% of cumulative annotated TE length (not considering different coverage depths or strandedness), thus showing a significant improvement over the size-selected library prepared from total RNA (4.9%). For the vast majority of active TE copies, sRNA coverage is sufficient to target them for silencing, since a TE does not have to be fully covered by sRNAs over its entire length to be subjected to silencing. Those copies that do not yield significant sRNA coverage (<10 counts) usually belong to transcriptionally inactive and decaying families. The distribution of sRNA reads is rather uniform over the entire TE length, not displaying any bias toward specific TE regions such as coding or repeated sequences. While sRNA mapping in most cases agrees with preexisting TE annotations, in some cases we were able to extend host–TE boundaries for previously unknown TE types by relying on small RNA coverage (El Baidouri *et al.* 2015; I. R. Arkhipova, F. Rodriguez, and I. A. Yushenova, unpublished results). Nevertheless, piRNAs show certain differences in numbers per class or superfamily. The bottom part of Figure 1C shows the distribution of small RNAs, in sense or reverse orientation, across the major *A. vaga* TE subclasses. The terminal inverted repeat (TIR)-containing and Athena elements yield a substantial share of reverse sRNAs (1.63 M and 1.18 M reads, respectively), which is expected because both subclasses occupy a bigger proportion of the genome than the others, as may be seen in the upper panel. However, LTR and non-LTR retrotransposons, which show relatively moderate genome coverage when compared to TIR or Athena, also comprise a substantial share of reverse reads (0.88 M and 1.39 M, respectively), which is indicative of higher density of sRNA coverage in these retrotransposons or perhaps more frequent collapse of several copies into a single contig.

To assess the transcriptional activity of TEs, we used transcriptome profiling. To this end, 76-bp Illumina reads from cDNA libraries were mapped to the *A. vaga* genome as described in Flot *et al.* (2013) and used to compare distributions of aligned RNA-seq reads and sRNA reads in annotated TE families. Figure S4 shows RNA-seq and sRNA count plots for previously described individual *A. vaga* retrotransposon (LTR, non-LTR, Penelope) and DNA TE (TIR and Helitron) families (Arkhipova and Meselson 2005b; Gladyshev and Arkhipova 2010; Arkhipova *et al.* 2013; Flot *et al.* 2013). In most cases, TE families with higher antisense sRNA counts also display lower RNA-seq counts, with a few exceptions. For instance, the high transcript levels in Zator1 are confined to the N-terminal part of the single ORF containing an extended coiled-coil motif and yield a truncated inactive transposase, which may potentially act as a repressor (Majumdar and Rio 2015). In other cases where TEs represent recent arrivals and may be undergoing an initial increase, such as Hebe or Tc/mariners (Arkhipova and Meselson 2005b; Gladyshev and Arkhipova 2010), the distribution of RNA-seq counts is biased toward several copies, which may have been collapsed together. A curious exception is the rotifer-specific Soliton clade of non-LTR retrotransposons: several Soliton families display atypically high RNA-seq/sRNA count ratios (Figure S4B), but are nevertheless characterized by particularly low copy numbers and by virtual absence of fragmented copies. Apparently, for these TEs, higher transcriptional activity is not correlated with higher proliferative capacity, perhaps being limited to a certain somatic tissue.

Small RNAs and genes of foreign origin

In the sRNA population that is not derived from TEs, a significant fraction (30%) originates from predicted CDS (Figure 1A), and 26% of all non-TE hits matched genes with AI > 0 (Figure S3B), many of which may have been acquired by HGT. Characteristically, sRNA coverage of such genes generally exhibits patterns that are similar to those observed in TEs. Specifically, 81% of sRNAs mapping to putatively foreign genes are represented by 25- to 32-nt reads, with 29-nt reads being the most abundant (either in reverse or sense orientation), and >79% of reads in the 25- to 32-nt size range are mapped to annotated CDS regions.

To characterize small RNAs matching foreign genes, we associated each GeneID with the AI value, distinguishing between reads corresponding to sense (stranded) direction or antisense (reverse) direction in each gene feature (UTRs, introns, and CDS). Small RNAs from the reverse strand of transcripts comprise the majority of reads (77%) in AI ≥ 0 genes, for which CDS features are precisely targeted (67%) by antisense reads. For sense reads in the 25- to 32-nt size range, 39% targeted CDS features, while a noticeable input was provided by intron regions (41%).

We also observed that small RNAs associated with foreign genes in *A. vaga* have a strong bias for 5′-U (Figure 1E), which is the terminal nucleotide in 87% of reads in the 25- to 32-nt size range. Again, as in pi-like small RNAs mapping to TEs, there is no obvious 5′-nucleotide bias in sense reads (data not shown). Nor could we observe a ping-pong amplification signature with a constant overlap between reverse and sense reads (Figure S2B).

To analyze sRNAs associated with the gene dataset, we divided it into four categories: AI ≥ 30; 0 ≤ AI < 30; −30 < AI < 0; and AI ≤ −30 (Figure S3, B and C; *Materials and Methods*). All sRNAs mapped to annotated genes were subdivided into UTR-, intron-, or CDS-matching reads. At the same time, a distinction was set based on read orientation (sense or antisense). Notably, while the *bona fide* foreign genes (AI ≥ 30) account for ∼10% of genes with BLAST hits (Figure S3B), they constitute nearly a quarter of all genes with sRNA coverage. The proportion of genes with piRNAs mapping to CDS for each AI category is shown in Figure S3C, where the number of genes with piRNA coverage (from two biological replicates with a minimum count of 10 per gene) is plotted for each AI group, showing that the proportion of CDS-matching antisense reads in AI ≥ 30 genes does not depend on a few genes with high sRNA count.

Notably, Figure 1D shows that across the AI categories, *i.e.*, from nonmetazoan to host genes, a transition is clearly observed in antisense reads, featuring a shift from CDS (over 70% in foreign AI ≥ 30) to UTR ($\sim$60% in metazoan AI ≤ -30). Thus, while our results for metazoan genes agree with previous observations that genic piRNAs are mostly 3′-UTR directed (Robine *et al.* 2009; Saito *et al.* 2009; Siomi *et al.* 2011; Le Thomas *et al.* 2014), the putatively foreign genes largely display piRNA coverage patterns that are similar to TEs in uniformly targeting the gene body, rather than the 3′ UTR.

Overall, we find that the patterns of piRNAs matched to numerous alien gene families resemble the patterns derived from TEs in read distribution and orientation, indicating that the RNA-mediated silencing machinery may be acting to regulate expression of putatively foreign genes.

Small RNA profiles in foreign multigene families

To further understand the interactions between sRNA machinery and foreign genes and to place them into phylogenetic context, we sought to investigate the immediate genomic environments and phylogenetic relationships in sufficiently expanded gene families and their nonmetazoan counterparts and to compare RNA profiles (RNA-seq and sRNA) in each group. Analysis of multigene family members in the phylogenetic context is particularly informative, since it allows one to discriminate between relatively recent insertion/duplication events and more ancient gene acquisitions residing in the host genome for a long time and to determine whether the more recent copies are more likely to be subjected to piRNA response. The presence of repetitive and/or modular structures in many expanded gene families complicates their proper taxonomic assignment and phylogenetic analysis, and therefore such families could not be considered in the phylogenetic context. Here, we focus on two foreign multigene families devoid of repetitive and modular structures: β-lactamases (*bla*) and NAD:arginine ADP-ribosyltransferases (ART).

Serine β-lactamases are known to provide resistance to β-lactam antibiotics such as penicillin or cephalosporin (Ambler 1980). They are found in the majority of principal bacterial groups or taxonomic subdivisions. We extracted all *A. vaga* CDS homologous to serine β-lactamases (36 copies, with AI between 22 and 304; File S1, sheet 2) and subjected them to phylogenetic analysis (Figure 2A), including representatives of known bacterial β-lactamase classes (Hall and Barlow 2004). Each node giving rise to an *A. vaga* β-lactamase clade that also includes a bacterial counterpart with significant clade support value apparently corresponds to an independent horizontal transfer, yielding at least nine such events (and possibly more; see below).

Not every copy of β-lactamase (*bla*) in *A. vaga* exhibits the same RNA profile, either in sRNA or RNA-seq profiles. Figure 2B provides a dendrogram representation of the clade bracketed in the β-lactamase tree of Figure 2A. Nucleotide sequence identity is shown for nearest neighbors, which can be grouped into allelic pairs after checking for collinearity of adjacent genes. The sRNA profiles are clearly observed in two groups in the left part of the chart, represented by seven gene models that form the most recent branches in the phylogenetic tree. Each of the two groups contains allelic pairs with high nucleotide sequence identity (>98%) located in collinear gene environments and an additional duplicated copy. The additional annotated *bla* copies in each group may represent assembly variants due to polymorphic TE insertions; however, they still draw sRNA and RNA-seq counts regardless of their activity status. GSADVT00009239001 does not share the local environment with three other members of the clade and does not have an allelic partner, but its coverage is approximately double in comparison with other members of this clade, and could therefore represent an allelic pair formed by gene conversion (Flot *et al.* 2013). The sRNA and RNA-seq profiles from each member of a pair display similar counts, indicating that identical reads are randomly assigned to either of the high-identity copies.

Importantly, all *bla* clades with sRNA coverage are likely to represent relatively recent insertions into subtelomeric regions, as may be judged from their phylogenetic placement and the presence of short stretches of telomeric repeats surrounding these *bla* copies (Figure 2C). The immediate genomic environment of *bla* genes in each group is different, indicating independent origin of each segment framed by telomeric repeats. Telomeric repeat signatures are not seen in the more basal branch, perhaps because of gradual erosion over time (Figure 2C, middle scaffold 32), and all of the earlier-branching clades, arranged in allelic pairs with syntenic environments (Figure 2B, right half), lack interstitial telomeric repeat stretches and do not display sRNA coverage, which may indicate full incorporation of these genes into the host environment.

