


ARTICLE

DOI: 10.1038/s42003-018-0141-4

OPEN

PIWI genes and piRNAs are ubiquitously expressed in mollusks and show patterns of lineage-specific adaptation

Julia Jehn¹, Daniel Gebert ¹, Frank Pipilescu¹, Sarah Stern¹, Julian Simon Thilo Kiefer¹, Charlotte Hewel¹ & David Rosenkranz¹

PIWI proteins and PIWI-interacting RNAs (piRNAs) suppress transposon activity in animals, thus protecting their genomes from detrimental insertion mutagenesis. Here, we reveal that *PIWI* genes and piRNAs are ubiquitously expressed in mollusks, similar to the situation in arthropods. We describe lineage-specific adaptations of transposon composition in piRNA clusters in the great pond snail and the pacific oyster, likely reflecting differential transposon activity in gastropods and bivalves. We further show that different piRNA clusters with unique transposon composition are dynamically expressed during oyster development. Finally, bioinformatics analyses suggest that different populations of piRNAs presumably bound to different *PIWI* paralogs participate in homotypic and heterotypic ping-pong amplification loops in a tissue- and sex-specific manner. Together with recent findings from other animal species, our results support the idea that somatic piRNA expression represents the ancestral state in metazoans.

¹Institute of Organismic and Molecular Evolution, Anthropology, Johannes Gutenberg University Mainz, Anselm-Franz-von-Bentzel-Weg 7, 55099 Mainz, Germany. These authors contributed equally: Julia Jehn, Daniel Gebert, Frank Pipilescu. Correspondence and requests for materials should be addressed to D.R. (email: rosenkranz@uni-mainz.de)

In virtually all animals, PIWI proteins protect germ cells from the steady threat of mobile genetic elements, so-called transposons^{1,2}. Based on sequence complementarity to their target transcripts, 23–31 nt non-coding RNAs, termed PIWI-interacting (pi-) RNAs, function as guide molecules for PIWI proteins that slice matching targets through their endonuclease activity. Besides post-transcriptional transposon control, PIWI proteins and piRNAs can trigger the establishment of repressive epigenetic DNA or chromatin modifications, thus inducing efficient transposon silencing on the transcriptional level^{3–6}.

Analyses of piRNA pathways in representatives of many animal taxa have unveiled a great diversity of lineage-specific adaptations, challenging the universal validity of insights obtained from model organisms^{7–19}. For a long time, PIWI proteins and piRNAs were thought to be dispensable for female germ cell development in mammals until it became clear that the model organisms mouse and rat represent an exception from the mammalian rule in that they employ an oocyte specific Dicer isoform for transposon control instead of *Piwil3* which is expressed in the bovine and human female germline^{15,20}. Similarly, evidence for a gene-regulatory role of piRNAs^{14,21–27} and their widespread somatic expression in many animals^{19,28–35} have eroded the dogma that the piRNA pathway is restricted to the germline, being exclusively responsible for silencing of transposons. Indeed, it has been shown that piRNAs are essential for regeneration and stem cell maintenance in the flatworm *Schmidtea mediterranea*²⁸, provide an adaptive immunity against virus infections in *Aedes aegypti*³⁶, are responsible for sex determination in *Bombyx mori*³⁷ and memory-related synaptic plasticity in *Aplysia californica*³⁸.

Despite the likely more than seventy thousand living molluscan species³⁹ there exist only a few functional descriptions of PIWI proteins or piRNAs for this taxon based on experiments in the sea slug *Aplysia californica*³⁸, the Farrer's scallop *Chlamys farreri*⁴⁰ and in the dog whelk *Nucella lapillus*⁴¹. Importantly, Waldron et al.⁴¹ recently showed that piRNA-like small RNAs matching virus and transposon sequences are somatically expressed in *Nucella lapillus*. However, the available data do not allow us to draw any conclusions on whether this represents a conserved or lineage-specific feature of the PIWI/piRNA system within mollusks. In order to further elucidate the evolution of the PIWI/piRNA system in mollusks, we have reconstructed the evolution of *PIWI* genes in this phylum based on 11 sequenced genomes showing that *Piwil1* and *Piwil2* are conserved in mollusks. We perform quantitative real-time PCR experiments to analyze the expression patterns of the identified *PIWI* paralogs across a representative set of tissues from the great pond snail *Lymnaea stagnalis* (*L. stagnalis*) and the pacific oyster *Crassostrea gigas* (*C. gigas*). We apply high-throughput sequencing of small RNAs from *L. stagnalis* to verify the presence of piRNAs in germline and muscle tissue. We further reanalyze published small RNA sequence data from *C. gigas* to characterize the dynamic expression of piRNAs from distinct piRNA clusters during oyster development. Finally, we use bioinformatics approaches to show that different piRNA populations and *PIWI* paralogs participate in the ping-pong amplification loop in a tissue-specific and sex-specific manner.

Results

The molluscan *PIWI* gene repertoire. Many *PIWI* gene tree reconstructions have been published in the past years, however, they do not provide a coherent picture regarding the evolution of *PIWI* genes in early bilaterians. Thus, we first wanted to characterize the *PIWI* protein equipment of sequenced mollusks to infer the ancestral molluscan state and subsequent evolution of

PIWI paralogs within the molluscan clade. To this end, we used available *PIWI* protein sequence data from six molluscan species (*Biomphalaria glabrata*, *Aplysia californica*, *Crassostrea gigas*, *Crassostrea virginica*, *Mizuhopecten yessoensis*, *Octopus bimaculoides*) and further manually annotated *PIWI* genes based on five publicly available but not yet (sufficiently) annotated genomes (*Lymnaea stagnalis*, *Radix auricularia*, *Lottia gigantea*, *Bathymodiolus platifrons*, *Pinctada martensii*). We found that the *PIWI* family members *Piwil1* and *Piwil2* are conserved in mollusks and are orthologous to *Piwil1* and *Piwil2* in vertebrates, suggesting a duplication event in an early bilaterian ancestor prior to the split of protostomes and deuterostomes. According to our results and in consistency with a number of previously published gene trees, *Drosophila* AGO3 shares a common ancestral gene with *Piwil2* clade members^{18,42–44}. However, the insect-specific *PIWI* genes *Piwi* and *Aubergine*, the latter one resulting from a duplication event in dipteran flies^{44,45}, do not group with the *Piwil1* clade (Fig. 1a). It is worth mentioning in this context that different rates of sequence evolution, selective regimes, and gene turnover for Argonaute subfamilies make it difficult to infer their ancient evolutionary history, which is mirrored by numerous published but contradicting *PIWI* gene trees, none of which correctly mirrors the phylogenetic relationship of the included species. Consequently, the presented gene tree reconstruction aims to provide a reliable reconstruction of molluscan *PIWI* gene evolution while the deeper topology should be considered with caution.

While we did not observe further gene duplication events within the molluscan *Piwil2* clade, several duplication events are present in the *Piwil1* clade resulting in two *Piwil1* paralogs in *Bathymodiolus platifrons* and even three *Piwil1* paralogs in *Lymnaea stagnalis* and *Radix auricularia*. Generally, *PIWI* gene duplication events are in line with the previously described erratic evolution of *PIWI* family genes in arthropods^{19,44–46}. Noteworthy, it was also a successive duplication of *Piwil1* on the eutherian lineage that gave rise to *Piwil3* (with subsequent loss on the murine lineage) and *Piwil4*^{47,48} (Fig. 1a).