Another expanded foreign multigene family of relatively simple structure is ART. While its bacterial counterparts are primarily annotated as uracil-DNA glycosylases (UDG), most of these genes in *A. vaga* lack the UDG moiety, and contain only the C-terminal ADP-ribosylating domain of certain bacterial UDG-like enzymes, which is most similar to the VIP2 family of protist and bacterial actin-ADP-ribosylating insecticidal binary toxins. Of 156 *A. vaga* genes of probable nonmetazoan origin (AI > 0) with ART domain, 69 copies formed an ortholog group. As in β-lactamases, a variety of RNA profiles is observed in *A. vaga* ART homologs (Figure 3B). Two of these show a particularly high sRNA peak. After inspection of their genomic context, they can be considered as an allelic pair with collinearity in the region of scaffold overlap (Figure 3A) and, being 95% identical, share most of the RNA reads. Both have an oppositely oriented insertion of the same TIR fragment (*Avmar1a*) at their 3′ flanks, which may have played a role in establishing piRNA production from this gene. In addition, GSADVT00063672001 has a unique *P*-element insertion at the 5′-flank, which, however, shows little piRNA coverage. The decaying pseudogene retained in one of the haplotypes has the footprints of microhomology-mediated deletion and is likely on the way out. Again, the presence of

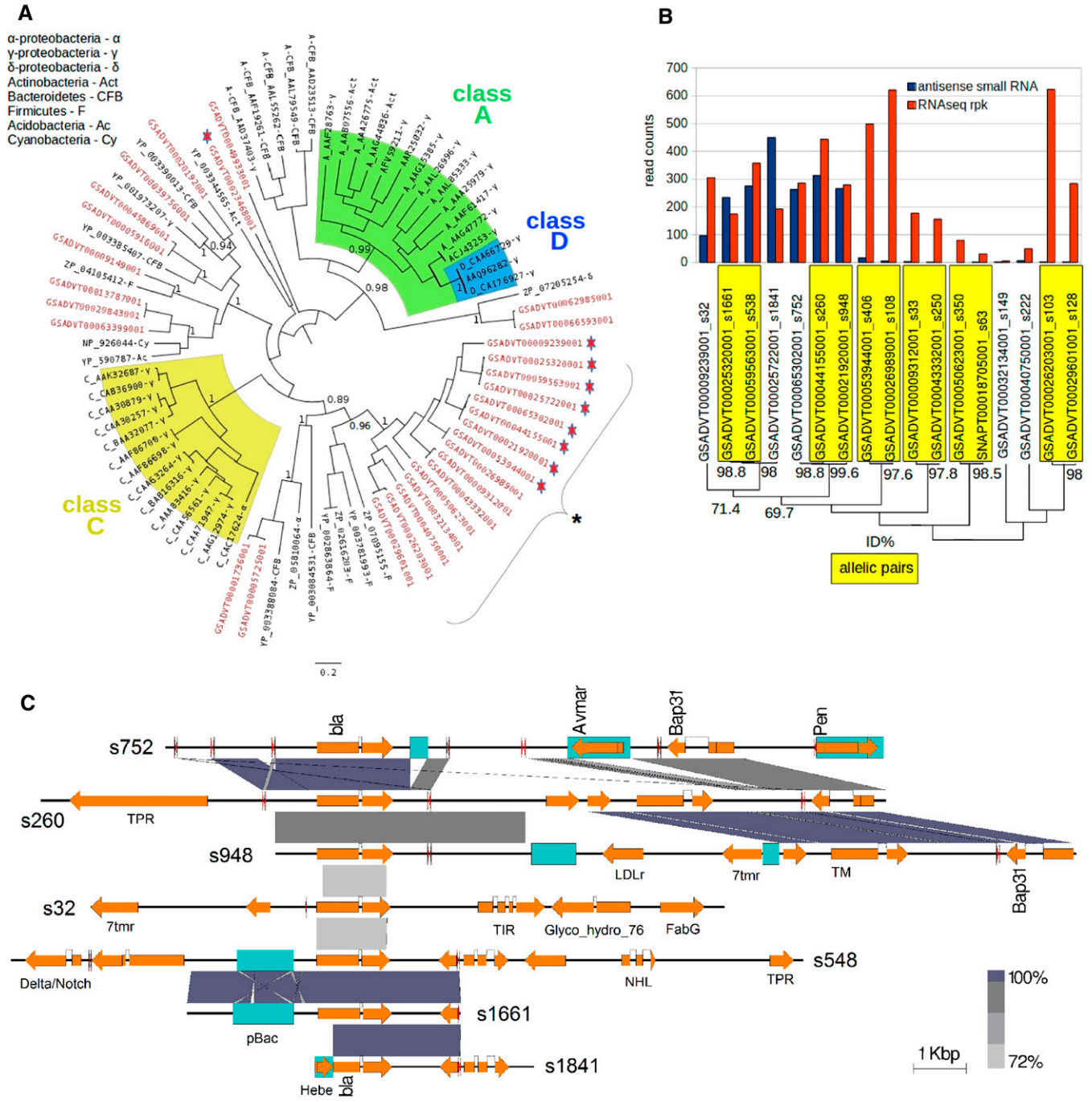


Figure 2 Small RNAs in the *A. vaga* β-lactamase family. (A) Phylogenetic relationships of *A. vaga* serine β-lactamase homologs and their bacterial counterparts. The codon-based maximum likelihood phylogenetic tree of serine β-lactamases includes bacterial classes A, C, and D (in black) and the *A. vaga* homologs (in red). Red stars denote *A. vaga* genes with small RNA coverage. The node with an asterisk (*) is expanded in B to show RNA profiles. The key for bacterial subdivisions is shown in the top left corner. (B) Dendrogram and RNA profile histogram of *A. vaga* β-lactamase homologs. The top histogram shows antisense sRNA counts and RNA-seq counts for the clade in A marked with *. The dendrogram below shows phylogenetic relationships of bla genes and nucleotide sequence identity (%) between nearest neighbors, which can be grouped into allelic pairs (yellow boxes) by synteny analysis. The *ab initio* gene prediction model SNAPT00018705001 was included as a member of the allelic pair. (C) Genomic environments of *A. vaga bla* genes with sRNA coverage. The alignment is centered on the seven leftmost *bla* genes from B, organized in two high-identity groups, with the unpaired scaffold 32 in the middle. Predicted genes, yellow block arrows; TE insertions, light blue bars; and telomeric repeats, small red arrowheads. Regions of homology are gray shaded, and the intensity of shading corresponds to nucleotide sequence identity (%) as indicated. BLASTN output was used for pairwise comparison of scaffolds, plotting similarities with *e*-values <0.001. The figure was produced using Easyfig 2.2.2 (Sullivan *et al.* 2011).

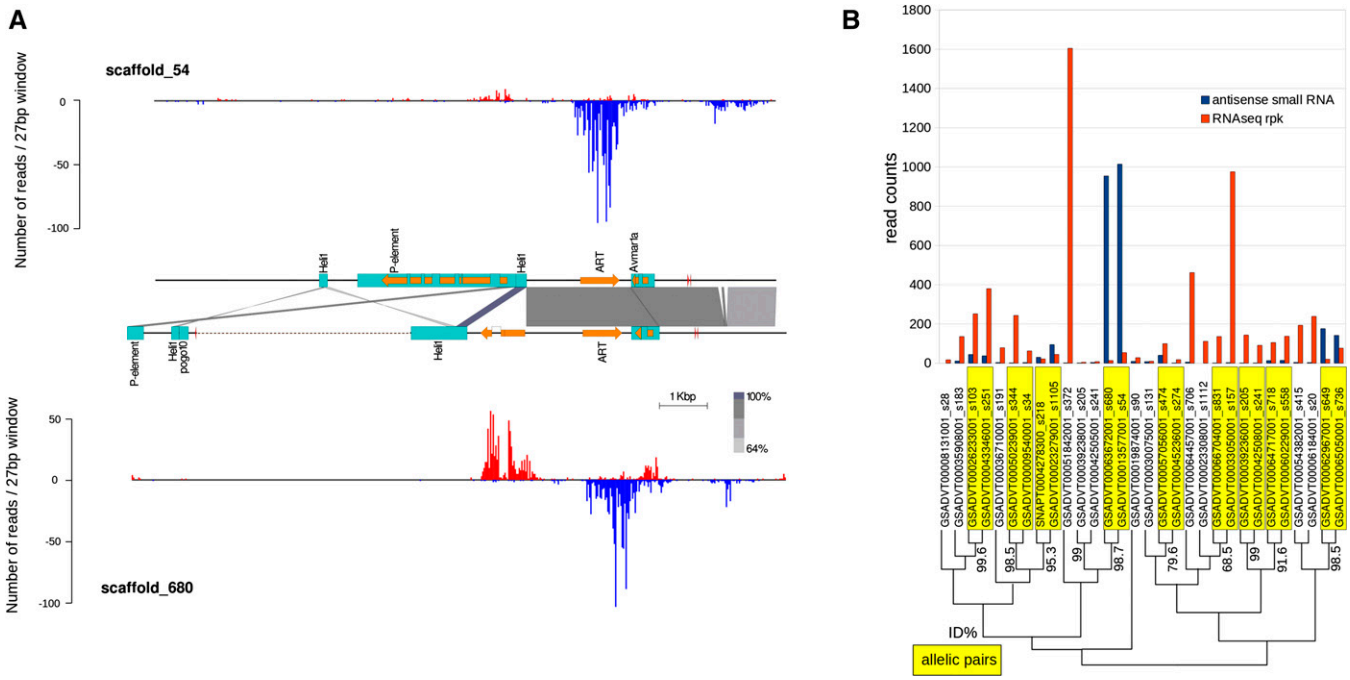


Figure 3 Small RNAs in the *A. vaga* ART family. (A) Synteny analysis and sRNA plots in two ART-containing scaffolds. Annotation and genomic context of gene models GSADVT00063672001 on scaffold 680 (1–13,600) and GSADVT00013577001 on scaffold 54 (13,000–1) are shown. Features are designated as in Figure 2C; Ψ, pseudogene. Plots above each region show sRNA read counts (blue, forward strand; red, reverse strand). It is not known whether synteny continues beyond the 5′-end of the scaffolds, as both start with the region of interest. (B) The top histogram shows RNA profiles for *A. vaga* ART genes (AI > 0), which clustered together as an ortholog group; the dendrogram below shows their phylogenetic relationships; labels are as in Figure 2B. Transcripts with high sRNA counts, assigned to an allelic pair after inspection of their genomic environment, are compared in A.

interstitial telomeric repeats framing the ART-containing segment (Figure 3A, red triangles) supports its relocation to the subtelomeric region.