Expression of *PIWI* genes in *L. stagnalis* and *C. gigas*. To investigate the expression of *PIWI* genes in mollusks we chose two representative species, the pacific oyster *Crassostrea gigas* (*C. gigas*, Bivalvia) showing no *Piwil1* duplication, and the great pond snail *Lymnaea stagnalis* (*L. stagnalis*, Gastropoda), featuring three predicted *Piwil1* paralogs (Fig. 1a). We performed quantitative real-time PCR (qPCR) for each *PIWI* paralog on a representative set of tissues from both species.

For the great pond snail *L. stagnalis* we measured *PIWI* expression on the mRNA level in the hermaphroditic reproductive tract, comprising both male and female gametes, foot muscle, lung, and brain. Relevant expression was detectable for *Piwil1* and particularly *Piwil2*, while the *Piwil1* duplicates *Piwil1b* and *Piwil1c* were only expressed at very low levels (Fig. 1b,c and Supplementary Fig. 1) suggesting a spatiotemporal sub-functionalization. As expected, we observed the highest expression of *Piwil1* and *Piwil2* in the reproductive tract. However, both genes were significantly expressed in the other analyzed tissues as well, reaching 62%, 21%, and 15% of germline expression for *Piwil1* in muscle, lung, and brain respectively, and 36%, 53%, and 12% of germline expression for *Piwil2* in muscle, lung, and brain, respectively (Fig. 1d).

For the dioecious pacific oyster *C. gigas*, *PIWI* mRNA expression was measured in the male gonad, labial palps, gill, adductor muscle, and mantle. We detected significant expression of *Piwil1* and *Piwil2* across all analyzed tissues, particularly in gonadal tissue (Fig. 1e, f), confirming data on *Piwil1* expression in

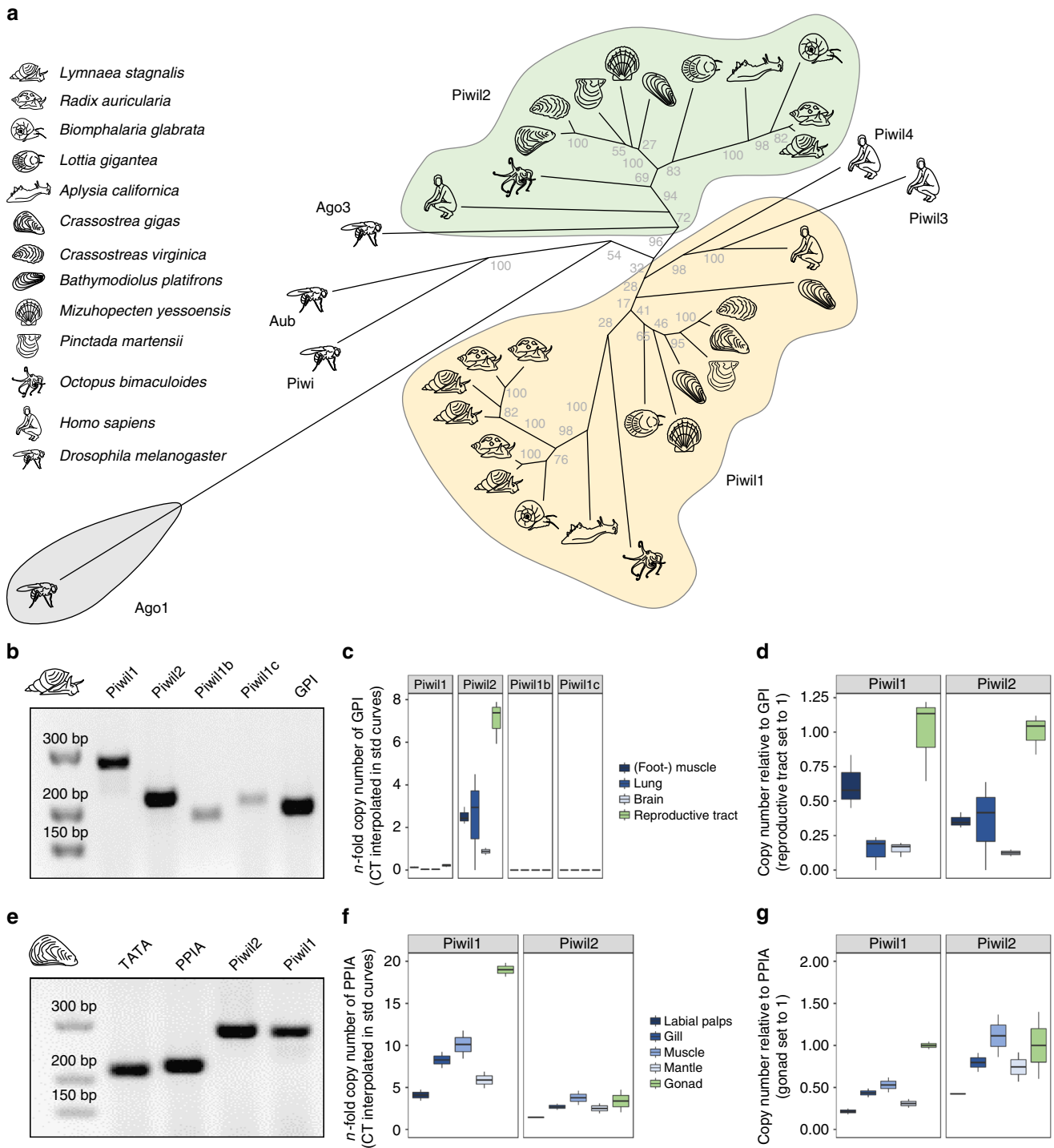


Fig. 1 Evolution and expression of *PIWI* genes in mollusks. **a** *PIWI* gene tree reconstruction of molluskan *PIWI* genes. **b** Control PCR with *PIWI* paralog specific primers and *L. stagnalis* cDNA from the reproductive tract. The complete gel is shown in Supplemental Fig. 1g. **c** qPCR results for *PIWI* paralog expression in different tissues of *L. stagnalis*, measured as *n*-fold expression of the housekeeping gene *GPI*. Center line indicates median, box limits represent the 50th percentile, whiskers show the upper and lower extremes. **d** *PIWI* paralog expression in different tissues of *L. stagnalis*, normalized by the expression of the housekeeping gene *GPI*, values from reproductive tract set to 1. Center line indicates median, box limits represent the 50th percentile, whiskers show the upper and lower extremes. **e** Control PCR with *PIWI* paralog specific primers and *C. gigas* cDNA from the adductor muscle. The complete gel is shown in Supplemental Fig. 1h. **f** qPCR results for *PIWI* paralog expression in different tissues of *C. gigas*, measured as *n*-fold expression of the housekeeping gene *PPIA*. Center line indicates median, box limits represent the 50th percentile, whiskers show the upper and lower extremes. **g** *PIWI* paralog expression in different tissues of *C. gigas*, normalized by the expression of the housekeeping gene *PPIA*, values from male gonad set to 1. Center line indicates median, box limits represent the 50th percentile, whiskers show the upper and lower extremes

the Hong Kong Oyster *Crassostrea hongkongensis*⁴⁹. In relation to gonadal expression, *Piwill1* and *Piwill2* were expressed in levels ranging from 21% (*Piwill1* in labial palps) to 111% (*Piwill2* in adductor muscle, Fig. 1g). The observed expression patterns suggest that a functional PIWI machinery acting in the soma and the germline is conserved in mollusks. Considering the somatic expression of PIWI proteins and piRNAs in many arthropod species¹⁹, it is parsimonious to assume that somatic PIWI/piRNA expression represents the ancestral state that was established in an early protostomian ancestor.

piRNAs in *L. stagnalis* muscle and reproductive tract. In order to characterize molluscan piRNAs, we sequenced small RNA transcriptomes from *L. stagnalis* extracted from the hermaphroditic reproductive tract and (foot-) muscle, since muscle tissue was found to exhibit the highest somatic *PIWI* expression in both *L. stagnalis* and *C. gigas*. Importantly, we want to clarify that we will use the term piRNA bona fide, without formal evidence for physical interaction with PIWI proteins but based on the evidence provided in the following.

The sequence read length profiles for both tissues show a maximum for 21 nt RNAs, with a considerable amount of 22 nt RNAs being present in the muscle, but not in the reproductive tract. We further observed a smaller fraction of RNAs in the range of 24–29 nt in both samples (Fig. 2a). Annotation of sRNA sequences with unitas⁵⁰ revealed a similar proportion of different sRNA classes in each tissue type, with miRNAs accounting for 47% and 53% of reads in the reproductive tract and muscle, respectively (Fig. 2b, Supplementary Table 1). Interestingly, we found a substantial difference in the abundance of tRNA fragments (tRFs). In both samples, 21 nt RNAs derived from the 3' end of tRNAs (3' tRFs, particularly from tRNA-Gly-TCC) constitute the vast majority of tRNA fragments. However, the share of 3' tRFs in the reproductive tract is considerably higher compared to muscle (17 and 10%, respectively, Supplementary Table 1). Recently, 3' tRFs were found to silence long terminal repeat (LTR) retrotransposons in mouse stem cells by targeting their functionally essential and highly conserved primer-binding sites⁵¹. The remarkable amount of 3' tRFs in the analyzed samples supports the idea proposed by Schorn et al.⁵¹ who assume that this mechanism could be highly conserved across different species, providing an innate immunity against LTR propagation.

Focusing on putative piRNAs, we analyzed the fraction of sequence reads that did not match to any other class of non-coding RNA nor mRNA. This dark matter of intergenic sRNAs comprises 27% and 23% of sequence reads in the reproductive tract and in muscle, respectively, and is enriched for transposon sequences, suggesting a role in transposon control (Fig. 2b). Analyses of their sequence read length distribution revealed a prominent class of 22 nt molecules in muscle and to a lesser extent in the reproductive tract, suggesting that transposon defense in *L. stagnalis* involves 22 nt siRNAs in addition to piRNAs (Fig. 2a). To verify the presence of piRNAs, we checked for the so-called ping-pong signature (bias for 10 bp 5' overlap of mapped sequence reads), which is a hallmark of secondary piRNA biogenesis and requires the catalytic activity, and thus expression, of PIWI proteins⁵². Remarkably, we detected a significant ping-pong signature in both, the reproductive tract and muscle (Fig. 2c), suggesting active PIWI/piRNA-dependent transposon silencing in the germline and in the soma. In addition, a ping-pong signature can also be observed for sequence reads that match protein-coding genes, indicating piRNA-dependent gene regulation (Fig. 2c).

Next, we used proTRAC⁵³ to identify 308 piRNA-producing loci in the reproductive tract, and 246 piRNA-producing loci in

muscle tissue. Merging of independently annotated contiguous (<10 kb distance) or overlapping piRNA-producing loci revealed a total of 307 distinct piRNA clusters in *L. stagnalis*, covering 0.27% of the genome (Fig. 2d, Supplementary Data 1). More precisely, all piRNA-producing loci identified in muscle tissue correspond to predicted piRNA clusters based on piRNAs from the reproductive tract, which illustrates that piRNAs in muscle originate from the same set of piRNA clusters compared to the reproductive tract. Nonetheless, there exist 12 clusters whose expression is 14-fold to 36-fold higher in the reproductive tract compared to muscle tissue, while no clusters show muscle-specific expression to a comparable extent. We found that 15.9% of sequence reads from the reproductive tract map to piRNA clusters, while only 6.7% of sequence reads from muscle do so, indicating rather moderate production of primary piRNAs in the soma compared to the germline (Fig. 2e). Besides the presence of primary piRNAs, we found that the number of piRNAs that participate in ping-pong amplification (measured as ping-pong reads per million bootstrapped reads, ppr-mbr) is slightly higher in muscle (~39k ppr-mbr) compared to the situation in the reproductive tract (~35k ppr-mbr), suggesting higher amounts of secondary piRNAs and emphasizing the functional importance of somatic PIWI/piRNA expression (Fig. 2e). In line with the transposon-suppressive role of piRNAs, the identified piRNA clusters show a twofold enrichment for transposon sequences compared to the whole genome situation (59 and 31%, respectively, Fig. 2f, g), whereas only 1.7% of piRNA cluster sequence represents protein-coding sequence. Interestingly, the transposon composition in piRNA clusters does not at all reflect the transposon landscape of the genome. Instead, piRNA clusters are enriched for Gypsy retrotransposons and particularly DNA transposons such as Kolobok, hAT5, or hATw showing up to 108-fold enrichment in piRNA clusters (Fig. 2g, h). This non-random distribution suggests a selective regime that favors insertion events of transposons with low divergence from their consensus sequence, likely representing evolutionary young and active elements.

Ubiquitous and dynamic expression of piRNAs in *C. gigas*.

Based on our observation that *PIWI* genes and piRNAs are expressed in the soma and the germline of *L. stagnalis*, we re-analyzed previously published small RNA data sets from *C. gigas* that were used to investigate the dynamic expression of miRNAs during oyster development without further examination of a putative piRNA fraction⁵⁴ (NCBI Sequence Read Archive Project ID SRP007591). We annotated *C. gigas* sRNAs from the male and female gonad, different developmental stages ranging from the egg to juvenile, and a representative set of somatic tissues from adult animals (Supplementary Table 2). In all data sets, particularly in gonadal tissues, eggs, and early embryo stages but also in hemolymph we detected a large amount of sequence reads that did not match to any known ncRNA class but was instead enriched for transposon sequences. The transposon-matching sub-fraction itself was enriched for antisense sequences (Supplementary Table 2). Analogous to the procedure applied for the *L. stagnalis* data sets, we verified the presence of primary and secondary piRNAs by analyzing the ping-pong signature of each data set. Remarkably, we detected a significant ping-pong signature across all analyzed data sets (Fig. 3a, Supplementary Fig. 2), but also found that the number of ping-pong reads (measured as ppr-mbr) differs considerably depending on the tissue and developmental stage (Fig. 3a, Supplementary Fig. 3). Noteworthy, as is the case with *L. stagnalis*, a ping-pong signature is also detectable when taking only those reads into account that match protein-coding sequences, suggesting a