Finally, an example of a foreign gene family with very few members also illustrates the greater likelihood of piRNA response in genes recently relocated to subtelomeres. The *A. vaga* RVT genes are domesticated single-copy reverse transcriptases of fungal/microsporidian origin (Gladyshev and Arkhipova 2011). Figure 4A shows that the two RVT_A copies lack introns, are embedded in a genomic environment rich in TEs and other foreign genes, are surrounded by telomeric repeats, and elicit a pronounced piRNA response, which is possibly driven by an LTR remnant at the 3′-end present in both alleles. In contrast, the two alleles in the transcriptionally active RVT_B lineage are located in a metazoan gene-rich environment, lack telomeric repeats in the vicinity, have acquired introns, and are virtually devoid of piRNA coverage (Figure 4B).

Host genes targeted by small RNAs

While the majority of genic sRNAs target CDS of foreign origin, a number of metazoan CDS are also serving as sRNA targets. It should be pointed out that not all HGTs can be detected by the AI metric—only those derived from nonmetazoan species. We have identified a subset of host genes with AI ≤ -30 (of likely metazoan origin) displaying sRNA coverage over the entire CDS region (a group of 327 genes, see *Materials and Methods*). As mentioned above, TEs and foreign genes tend to colocalize in subtelomeric regions of the *A. vaga* genome (Flot *et al.* 2013). It was therefore of interest to investigate possible colocalization between TEs and host genes with sRNA coverage. This co-occurrence was tested statistically, using the annotated TEs and host genes with AI ≤ -30. First, we counted the number of TEs in windows of 500, 1000, and 2000 bp around each of the selected host genes (a group of 327 genes with sRNA counts) and around host genes without sRNA coverage. Then, we compared the distribution of TE density around host genes with sRNA coverage and around the remaining host genes, using a *t*-test in three different window sizes. According to the *t*-test, the density of TEs is significantly higher around the selected group (AI ≤ -30 and sRNA coverage) for genomic windows as small as 500 bp (Table 1).

Another way of determining whether the selected group of host genes is located in the genomic environment similar to that of foreign genes is to employ the number of interstitial telomeric repeats in the vicinity of a gene as a proxy for localization in subtelomeric regions. We counted the number of telomeric repeats in windows of different sizes around annotated TEs, putatively foreign genes, selected metazoan genes with sRNA coverage, and the remaining bulk of metazoan genes. Table 2 shows that, in contrast to the majority of metazoan genes, the sRNA-covered host genes are about as likely to reside in the vicinity of telomeric repeats as are genes of putatively foreign origin.

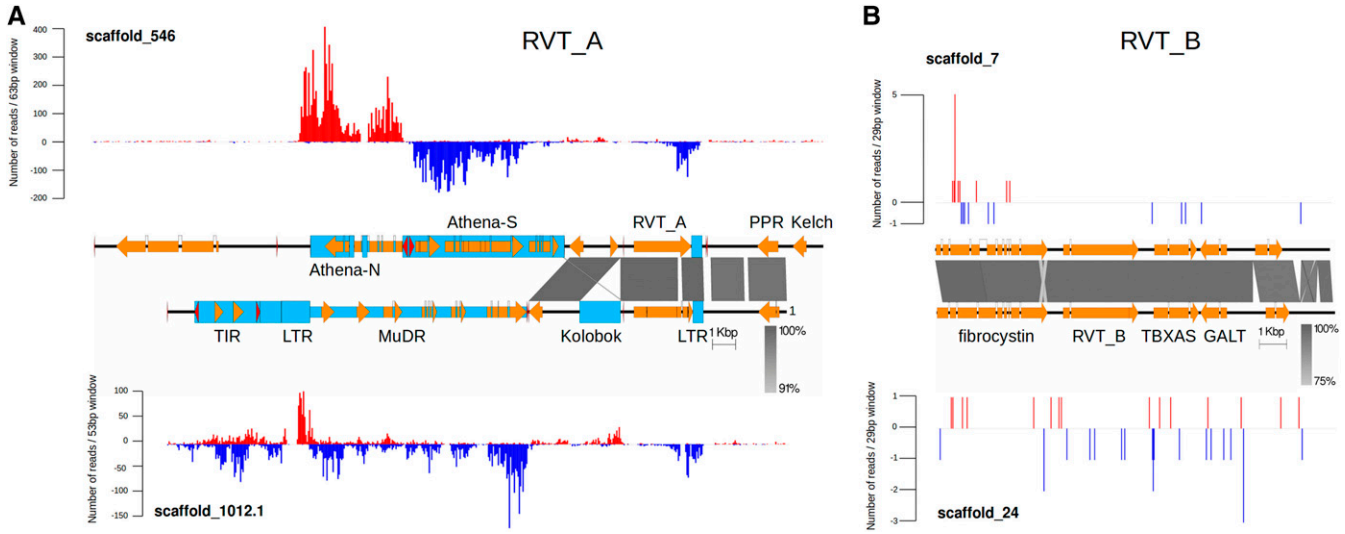


Figure 4 Small RNAs in the *A. vaga* RVT family. The sRNA coverage plots are shown for allelic scaffolds s546(18,500–50,000)/s1012(26,700–1) (A) and s7(1–14,500)/s24(20,700–35,300) (B), comprising the *rvtA* and *rvtB* gene pairs, respectively. Features are designated as in Figure 2C. Scaffold 1012 has an assembly gap (mostly filled with scaffold 6459) within the *rvt* gene, which was previously sequenced in full for both members of the allelic pair from a fosmid library (JN235987.1 and JN235988.1; Gladyshev and Arkhipova 2011). An additional pseudogenized pair (*rvtC*) does not yield transcripts or sRNAs and is not shown.

Host multigene families with small RNA coverage

To find out whether the limited subset of metazoan CDS yielding sRNA coverage can also be regarded as relatively recent, we focused on the sufficiently large multigene family that permits informative phylogenetic analysis, which was identified as an overrepresented category in functional GO annotations (File S1, sheet 3). The CYP450 enzymes catalyze oxidative transformation, which contributes to a broad array of biological functions in living organisms, such as activation or inactivation of many endogenous and exogenous bioactive compounds, particularly xenobiotic detoxification, or even modification of nonribosomal peptides (Guengerich 1991; Haslinger *et al.* 2015). CYP enzymes are found in all kingdoms and domains of life (metazoans, plants, fungi, protists, bacteria, and archaea) and even in a virus. The number of putatively functional full-length CYP genes in metazoan genomes can reach a few dozen copies (57 in humans, 102 in mice, 84 in *D. melanogaster*, 74 in *Caenorhabditis elegans*) (Nelson 2009). After ortholog clustering analysis, which helps to distinguish orthologs and paralogs in large metazoan gene families, we identified 92 *A. vaga* CYP forming an ortholog group (File S1, sheet 4). Despite the fact that none of the copies appear to yield truncated proteins, they differ in transcriptional activity (RNA-seq counts), and, importantly, their origins are quite dissimilar. BLASTN and phylogenetic analysis (Figure 5A) showed AvCYP genes to be related to different CYP clans previously identified in other taxa. The AI values also indicate that some of them could have been acquired by independent HGT events.

Analysis of small RNA reads and their mapping patterns reveals a patchy distribution (Figure 5A), with one of the expanded clades displaying higher levels of sRNA coverage. This expanded node, marked with an asterisk (16 CYP genes, 12 of which display sRNA coverage), is shown in Figure 5B as a dendrogram with small RNA counts and RNA-seq reads per kilobase (rpk) values per CYP copy. An allelic pair with the highest transcriptional activity lacks sRNA coverage, while in the remaining allelic pairs the transcript levels are much lower, in agreement with increased levels of sRNA coverage. Interestingly, in this group, sRNA coverage is correlated with the presence of an oppositely oriented transcriptional unit in the 3′-flanking region (File S5, sheet 5).

Discussion

Bdelloid rotifers represent a particularly interesting system for investigating RNA-mediated silencing phenomena. Our earlier work demonstrated that the genome of the sequenced laboratory strain of the bdelloid rotifer *A. vaga* has a large number of TE families, but very low copy numbers of TEs within each family, hinting at the existence of mechanisms which can efficiently suppress proliferation of incoming TEs (Flot *et al.* 2013; reviewed in Hayes 2013). Furthermore, our genome-wide analysis showed an unusually high degree of expansion and diversification of the principal gene families involved in RNA-mediated silencing, such as Dicers, Ago/Piwi, and RdRPs. The present study investigates genome-wide patterns of distribution of small RNAs in *A. vaga*, focusing specifically on the piRNA-like population, which is known to be responsible for TE silencing in the germ line at both transcriptional and post-transcriptional levels to ensure genome integrity and shows *trans*-generational effects (Ashe *et al.* 2012; Shirayama *et al.* 2012). In agreement with our expectations, 71% of the reads mapping to annotated features were TE

Table 1 Transposon distribution around host genes with and without piRNA coverage

TE counts	Window size	500 bp	1000 bp	2000 bp
Host genes no piRNA (15,821)	Number	1566	1871	2460
	Average	0.0990	0.1183	0.1555
	SD	0.3706	0.4051	0.4690
Host genes piRNA (327)	Number	68	89	129
	Average	0.2080	0.2722	0.3945
	SD	0.5011	0.5615	0.7008
t-test	*P*-value	1.12*E*-004	1.30*E*-006	2.37*E*-009

The difference was tested statistically using a two-tailed *t*-test. TE numbers (including full-length and fragmented copies) were counted for each CDS feature (AI < 0) in three different window sizes (500, 1000, and 2000 bp). TE density is significantly higher around selected host genes with small RNA profiles. SD, standard deviation.

Table 2 Occurrence of telomeric repeats near transposons and genes

Category	Window size	500 bp	1000 bp	2000 bp
TEs (10,183)	Number	749	974	1304
	Average	0.0736	0.0956	0.1281
	SD	0.3436	0.3943	0.4664
Genes AI > 0 (8954)	Number	203	347	522
	Average	0.0227	0.0388	0.0583
	SD	0.1801	0.2478	0.3156
Host genes piRNA (327)	Number	9	18	29
	Average	0.0275	0.0550	0.0887
	SD	0.1978	0.2878	0.3522
Host genes no piRNA (15,821)	Number	51	90	140
	Average	0.0032	0.0057	0.0088
	SD	0.0679	0.0939	0.1208

The number of interstitial telomeric repeat annotations (each annotation including one or more consecutive hexamer units TGTGGG or TGAGGG) was counted in three different window sizes (500, 1000, and 2000 bp) around annotated TEs (including full-length and partial copies); CDS of putatively foreign genes (AI > 0); host genes with small RNA coverage; and host genes without small RNA coverage. SD, standard deviation.

derived. Furthermore, we found that most of TE-derived reads are in the antisense orientation and exhibit 5′-uridine bias, indicating that the investigated sRNA population largely consists of primary piRNAs. For overlapping sense–antisense reads, no ping-pong amplification signature could be detected, resembling the situation with "21U" piRNAs in *C. elegans* (Das *et al.* 2008). A plausible explanation for the lack of ping-pong signatures and for the overall paucity of sense piRNAs in our datasets is that amplification of the silencing signal, if it occurs in *A. vaga*, could be achieved via the expanded set of RdRP proteins, rather than via the ping-pong cycle involving secondary piRNAs bound to Ago proteins.

In the course of this work, we discovered that a significant proportion of sRNAs can be mapped to host genes of predominantly foreign origin, which were previously shown to colocalize with TEs in the *A. vaga* genome (Flot *et al.* 2013). Upon inspection, foreign genes with the strongest sRNA signal could be classified as more recent insertion events, as evidenced by their placement in the more recent branches of the phylogenetic tree, and displayed association with short stretches of interstitial telomeric repeats. Notably, the chances of occurrence of a telomeric repeat in the vicinity of putatively foreign genes are about 10-fold higher than in the vicinity of metazoan genes (Table 2). Interestingly, DNA segments flanked by telomeric repeats do not show microsynteny in the immediate genomic environments of β-lactamase genes from neighboring branches of the phylogenetic tree (Figure 2C), suggesting that these segments were not generated by large segmental duplications, but may have been relocated individually or acquired by HGT. Overall, sRNA-mediated silencing of foreign genes could represent a convenient mechanism for their adaptation to new genomic environments.

Furthermore, 327 host genes of apparently metazoan origin (AI ≤ -30) also exhibited sRNA coverage and were often colocalized with TEs (Table 1). These host genes also display higher incidence of co-occurrence with interstitial telomeric repeats (Table 2). While these genes may have originated from intragenomic amplification events placing them into subtelomeric regions, their origin by rotifer-to-rotifer HGT cannot be ruled out, since the methodology does not permit detection of transfers between metazoans. Intriguingly, among the expanded set of metazoan CYP450 genes, those with the highest levels of sRNA coverage exhibited strong association with the presence of an antisense transcriptional unit in the 3′-flanking region. Convergent transcription from the oppositely oriented units is expected to yield dsRNAs, which can be processed into siRNAs in a Dicer-dependent manner (Czech *et al.* 2008; Okamura *et al.* 2008). While the correlation between the production of siRNAs and piRNAs from the corresponding loci needs to be investigated in further studies, previous work in *C. elegans* reported that dsRNA could trigger the establishment of silencing chromatin modifications at the corresponding loci (Gu *et al.* 2012).

It is noteworthy that the pattern of piRNA distribution over the entire gene body in putatively foreign genes (Figure 1D) resembles that of TEs, rather than that of the previously reported genic piRNAs, which map primarily to the 3′-UTR regions (Robine *et al.* 2009; Saito *et al.* 2009; Iwasaki *et al.* 2015). Indeed, the bulk of sRNA-covered metazoan genes display a pronounced peak of sRNAs in the 3′-UTRs (Figure 1D). In some cases, however, the pattern appears uniform, as in selected CYP450 genes (Figure 5). In the future, when the gene knockout and/or knockdown methods are developed for bdelloids, these differences could be explored by perturbing the activity of different Piwi variants.

Silencing of TEs in the germ line is essential for preserving genomic integrity, and in *A. vaga* it evidently helps to prevent uncontrolled TE proliferation, as expected from the low TE content and diversified sRNA silencing machinery (Flot *et al.* 2013). Almost every active TE family displays measurable levels of sRNA coverage, which is largely accompanied by relatively low transcriptional activity. In fact, the only TE family in Figure S4 that is virtually devoid of piRNA coverage is the DNA TE Zator2, the two copies of which lack TIRs and are located in a syntenic environment, and are therefore likely to represent two alleles of a domesticated TE-derived gene. Nevertheless, the overall transcriptional activity of certain TE families, as well as of sRNA-covered host genes, is not

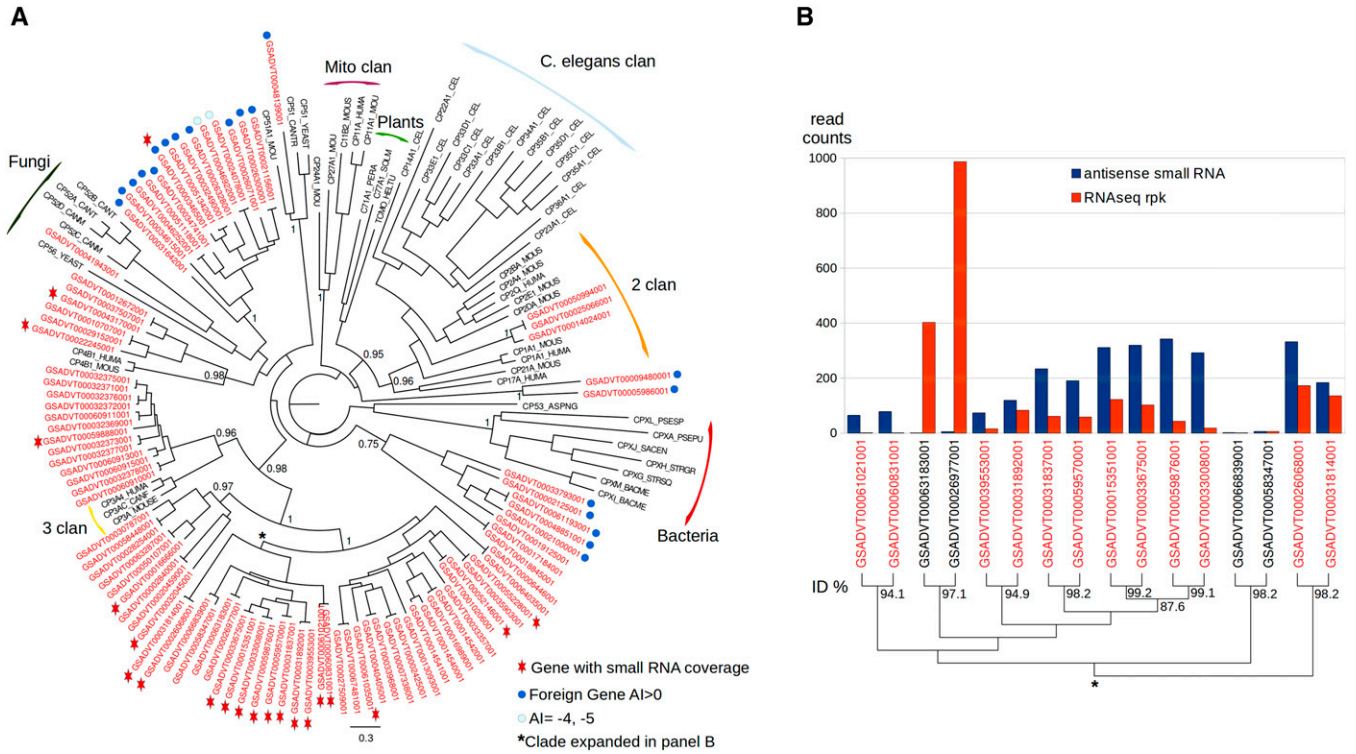


Figure 5 Small RNAs in the *A. vaga* cytochrome P450 (CYP) gene family. (A) The maximum likelihood phylogenetic tree includes *A. vaga* CYP homologs (in red) and CYP genes from other domains of life. CYP gene sequences and names were taken from the CytP450 homepage (Nelson 2009). Red stars denote genes with small RNA coverage. Genes marked with blue dots indicate probable foreign origin (AI > 0). The node marked with an asterisk (*), which includes most genes with significant antisense read counts (>100), is shown in B. (B) The histogram on the top shows RNA profiles from the AvCYP node marked in A, with antisense sRNA counts and RNA-seq rpk counts. The phylogenetic relatedness and nucleotide sequence identity (%) between nearest neighbors is shown in the dendrogram below. Two pairs grouped with 87.6% identity do not exhibit synteny in their genomic environments. Genes with sRNA reads, marked by red stars in A, are shown in red.