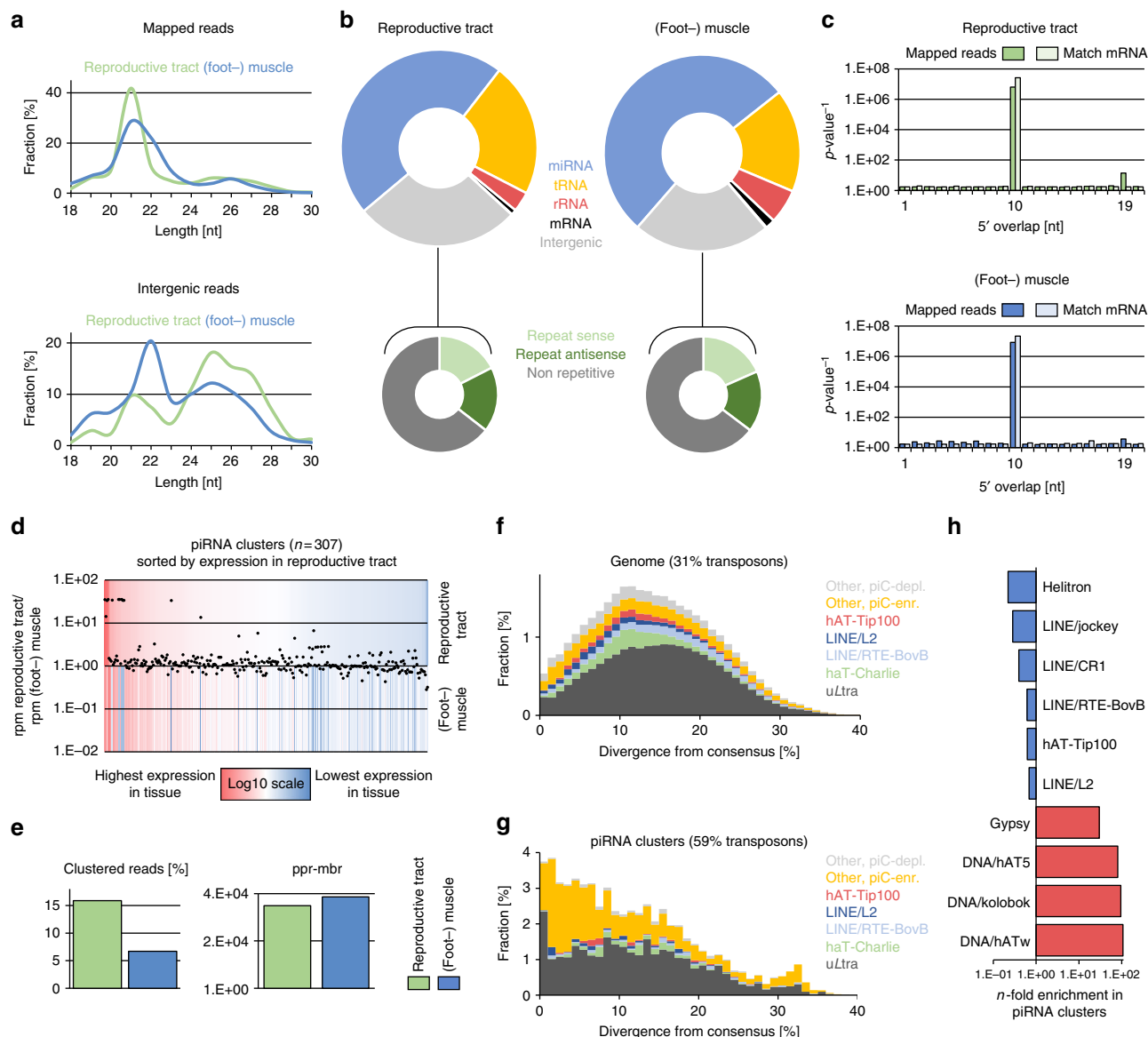


Fig. 2 Characterization of small RNAs from *L. stagnalis* (foot-) muscle and reproductive tract. **a** Sequence read length distribution of mapped (top) and unannotated (intergenic) reads (bottom). **b** Results from small RNA annotation with unitas (top) and transposon content of intergenic reads (bottom). **c** Ping-pong signature. P -values are deduced from the corresponding Z -scores. P -values for all reads and reads that match mRNA are shown. **d** Differential expression of 307 predicted piRNA clusters. Colors refer to expression relative to highest/lowest expression within one tissue. Dots indicate n -fold expression of a given cluster in reproductive tract relative to muscle. **e** Amount of clustered reads and ping-pong reads per million bootstrapped reads (ppr-mbr). **f** Representation of transposons in the genome of *L. stagnalis*, plotted by divergence [%] from transposon consensus. **g** Representation of transposons within piRNA clusters of *L. stagnalis*, plotted by divergence [%] from transposon consensus. **h** Prominent transposons that are enriched or depleted in *L. stagnalis* piRNA clusters

relevant and conserved role of the PIWI/piRNA pathway in post-transcriptional regulation of protein-coding genes in gonads, egg, blastula, digestive gland, and hemolymph (Supplementary Table 3). We further used sequences without ncRNA annotation to predict piRNA clusters with proTRAC (Supplementary Data 2) and checked whether we can observe a differential expression of specific piRNA clusters in time and space (Fig. 3a).

In contrast to the situation in *L. stagnalis*, we found that different genomic loci are responsible for production of primary piRNAs in the germline and in the soma, but also during different developmental stages, which is similar to the situation in the sea anemone *Nematostella vectensis*¹⁸ and the German cockroach *Blattella germanica*⁵⁵. A clustering approach based on average linkage⁵⁶ revealed four distinct groups of piRNA clusters which

we named class 1–4 piRNA clusters (Fig. 3a). Class 1 piRNA clusters are active in the adult germline (male and female) and in the early embryo until the D-shaped veliger stage where larvae are ~14 h old. The same applies to class 2 piRNA clusters, however, following the D-shape veliger stage, class 1 piRNA clusters become inactive, while class 2 piRNA clusters remain active and class 3 piRNA clusters start piRNA production. Both, class 2 and class 3 piRNA cluster activity is measurable until the juvenile stage, where oysters are ~20 days old. In somatic tissues of adult oysters, class 4 piRNA clusters represent the main source of primary piRNAs (Fig. 3a, bottom). Interestingly, all four classes of piRNA clusters are active in hemocytes, which also feature the highest amount of clustered reads, and ping-pong reads compared to other somatic tissues. This might reflect the