always strictly anticorrelated with the degree of sRNA coverage. In other systems, piRNAs are known to silence TEs primarily in the germ line and in the adjacent somatic cells, and their transcriptional activity may be higher in other somatic tissues. Construction of tissue-specific RNA-seq and sRNA libraries is not feasible in rotifers due to their microscopic size and prevalence of soft tissues not amenable to dissection. It should also be kept in mind that a certain level of transcriptional activity is necessary to initiate piRNA biogenesis in the first place. Furthermore, a subset of small RNAs in *C. elegans* has been shown to possess antisilencing properties, targeting selected genes for expression in the germ line (Lee *et al.* 2012; Shirayama *et al.* 2012). Nevertheless, despite the complexity of transcriptional responses mediated by sRNAs together with different protein complexes, in general our data agree with the interpretation that the majority of sRNAs in *A. vaga* act to silence expression of active TEs, as well as a subset of host genes.

Silencing of relocated genes may be advantageous for rotifers, because such genes are mostly providing functions that are optional and/or redundant, and their expression could be favored only under certain conditions, for instance following environmental stresses or pathogen attacks. Their colocalization with TEs in the extended subtelomeric regions, while placing them at a higher risk of loss, disruption, and pseudogenization, can also provide an easy mechanism for diploidization of an introduced single-copy gene via break-induced replication at telomeres, which does not require more than one template switch, as with other conversion events. Note that in each allelic piRNA-producing pair, both gene copies could be affected by silencing, even if a TE insertion occurs in only one of the copies, thus initiating changes akin to paramutation. In *Drosophila*, paramutation has been shown to involve subtelomeric piRNA-producing loci (de Vanssay *et al.* 2012). Thus, gene acquisition at telomeres, which in bdelloids may involve relatively short segments of DNA (a few kilobases in length) containing one or several ORFs and flanked by short stretches of telomeric repeats, as shown in Figure 2C, Figure 3A, and Figure 4A and our previous work (Gladyshev *et al.* 2008; Gladyshev and Arkhipova 2011), could be accompanied by piRNA-mediated silencing, with the potential to modulate expression of acquired genes in subsequent generations.

Acquisition of piRNA coverage by foreign genes may reflect the advantage of mechanisms that permit piRNA production without the need for incorporation into extended piRNA clusters, which are the main producers of piRNAs in *Drosophila*, but instead require only a nearby TE insertion (or another oppositely oriented transcriptional unit), which could give rise to an antisense transcript with the potential to be

processed into piRNAs (Shpiz *et al.* 2014). Even TEs inserted in the same orientation have the potential to promote aberrant transcription due to widespread presence of antisense promoters (Russo *et al.* 2015). While *A. vaga* harbors many extended genomic regions enriched in TEs that yield sRNA coverage over large distances, thereby resembling extended piRNA clusters in other species, a significant share of piRNA production also comes from much smaller regions where it apparently can be acquired on a gene-by-gene basis. While it is not possible to assign piRNA-producing loci to defined chromosomal locations due to the fragmentary nature of the genome assembly, especially in telomeric regions, which are notoriously hard to assemble, the observed association between TEs, foreign genes, and telomeric repeats indicates that the extended subtelomeric regions could play a special role in acquisition of foreign genes and their adaptation to new genomic environments. Members of multigene families with sRNA coverage show characteristics of more recent acquisition, as inferred by phylogenetic analysis, and may be tentatively assigned to the more distal regions of the chromosomes, as follows from the decreased scaffold lengths typical for poorly assembled telomeric regions and from the increased frequency of telomeric repeat occurrence in the vicinity. In contrast, the earlier branches of phylogenetic trees of multigene families, which lack sRNA coverage, are usually located on longer TE-poor scaffolds forming the core genome, and harbor very few if any telomeric repeats, implying gradual incorporation into the core genome accompanied by erasure of telomeric repeat signatures.

Future population genomic studies should shed light on the evolutionary histories of modular/repetitive foreign genes not easily amenable to phylogenetic analysis, by investigating their presence/absence across populations. In sum, the present work helps to understand the complexity of interactions between TEs and coding sequences in the processes of gene acquisition, expansion, and diversification in the genomes of taxonomic groups with distinctive properties, which belong to the still poorly studied branches on the Tree of Life.

Acknowledgments

We thank Nelson Lau for advice and help with small RNA library construction protocols and stimulating discussions, John Stegeman for advice on P450, and the Genomics Core Facility at Brown University for help with sequencing. This research was supported by National Science Foundation grant MCB-1121334 (RVT genes) and National Institutes of Health grant GM111917 (horizontal gene transfer) to I.R.A.

Literature Cited

Ambler, R. P., 1980 The structure of beta-lactamases. Philos. Trans. R. Soc. Lond. B Biol. Sci. 289: 321–331.

Anders, S., P. T. Pyl, and W. Huber, 2014 HTSeq: a Python framework to work with high-throughput sequencing data. Bioinformatics 31: 166–169.

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

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

Arkhipova, I., and M. Meselson, 2000 Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. USA 97: 14473–14477.

Arkhipova, I. R., and M. Meselson, 2005a Deleterious transposable elements and the extinction of asexuals. BioEssays 27: 76–85.

Arkhipova, I. R., and M. Meselson, 2005b Diverse DNA transposons in rotifers of the class Bdelloidea. Proc. Natl. Acad. Sci. USA 102: 11781–11786.

Arkhipova, I. R., and F. Rodriguez, 2013 Genetic and epigenetic changes involving (retro)transposons in animal hybrids and polyploids. Cytogenet. Genome Res. 140: 295–311.

Arkhipova, I. R., I. A. Yushenova, and F. Rodriguez, 2013 Endonuclease-containing Penelope retrotransposons in the bdelloid rotifer *Adineta vaga* exhibit unusual structural features and play a role in expansion of host gene families. Mob. DNA 4: 19.

Ashe, A., A. Sapetschnig, E.-M. Weick, J. Mitchell, M. P. Bagijn *et al.*, 2012 piRNAs can trigger a multigenerational epigenetic memory in the germline of *C. elegans*. Cell 150: 88–99.

Bailly-Bechet, M., A. Haudry, and E. Lerat, 2014 "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mob. DNA 5: 13.

Boschetti, C., A. Carr, A. Crisp, I. Eyres, Y. Wang-Koh *et al.*, 2012 Biochemical diversification through foreign gene expression in bdelloid rotifers. PLoS Genet. 8: e1003035.

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

Chambeyron, S., A. Popkova, G. Payen-Groschêne, C. Brun, D. Laouini *et al.*, 2008 piRNA-mediated nuclear accumulation of retrotransposon transcripts in the Drosophila female germline. Proc. Natl. Acad. Sci. USA 105: 14964–14969.

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

Czech, B., C. D. Malone, R. Zhou, A. Stark, C. Schlingeheyde *et al.*, 2008 An endogenous small interfering RNA pathway in Drosophila. Nature 453: 798–802.

Das, P. P., M. P. Bagijn, L. D. Goldstein, J. R. Woolford, N. J. Lehrbach *et al.*, 2008 Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the *Caenorhabditis elegans* germline. Mol. Cell 31: 79–90.

de Vanssay, A., A.-L. Bougé, A. Boivin, C. Hermant, L. Teysset *et al.*, 2012 Paramutation in Drosophila linked to emergence of a piRNA-producing locus. Nature 490: 112–115.

Donlin, M. J., 2007 Using the Generic Genome Browser (GBrowse). Curr. Protoc. Bioinformatics Chapter 9: Unit 9.9.

El Baidouri, M., K. D. Kim, B. Abernathy, S. Arikit, F. Maumus *et al.*, 2015 A new approach for annotation of transposable elements using small RNA mapping. Nucleic Acids Res. 43: e84.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792–1797.

Fischer, S., B. P. Brunk, F. Chen, X. Gao, O. S. Harb *et al.*, 2011 Using OrthoMCL to assign proteins to OrthoMCL-DB groups or to cluster proteomes into new ortholog groups. Curr. Protoc. Bioinformatics Chapter 6: Unit 6.12, 1–19.

Flot, J.-F., B. Hespeels, X. Li, B. Noel, I. Arkhipova *et al.*, 2013 Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. Nature 500: 453–457.

Francia, S., F. Michelini, A. Saxena, D. Tang, M. de Hoon *et al.*, 2012 Site-specific DICER and DROSHA RNA products control the DNA-damage response. Nature 488: 231–235.

Gladyshev, E., and M. Meselson, 2008 Extreme resistance of bdelloid rotifers to ionizing radiation. Proc. Natl. Acad. Sci. USA 105: 5139–5144.

Gladyshev, E. A., and I. R. Arkhipova, 2010 A subtelomeric non-LTR retrotransposon Hebe in the bdelloid rotifer *Adineta vaga* is subject to inactivation by deletions but not 5′ truncations. Mob. DNA 1: 12.

Gladyshev, E. A., and I. R. Arkhipova, 2011 A widespread class of reverse transcriptase-related cellular genes. Proc. Natl. Acad. Sci. USA 108: 20311–20316.

Gladyshev, E. A., M. Meselson, and I. R. Arkhipova, 2008 Massive horizontal gene transfer in bdelloid rotifers. Science 320: 1210–1213.