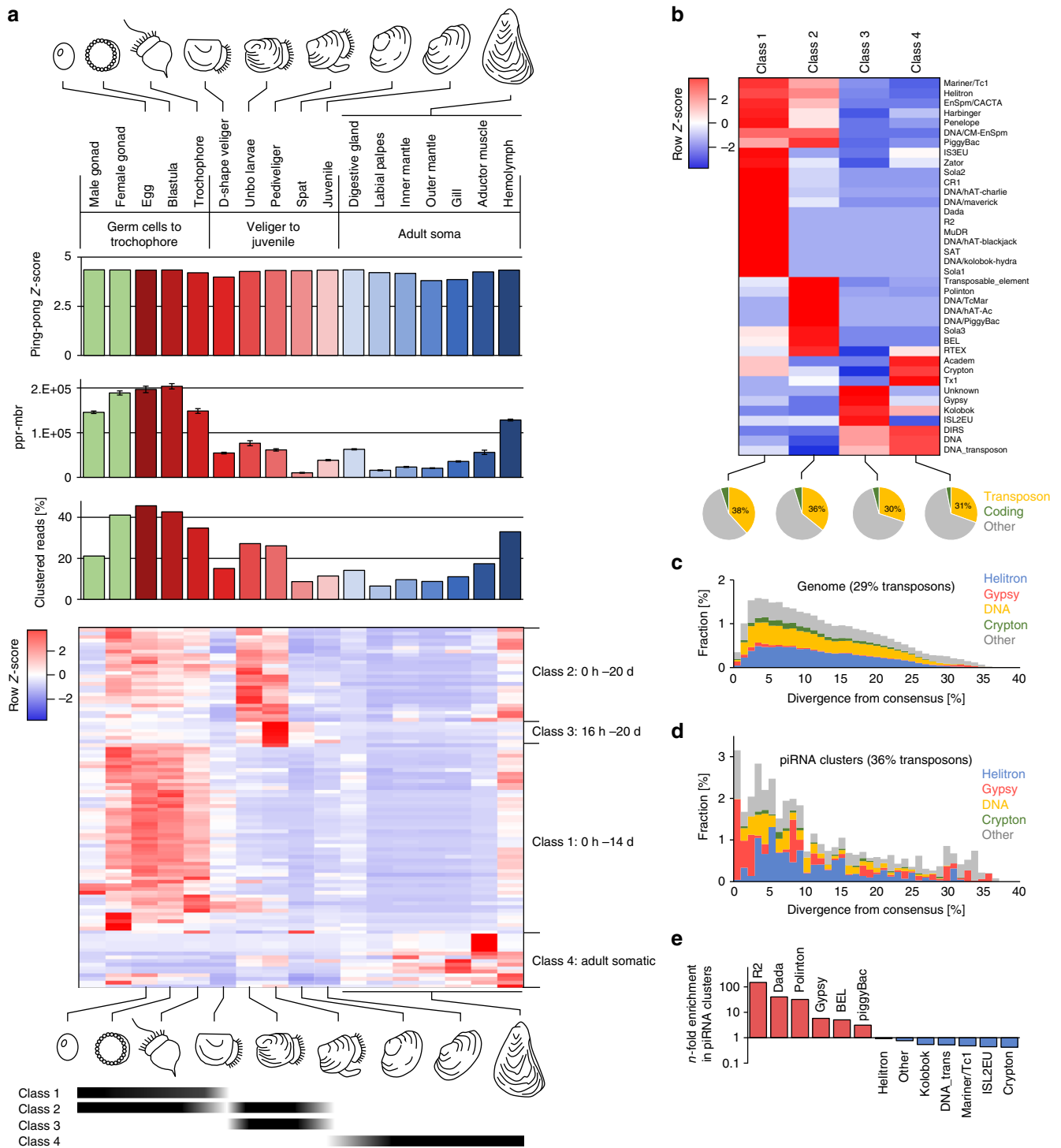


Fig. 3 Characterization of small RNAs and piRNA clusters from different *C. gigas* samples. **a** Sequence reads without annotation produce a significant ping-pong signature (top row of bars, only Z-scores for 10 bp 5' overlap are shown). The number of ping-pong reads per million bootstrapped reads (middle row of bars), and the number of clustered reads (bottom row of bars) differs considerably across the samples. Heatmap shows the differential expression of the top 100 piRNA clusters in terms of maximum rpm coverage. Different classes of piRNA clusters are expressed during oyster development and in adult somatic tissues (bottom). Error bars indicate standard deviation. **b** Transposon composition of piRNA clusters belonging to four different classes. **c** Representation of transposons in the genome of *C. gigas*, plotted by divergence [%] from transposon consensus. **d** Representation of transposons within piRNA clusters of *C. gigas*, plotted by divergence [%] from transposon consensus. **e** Prominent transposons that are enriched or depleted in *C. gigas* piRNA clusters

presence of stem cells within the hemocyte cell population, which are subject to complex differentiation processes^{57,58}.

Interestingly, the four classes of piRNA clusters differ considerably regarding the overall transposon content as well as the specific transposon composition (Fig. 3b–d). Class 1 and class

2 piRNA clusters are generally enriched for transposon sequences showing 38 and 36% transposon-derived sequences, respectively, compared to a genomic transposon content of 29%. The surprisingly high accumulation of young (as deduced from the divergence from their consensus) Gypsy elements in piRNA

clusters, suggests a strong selection for Gypsy element insertions, probably as a consequence of Gypsy activity in *C. gigas*. Noteworthy, the accumulation of young transposons in molluscan piRNA clusters sharply contrasts the situation in *Drosophila* and human, where older transposons are more abundant in piRNA-producing loci^{59,60}. Considering transposons that are generally enriched in piRNA clusters, we found that R2 retrotransposons (149-fold enrichment in piRNA clusters) and Dada DNA transposons (40-fold enrichment in piRNA clusters) are most abundant in class 1 piRNA clusters (Fig. 3e). In contrast, Polinton DNA transposons (32-fold enrichment in piRNA clusters) and BEL retrotransposons (fivefold enrichment in piRNA clusters) are most abundant in class 2 piRNA clusters. Different from class 1 and class 2 piRNA clusters, class 3, and class 4 piRNA clusters display only slight transposon enrichment (30 and 31%, respectively). Noteworthy, high copy number Gypsy retrotransposons (fivefold enrichment in piRNA clusters) are most abundant in class 3 piRNA clusters, while Academ, Crypton, and Tx1 transposons are most abundant in class 4 piRNA clusters.

The fact that different piRNA clusters are expressed in the germline (class 1 and class 2) and in adult somatic tissues (class 4) of *C. gigas* contrasts with the situation in *L. stagnalis*, where identical piRNA-producing loci are active in the germline and in the soma. Moreover, we can observe considerable differences in the transposon composition of piRNA clusters in the two species, which likely reflect a divergent transposon activity in gastropods and bivalves, resulting in varying selective constraints on the different phylogenetic lineages.

Homotypic and heterotypic ping-pong amplification. The ping-pong amplification loop describes a process that is responsible for the post-transcriptional silencing of transposable elements⁵². In *Drosophila* and mouse, this process typically involves two PIWI paralogs (heterotypic ping-pong), one loaded with antisense piRNAs targeting transposon transcripts, and the other loaded with sense piRNAs targeting piRNA cluster transcripts, which contain transposon sequences in antisense orientation^{61,62}. Likely for steric reasons, premature piRNAs loaded onto the different PIWI paralogs are more or less rigorously trimmed at their 3' ends. This is why piRNA populations bound to different PIWI paralogs not only differ regarding the amount of sense-transposon and antisense-transposon sequences, but also in their sequence length profiles^{52,63,64}. In addition to the heterotypic ping-pong amplification, homotypic ping-pong has been shown to occur in *gin* mutant flies (Aub:Aub⁶⁵), and wild-type prenatal mouse testis (Miwi2:Miwi2, Mili:Mili⁶²).