Gu, S. G., J. Pak, S. Guang, J. M. Maniar, S. Kennedy *et al.*, 2012 Amplification of siRNA in *Caenorhabditis elegans* generates a transgenerational sequence-targeted histone H3 lysine 9 methylation footprint. Nat. Genet. 44: 157–164.

Guengerich, F. P., 1991 Reactions and significance of cytochrome P-450 enzymes. J. Biol. Chem. 266: 10019–10022.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52: 696–704.

Gunawardane, L. S., K. Saito, K. M. Nishida, K. Miyoshi, Y. Kawamura *et al.*, 2007 A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. Science 315: 1587–1590.

Hall, B. G., and M. Barlow, 2004 Evolution of the serine beta-lactamases: past, present and future. Drug Resist. Updat. 7: 111–123.

Haslinger, K., M. Peschke, C. Brieke, E. Maximowitsch, and M. J. Cryle, 2015 X-domain of peptide synthetases recruits oxygenases crucial for glycopeptide biosynthesis. Nature 521: 105–109.

Hayes, G., 2013 Genome guardians. Cell 154: 1167–1169.

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

Iwasaki, Y. W., M. C. Siomi, and H. Siomi, 2015 PIWI-Interacting RNA. Its Biogenesis and Functions. Annu. Rev. Biochem. 84: 405–433.

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

Lau, N. C., A. G. Seto, J. Kim, S. Kuramochi-Miyagawa, T. Nakano *et al.*, 2006 Characterization of the piRNA complex from rat testes. Science 313: 363–367.

Lee, H.-C., S.-S. Chang, S. Choudhary, A. P. Aalto, M. Maiti *et al.*, 2009 qiRNA is a new type of small interfering RNA induced by DNA damage. Nature 459: 274–277.

Lee, H.-C., W. Gu, M. Shirayama, E. Youngman, D. Conte *et al.*, 2012 *C. elegans* piRNAs mediate the genome-wide surveillance of germline transcripts. Cell 150: 78–87.

Li, L., C. J. Stoeckert, and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.

Majumdar, S, and D. C. Rio, 2015 P transposable elements in Drosophila and other eukaryotic organisms. Microbiol. Spectr. 3(2). pii: MDNA3–0004–2014.

Mark Welch, D. B., J. L. Mark Welch, and M. Meselson, 2008 Evidence for degenerate tetraploidy in bdelloid rotifers. Proc. Natl. Acad. Sci. USA 105: 5145–5149.

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

Munafó, D. B., and G. B. Robb, 2010 Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. RNA 16: 2537–2552.

Nelson, D. R., 2006 Cytochrome P450 nomenclature, 2004. Methods Mol. Biol. 320: 1–10.

Nelson, D. R., 2009 The cytochrome p450 homepage. Hum. Genomics 4: 59–65.

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

Poinar, G. O., and C. Ricci, 1992 Bdelloid rotifers in Dominican amber: evidence for parthenogenetic continuity. Experientia 48: 408–410.

Ricci, C., and G. Melone, 2000 Key to the identification of the genera of bdelloid rotifers. Hydrobiologia 418: 73–80.

Robine, N., N. Lau, S. Balla, Z. Jin, K. Okamura *et al.*, 2009 A broadly conserved pathway generates 3′ UTR-directed primary piRNAs. Curr. Biol. 19: 2066–2076.

Russo, J., A. W. Harrington, and M. Steiniger, 2015 Antisense transcription of retrotransposons in *Drosophila*: the origin of endogenous small interfering RNA precursors. Genetics 202: 107–121.

Saito, K., S. Inagaki, T. Mituyama, Y. Kawamura, Y. Ono *et al.*, 2009 A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature 461: 1296–1299.

Segers, H., 2007 Annotated checklist of the rotifers (Phylum Rotifera), with notes on nomenclature, taxonomy and distribution. Zootaxa 1564: 1–104.

Shirayama, M., M. Seth, H.-C. Lee, W. Gu, T. Ishidate *et al.*, 2012 piRNAs initiate an epigenetic memory of nonself RNA in the *C. elegans* germline. Cell 150: 65–77.

Shpiz, S., S. Ryazansky, I. Olovnikov, Y. Abramov, and A. Kalmykova, 2014 Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the drosophila germline. PLoS Genet. 10: e1004138.

Signorovitch, A., J. Hur, E. Gladyshev, and M. Meselson, 2015 Allele sharing and evidence for sexuality in a mitochondrial clade of bdelloid rotifers. Genetics 200: 581–590.

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

Siomi, M. C., K. Sato, D. Pezic, and A. Aravin, 2011 PIWI-interacting small RNAs : the vanguard of genome defence. Nat. Rev. Mol. Cell Biol. 12: 246–258.

Smit, A. F. A., R. Hubley, and P. Green, 2013 RepeatMasker Open-4.0. Available at: http://www.repeatmasker.org. Accessed January 7, 2015.

Sullivan, M. J., N. K. Petty, and S. A. Beatson, 2011 Easyfig: a genome comparison visualizer. Bioinformatics 27: 1009–1010.

Suyama, M., D. Torrents, and P. Bork, 2006 PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34: W609–W612.

Le Thomas, A., K. F. Tóth, and A. A. Aravin, 2014 To be or not to be a piRNA: genomic origin and processing of piRNAs. Genome Biol. 15: 204.

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

Wei, W., Z. Ba, M. Gao, Y. Wu, Y. Ma *et al.*, 2012 A role for small RNAs in DNA double-strand break repair. Cell 149: 101–112.

Communicating editor: S. Poethig

GENETICS

Supporting Information
www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.186734/-/DC1

Multitasking of the piRNA Silencing Machinery: Targeting Transposable Elements and Foreign Genes in the Bdelloid Rotifer *Adineta vaga*

Fernando Rodriguez and Irina R. Arkhipova

DOI: 10.1534/genetics.116.186734

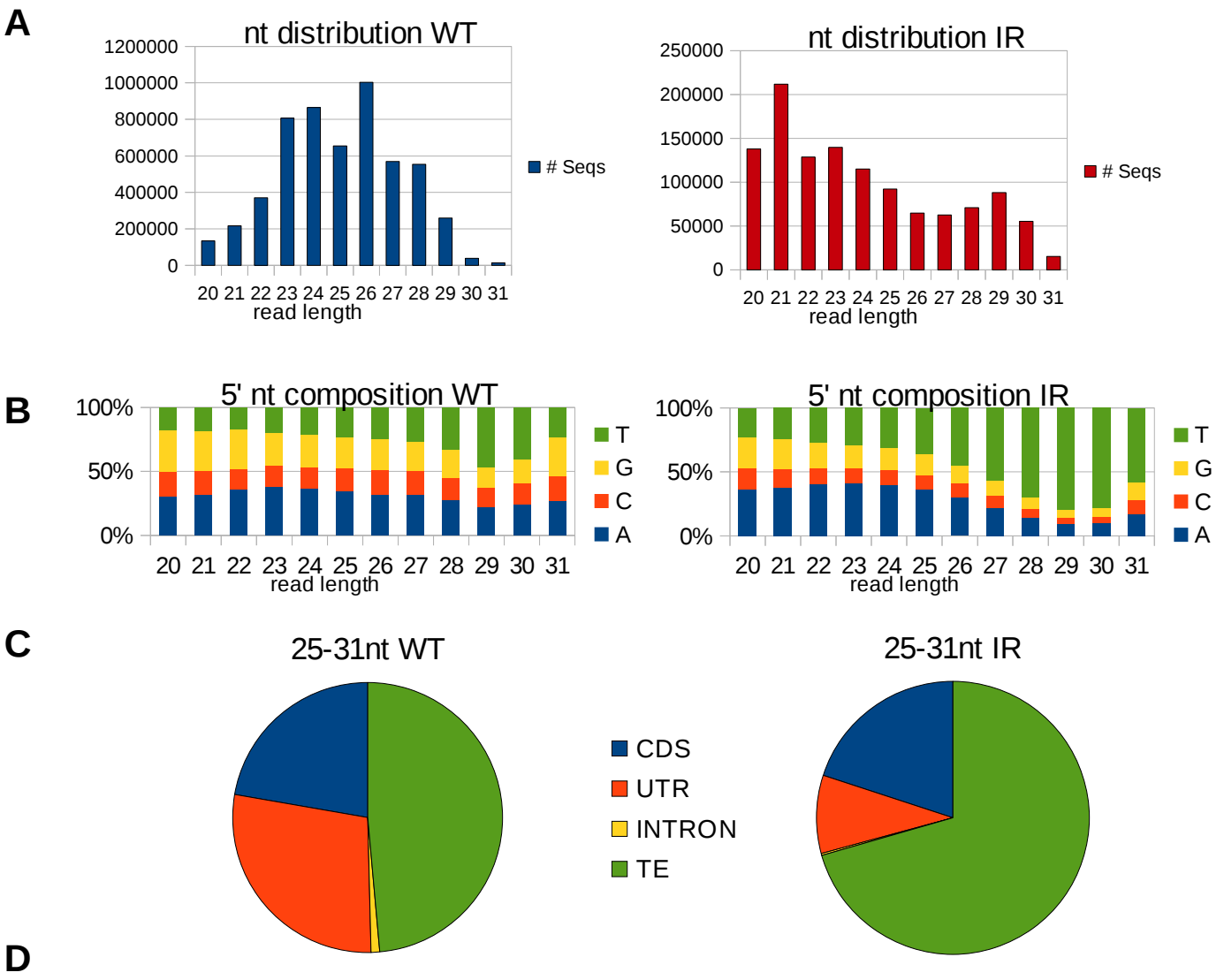


Figure S1. Comparison of wild type (WT) and irradiated (IR) small RNAs from total *A. vaga* RNA extractions. (A) Nucleotide length distribution (after filtering rRNA and tRNA-derived sequences) across WT and IR samples. (B) The 5' nucleotide composition across the read lengths in WT and IR samples. (C) Pie charts showing the proportion of 25-31 nt small RNAs mapped in reverse orientation to CDS, UTR, intron, or annotated TE features of the A. vaga genome in WT and IR samples. (D) Examples of 25-31 nt small RNAs coverage plots in WT and IR samples across different TE families.

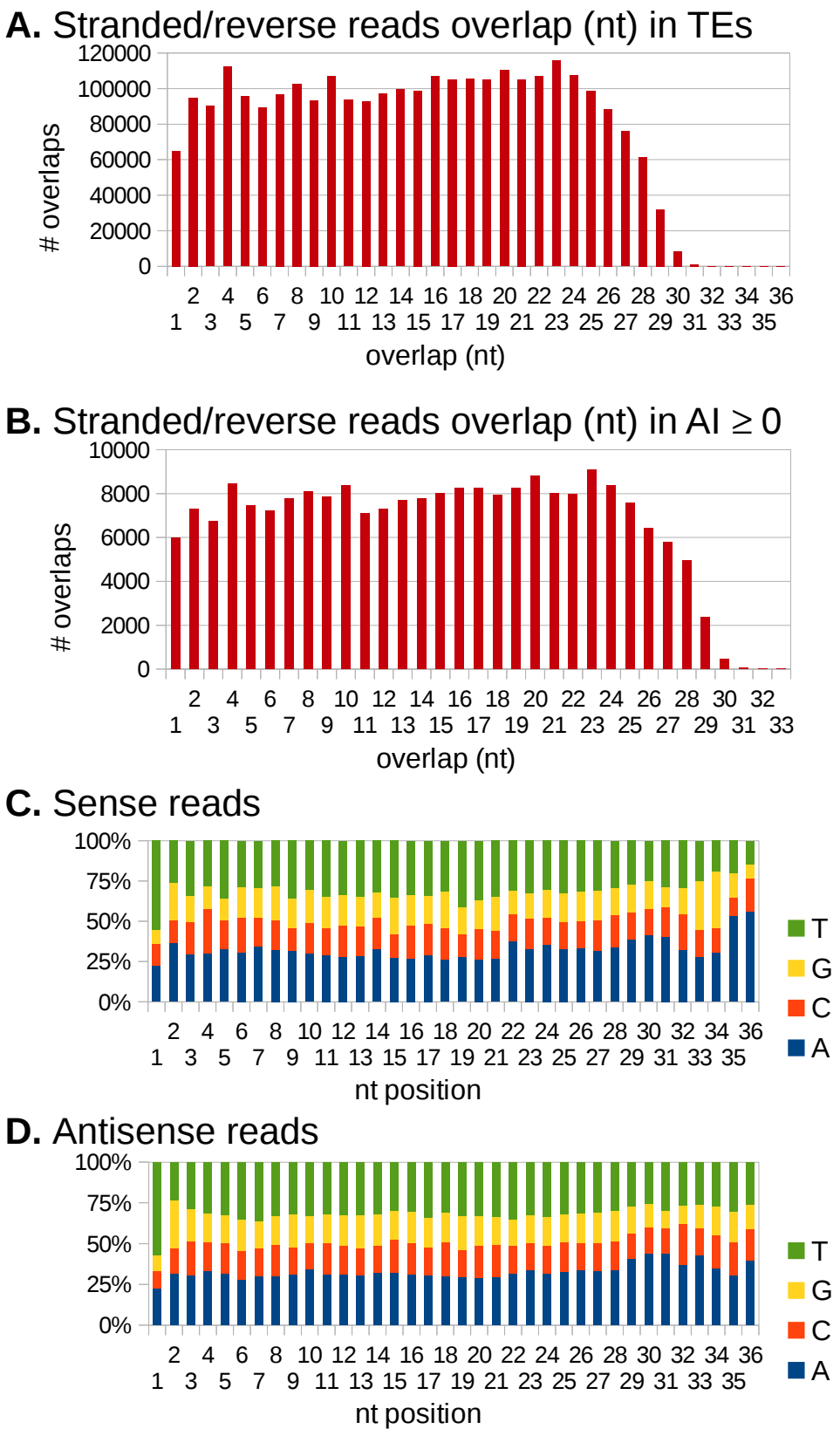


Figure S2. Lack of ping-pong signatures in *A. vaga*. Nucleotide overlaps were extracted from paired sense and antisense reads mapped to transposable elements (A) and to putatively foreign genes with AI ≥ 0 (B). Panels (C) and (D) show nucleotide composition of sense and antisense reads, respectively, as percentage of nucleotides (A,T,G,C) along small RNAs mapped to annotations on sense and antisense strands. Each nucleotide position represents he sum of all sRNA reads with the specified offset.