Since the typical molluscan genome encodes two ubiquitously expressed PIWI paralogs, *Piwil1* and *Piwil2*, we asked whether we can provide evidence for the participation of distinct piRNA populations and PIWI paralogs in the ping-pong cycle. We conducted a bioinformatics approach under the premise that *Piwil1*-bound and *Piwil2*-bound piRNAs exhibit different length profiles, which is the case for the corresponding mouse homologs *Piwil1* (Miwi) that preferentially binds 29/30 nt piRNAs, and *Piwil2* (Mili) which preferentially binds 26/27 nt piRNAs⁶⁶. A similar, yet not equally pronounced, difference between *Piwil1*-bound (Ziwi) and *Piwil2*-bound (Zili) piRNAs also exists in zebrafish, suggesting the evolutionary conservation of this pattern⁸. We analyzed pairs of mapped *C. gigas* and *L. stagnalis* sequence reads that showed a 10 bp 5' overlap (ping-pong pairs), with respect to the sequence length of each ping-pong partner (Fig. 4, Supplementary Fig. 4). In the female gonad of *C. gigas*, most ping-pong pairs combine piRNAs with a length of 25 nt and 29 nt (Fig. 4a), suggesting heterotypic *Piwil1*–*Piwil2*-dependent

ping-pong amplification as depicted in Fig. 4b. In support of this, 29 nt piRNAs, presumably bound to *Piwil1*, are heavily biased for a 5' uridine (a hallmark of primary piRNAs), whereas 25 nt piRNAs, presumably bound to *Piwil2*, show a stronger bias for an adenine at position 10 (typical for secondary piRNAs). In contrast, ping-pong pairs in *C. gigas* muscle predominantly combine two 29 nt piRNAs, suggesting homotypic, *Piwil1*-dependent ping-pong amplification (Fig. 4b). Generally, the observed patterns of ping-pong pairs are very diverse across the different samples, for instance displaying heterotypic ping-pong in the digestive gland and homotypic *Piwil2*-dependent ping-pong in hemolymph cells (Supplementary Fig. 4).

Since the expression of *Piwil1* compared to *Piwil2* is considerably lower in *L. stagnalis*, we were curious to check whether the corresponding ping-pong pairs might reflect this fact. Indeed, 26/26 nt pairs (homotypic, *Piwil2*-dependent ping-pong) represent the majority of ping-pong pairs in the reproductive tract (Fig. 4a). In addition, homotypic *Piwil2*-dependent ping-pong amplification with 24/25 nt ping-pong pairs is also dominant in the *L. stagnalis* muscle (Fig. 4b). However, we also observed differences in ping-pong patterns that do not correlate with the measured mRNA levels of *Piwil1* and *Piwil2*. For example, our data suggests homotypic *Piwil2*-dependent ping-pong amplification in the oyster gill but homotypic *Piwil1*-dependent ping-pong amplification in the oyster muscle (Supplementary Fig. 4), while both tissues display a very similar expression of both *PIWI* paralogs on the mRNA level (Fig. 1f). Thus, we assume that factors other than mere *PIWI* expression critically influence characteristics of the ping-pong amplification loop.

Moreover, we clearly cannot rule out the possibility that binding preferences of *PIWI* paralogs have changed on the molluscan lineage and are different from those observed in fly, fish, and mouse. This could mean that length profiles of piRNAs associated to each of the molluscan *PIWI* paralogs might be exactly reciprocal compared to our presumption. One could even speculate that both *PIWI* paralogs may bind the whole range of piRNAs, which is not possible to disprove without performing corresponding co-immunoprecipitation experiments. However, based on the presence of piRNA populations with different length profiles (Fig. 2a), their representation in ping-pong pairs together with the differences in their amount of 1U and 10A reads (Fig. 4a), we believe that the above made interpretations are a reasonable and parsimonious interpretation of the data at hand, yet not the only possible one.

Discussion

Our results reveal that mollusks utilize the *PIWI*/piRNA pathway as a defense against transposable elements in the germline and in the soma, which corresponds to the situation in arthropods and therefore suggests somatic *PIWI*/piRNA expression to represent a plesiomorphic protostomian character state. In fact, available data from deeper branching metazoans such as poriferans and cnidarians supports the view that this system was established in the soma even long before the split of protostomes and deuterostomes^{7,18,41}. In addition, based on the observation that a substantial fraction of arthropod and mollusk piRNAs targets messenger RNAs producing the generic ping-pong signature, it seems likely that the last common ancestor of arthropods and mollusks applied the *PIWI*/piRNA pathway also for post-transcriptional regulation of protein-coding genes. Recently, the Xenacoelomorpha phylum, a group of marine worms that were previously thought to belong to the Platyhelminthes clade, was found to represent the sister group of Nephrozoa which comprise protostomes and deuterostomes^{67,68}. Presently, piRNAs for this