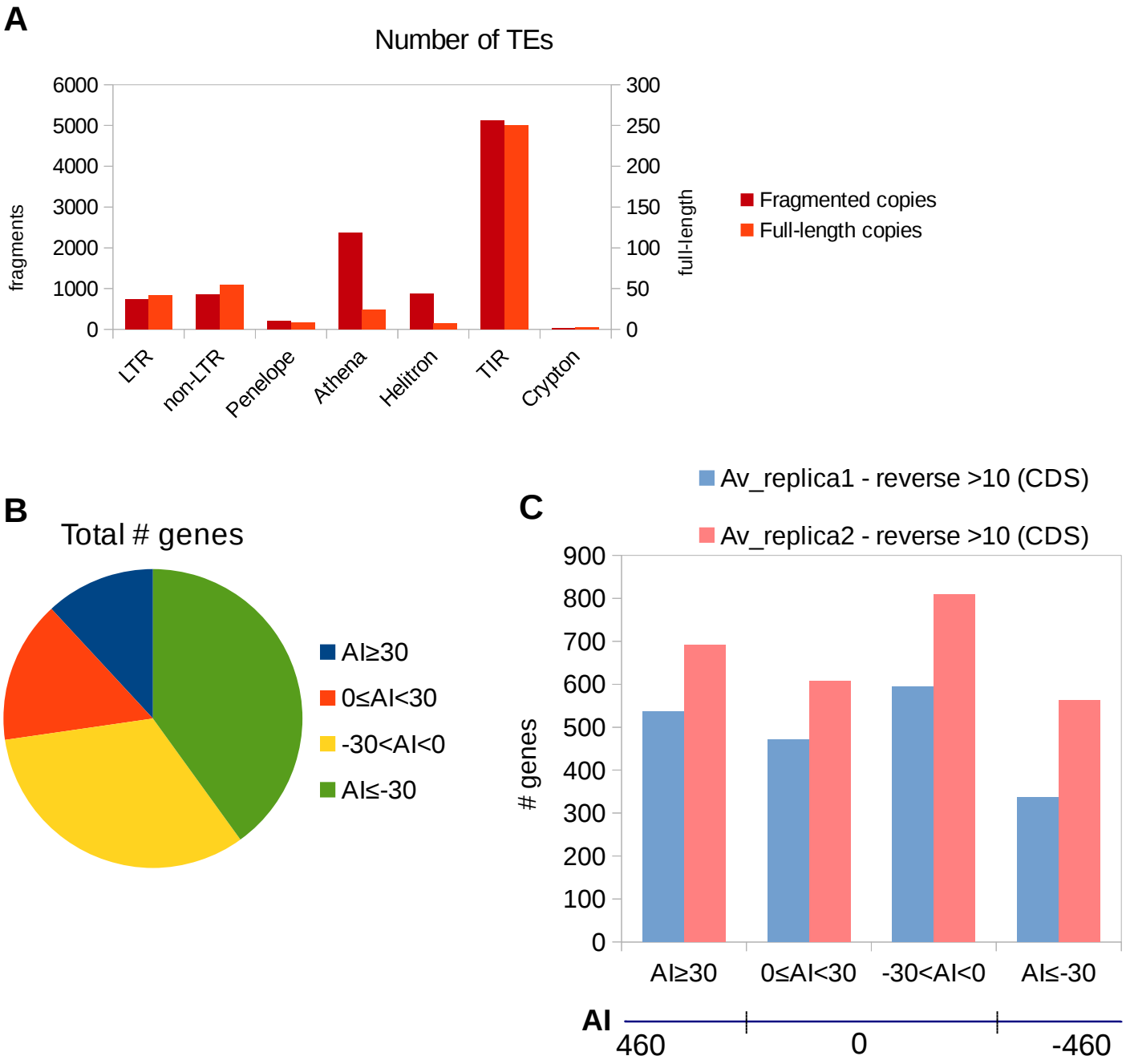
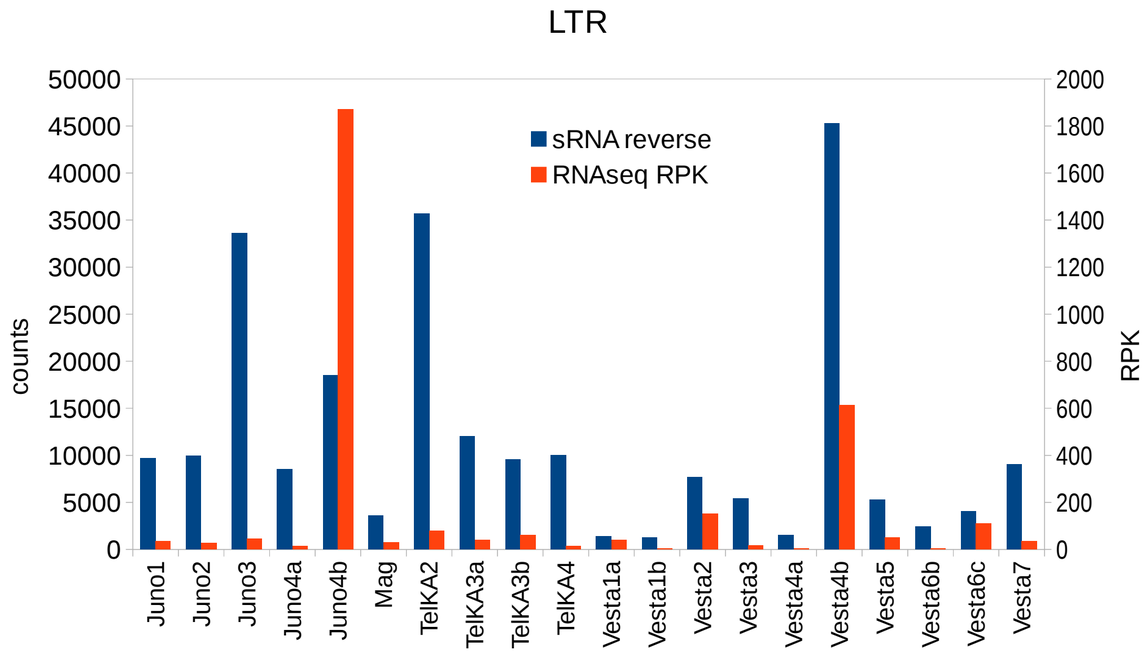
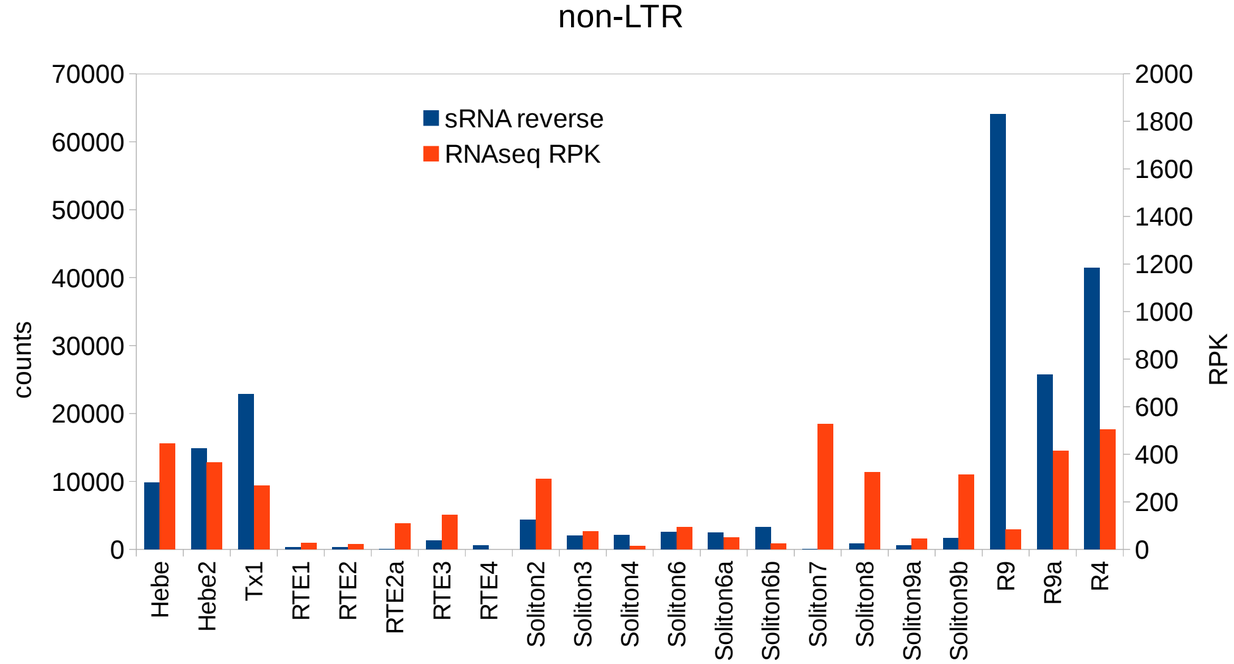
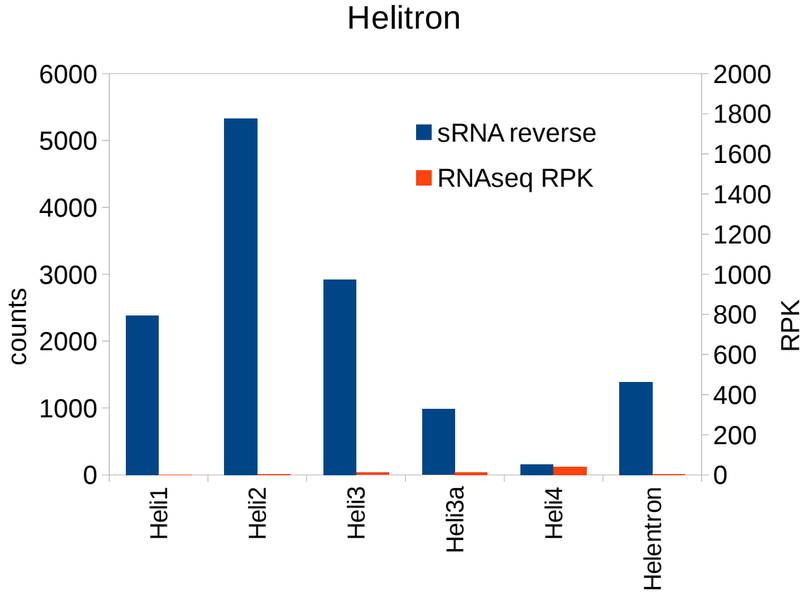


Figure S3. Breakdown of *A. vaga* TEs and CDS by categories. (A) Numbers of fragmented TE copies longer than 100 bp (left Y-axis) and full-length TE copies (right Y-axis; note the difference in scale) were estimated by BLAT, using full-length sequences as queries. (B) Total number of genes representing each AI category (not including genes without blast hits for index calculation). (C) Number of genes with small RNA coverage (in reverse orientation, at least 10 reads per CDS) by gene category in each of the replicas.







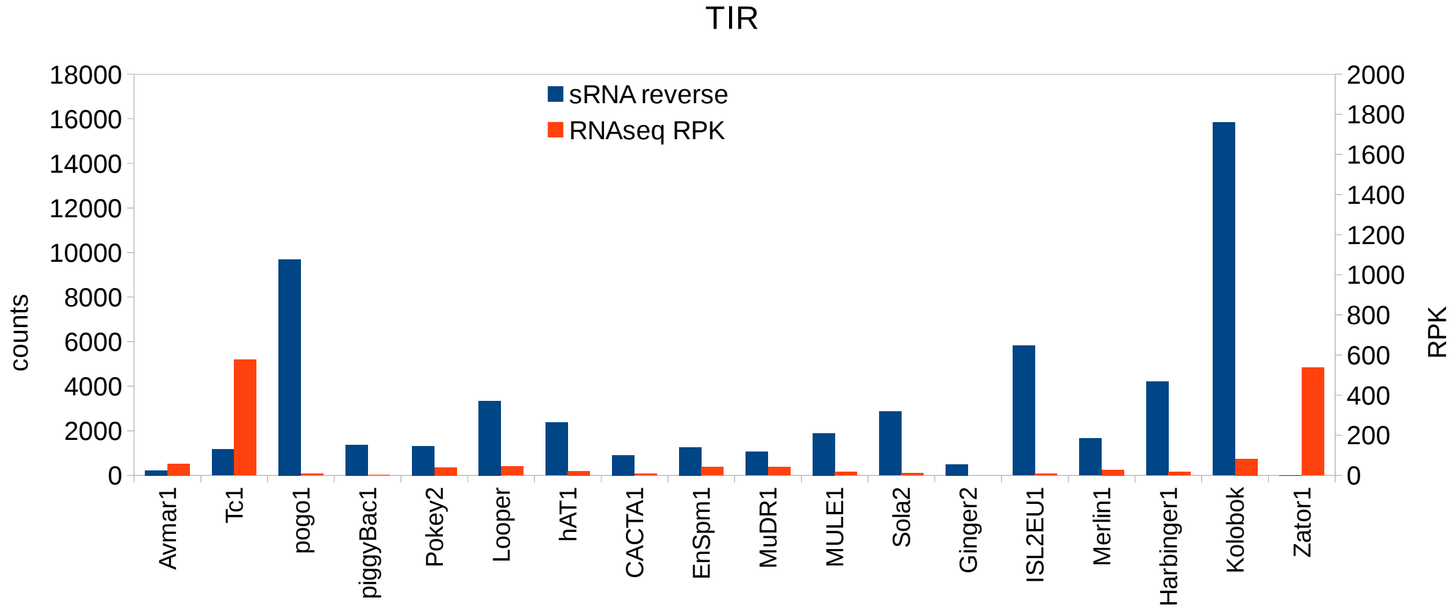


Figure S4. RNA profiles for different TE classes and families. Distribution of RNAseq reads with RPK (reads per kilobase, right Y-axis) values, and small RNAs in reverse orientation (total counts, left Y-axis) mapped to annotated TEs in the *A. vaga* genome. Shown are examples of full-length copies (after BLAT comparison) in individual families from LTR, non-LTR, TIR, and Helitron groups, which display significant sRNA and RNA-seq counts . Small RNA coverage of Penelope-like elements is being reported separately (Arkhipova, Rodriguez and Yushenova, in prep.; Arkhipova et al. 2013).

File S1: Small RNA counts and associated metadata for *A. vaga* CDS with AI>0 and for selected multigene families. (.xlsx, 2028 KB)

Available for download as a .xlsx file at:

http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.116.186734/-/DC1/FileS1.xlsx