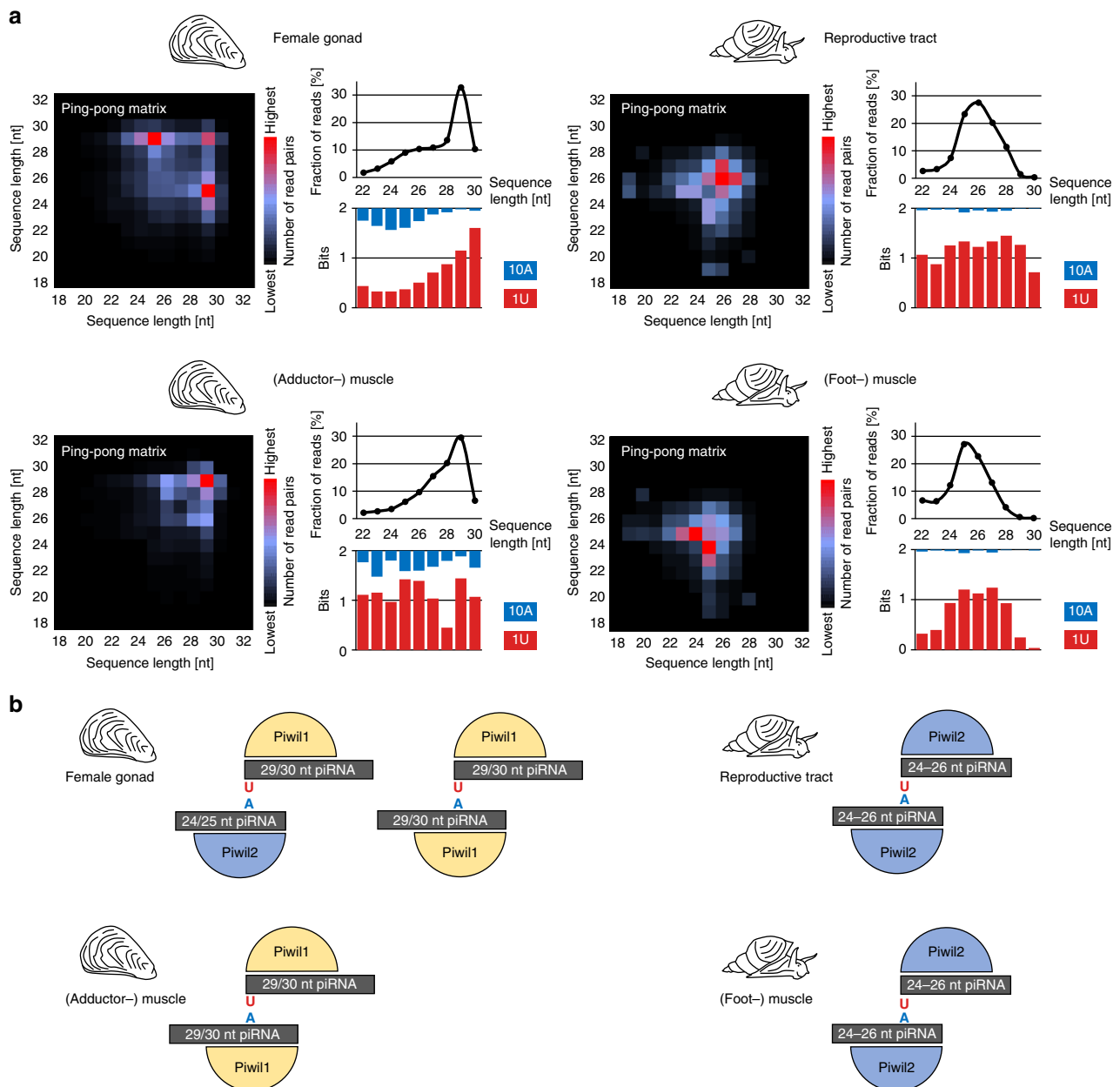


Fig. 4 Analysis of piRNAs that participate in the ping-pong amplification loop. **a** Ping-pong matrices illustrate frequent length-combinations of ping-pong pairs (sequences with 10 bp 5' overlap). Sequence read length distribution and 1U/10A bias [bits] for ping-pong sequences are shown. **b** Proposed model of ping-pong amplification in the germline and muscle of *C. gigas* and *L. stagnalis*

outgroup are not characterized but having such data would doubtlessly provide valuable insights and allow to draw conclusions regarding the function of the PIWI/piRNA system in the last common ancestor of all bilaterians, particularly with respect to an ancestral gene-regulatory role. Especially with regard to the latter, functional studies in non-model organisms are urgently needed since the pure bioinformatical evidence for piRNA-dependent processing of protein-coding genes does not give any information on its factual biological relevance this process might have in different species. In vertebrates, somatic PIWI/piRNA expression appears to have faded away and reports on somatically expressed piRNAs in mammals are often considered with skepticism for good reasons⁶⁹. However, remnants of the former somatic expression might have outlasted to fulfill special functions in specific cells and/or in narrowly defined timespans of

development or cell differentiation in the one or the other clade. Our results indicate that studying the PIWI/piRNA pathway in organisms outside of the main experimental models of *Drosophila* and mouse is necessary to fully understand its evolution and functions.

Methods

PIWI gene annotation and tree reconstruction. In order to reconstruct the phylogenetic relations of mollusk Piwi proteins, we first searched for *PIWI* genes in species with an available genome sequence that lack proper annotation (*Lymnaea stagnalis*, *Radix auricularia*, *Lottia gigantea*, *Bathymodiolus platifrons*, *Pinctada martensii*). To this end, we scanned the relevant genomes for sequences that are homologous to annotated *PIWI* paralogs of the pacific oyster (EKC35279 and EKC29295) by aligning translated DNA sequences using tblastx (v2.7.1+⁷⁰). Neighboring hits with a distance smaller than 10 kb were grouped as exons of distinct gene loci. Only groups containing the overall best hits for a given locus

were retained. Finally, the predicted gene sequences were checked for presence of PIWI and PAZ domains using NCBI conserved domain database⁷¹. Similarly, for *PIWI* expression analysis by qPCR in the pond snail, we identified the housekeeping gene *GPI* (glucose-6-phosphate isomerase) by comparison with the human ortholog (ARJ36701).

The predicted and annotated PIWI protein sequences of the 11 available molluscan species together with PIWI paralogs of human (Piwi1–4) and fly (Ago3, Piwi, Aub), as well as fly argonaute Ago1 were aligned using MUSCLE (v.3.8.31⁷²). Subsequently, the resulting protein alignment was curated with Gblocks (v.0.91b), allowing smaller final blocks with gap positions and less strict flanking positions. Using ModelGenerator (v.0.85⁷³), we determined LG + G + F⁷⁴ to be the best-fitting model of substitution for our data. The curated alignment (Supplementary Data 3) was then used for phylogenetic tree reconstruction with PhyML (v3.1⁷⁵), applying approximate likelihood-ratio test (SH-like) and LG substitution model, including empirical gamma distribution (G) and character frequencies (F). Support values were generated by bootstrap with 100 replicates.

qPCR. Experiments were performed on commercially available *C. gigas* animals from the western French Atlantic coast (Ile d'Oleron) and captured wild living *L. stagnalis* animals from South-western Germany (Heppenheim). To estimate the expression of the *Piwi* homologs in several tissues of *L. stagnalis* and *C. gigas* we performed qPCR with cDNA synthesized from the total RNA fraction of these tissues. Total RNA was isolated with TriReagent and the polyadenylated transcriptome was reversely transcribed with SuperScript IV using the RT-primer 5'-CGAATTCTAGAGCTCGAGGCGAGGACATGT₂₅VN-3'. Primers amplifying ~200 bp long products of the respective *Piwi* homologs and housekeeping genes were designed with the NCBI tool primer-BLAST on basis of the *L. stagnalis* genome assembly GCA_900036025.1 v1.0 and the *C. gigas* genome assembly GCA_000297895.1 oyster_v9. To prevent amplification of residual genomic DNA, primers were designed to be exon-junction spanning or to span at least several intronic regions. The respective biological replicates were analyzed as technical duplicates on a Corbett Rotor-Gene 6000 real-time PCR cyclers and the copy numbers of the genes of interest were quantified by standard curves of the individual primer pair amplicons. For each cDNA sample the calculated *Piwi* copy numbers were relativized by the calculated copy numbers of the housekeeping genes to calibrate for variabilities in sample preparation. These *n*-fold expression values were finally used to calculate the mean and standard deviation of the replicates. For an improved visualization, the *n*-fold expression values of each *Piwi* homolog are additionally displayed as a percentage of the respective gonad value.

Small RNA extraction and sequencing. We extracted total RNA from *L. stagnalis* reproductive tract (incl. ovotestis, oviduct, spermatheca, spermiduct, prostate, uterus, vagina, vas deferens) and foot muscle, and total RNA from *C. gigas* adductor muscle and gonadal tissue with TriReagent according to the manufacturer's instructions. For each species we sampled two different individuals per tissue. The small RNA fractions of each obtained total RNA sample were sequenced at BGI, Hong Kong, on a BGISEQ-500 unit. Small RNA sequence data sets for *L. stagnalis* and *C. gigas* are deposited at NCBI's Sequence Read Archive (SRA) and can be accessed under the SRA project IDs SRP130729 and SRP130745. We further used previously published small RNA sequence data from *C. gigas*⁵⁴ to analyze piRNA expression and characteristics with respect to different developmental stages.

Repeat annotation. We performed de novo prediction of repetitive elements in the genome of *L. stagnalis* with RepeatScout (v. 1.0.5⁷⁶). Predicted repetitive elements were classified with RepeatClassifier which is part of the RepeatModeler (v. 1.0.11) package. Transposons that failed to be classified based on known transposons from other species are referred to as unclassified *Lymnaea*-specific transposons (uLtra). The resulting repeat sequences, as well as a complete collection of currently available molluscan repeat sequences from RepBase⁷⁷ were used as reference sequences for repeat masking of the *L. stagnalis* and *C. gigas* genomes with RepeatMasker (v. 4.0.7) using the cross_match search engine and the option -s for most sensitive masking. Annotated repeats in the RepeatMasker output were analyzed with respect to transposon families and divergence from their consensus sequence using the Perl script TE_landscape.pl. Analysis was conducted with the entire repeat data set as well as with repeats localized in predicted piRNA clusters. TE_landscape.pl is freely available at <https://sourceforge.net/projects/protrac/files/tools/>.

Gene annotation. We performed de novo gene annotation of the *L. stagnalis* genome assembly gLs_1.0⁷⁸ using the MAKER genome annotation pipeline (v.2.31.8) in order to identify sRNAs that match protein-coding sequences⁷⁹. Initially, we masked the *L. stagnalis* genome with WindowMasker⁸⁰ using default settings including the duster option to mask low-complexity regions. Then, we used available molluscan cDNA data from Ensembl database (release 92) and available mRNA and protein data from *L. stagnalis* deposited at NCBI (Effective April 25, 2018) as input for MAKER. MAKER output files for separate scaffolds were merged using the Perl script mergeMAKERoutput.pl which is freely available at <https://sourceforge.net/projects/protrac/files/tools/>. The complete genome

annotation in GFF3 format and a corresponding mRNA sequence file in FASTA format are available as Supplementary Data 4 and Supplementary Data 5.

Processing and annotation of small RNA sequence data. Small RNA sequence data sets were collapsed to non-identical sequences, retaining information on sequence read counts using the Perl script *collapse*. Sequences > 36nt were rejected using the Perl script *length-filter*. Finally, low-complexity sequences were filtered using the Perl script *duster* with default parameters. All Perl scripts mentioned are part of the NGS toolbox⁸¹.

We then applied a customized mapping strategy of the remaining small RNA sequence reads based on the consideration that our data sets presumably contain considerable amounts of transposon-derived piRNAs as well as post-transcriptionally edited (e.g., A-to-I) or tailed miRNAs and piRNAs. Genomic mapping was performed with SeqMap⁸² using the option/output_all_matches and allowing up to three mismatches. The obtained alignments were further filtered using the Perl script *seqmap_filter.pl* that is freely available at <https://sourceforge.net/projects/protrac/files/tools/>. For the final alignments we allowed up to two non-template 3' nucleotides and up to one internal mismatch. For each sequence, we only considered the best alignments in terms of mismatch counts, but did not reject alignments with equal quality in case of multiple mapping sequences. Sequences that did not produce at least one valid alignment to the reference genome were rejected.

To improve small RNA sequence annotation, we performed de novo tRNA, rRNA, and miRNA prediction based on the available reference genome assemblies gLs_1.0 (*L. stagnalis*) and GCA_000297895.1 oyster_v9 (*C. gigas*). tRNA annotation was performed with a local copy of tRNAscan (v.1.3.1⁸³). Only tRNAs with <5% N's were taken for further analysis. rRNA sequences were predicted using a local copy of RNAmmer (v.1.2⁸⁴) and hmmer (v.2.2 g⁸⁵). Both tools were run with default parameters. We pooled small RNA sequence reads from different replicates and tissues for each species separately to perform miRNA *de novo* prediction with ShortStack (v.3.8.4⁸⁶) using default parameters. The predicted tRNA, rRNA, and miRNA precursor sequences, as well as previously published miRNA precursor sequences^{54,87,88}, were used as additional reference sequences for small non-coding RNA annotation with unitas (v.1.4.6⁵⁰) which was run with the option -riborase. For *L. stagnalis*, we also included predicted cDNA data based on MAKER annotation (see above). sRNA sequences that did not match to any ncRNA or mRNA of *C. gigas* or *L. stagnalis* were blasted against NCBI nucleotide collection (nr) to search for possible contaminants of parasitic species. Sequences that produced better alignments to genomes of species that possibly parasitized the sampled individuals (*Dicrocoelium*, *Legionella*, *Panagrellus*, *Thelazia*, *Trichobilharzia*) were considered as contaminants and not used for downstream analyses.

piRNA cluster identification. Sequences that did not produce a match to known non-coding RNAs were considered as putative piRNAs and were used for prediction of piRNA clusters with proTRAC (v. 2.4.0⁵³) applying default settings. piRNA clusters were predicted for each data set and species separately. The resulting piRNA cluster predictions for each species were condensed, merging clusters with <10 kb distance from each other using the Perl script *merge_clusters* which is freely available at <https://sourceforge.net/projects/protrac/files/tools/>. To preclude false positive annotation of, e.g., tRNA or rRNA genes as piRNA clusters, we validated predicted piRNA clusters by analyzing sRNA reads that mapped to them with respect to their relation to mRNA or other ncRNA classes (Supplementary Fig. 5a). To further check whether piRNA cluster calling may underestimate or overestimate the number of primary piRNAs in our data sets, we performed an arithmetical approach to estimate the fraction of genuine primary piRNAs based on the fraction of 5' U reads in annotated and non-annotated reads with 24–29 nt length which yields results very close to the number of clustered reads (Supplementary Methods, Supplementary Fig. 5b). We calculated the sequence read coverage (rpm) for each of the resulting piRNA clusters per data set. For *C. gigas* piRNA clusters, a heat map for the top 100 piRNA clusters in terms of maximum rpm coverage (accounting for 64% of summed rpm values) was constructed with Heatmapper⁵⁶ applying Pearson distance and average linkage clustering. Finally, predicted piRNA clusters were analyzed with respect to their repeat and gene content using the Perl script *pic_content.pl* which is freely available at <https://sourceforge.net/projects/protrac/files/tools/>.

Ping-pong quantification. In order to compare ping-pong signatures across multiple data sets with different sequencing depth, we constructed a software tool, PPMeter (v.0.4), that creates bootstrap pseudo-replicates from original data sets and subsequently analyzes the ping-pong signature and number of ping-pong sequence reads of each pseudo-replicate (default: 100 pseudo-replicates each comprising one million sequence reads). The obtained parameters 'ping-pong score per million bootstrapped reads' (pps-mbr) and 'ping-pong reads per million bootstrapped reads' (ppr-mbr) can be used for quantification and direct comparison of ping-pong activity in different small RNA data sets. The software is freely available at <http://www.smallRNAGroup.uni-mainz.de/software.html> and <https://sourceforge.net/projects/protrac/files/tools/>.

Code availability. Source code of software that has been written for data processing and analysis is freely available at <https://sourceforge.net/projects/protrac/files/tools/>.

Data availability

Sequence data have been uploaded to NCBI's Sequence Read Archive and can be accessed via the accessions SRP130729 and SRP130745.

Received: 21 March 2018 Accepted: 17 August 2018

Published online: 07 September 2018

References

- Thomson, T. & Lin, H. The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annu. Rev. Cell Dev. Biol.* **25**, 355–376 (2009).
- Iwasaki, Y. W., Siomi, M. C. & Siomi, H. PIWI-interacting RNA: its biogenesis and functions. *Annu. Rev. Biochem.* **84**, 405–433 (2015).
- Reuter, M. et al. Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature* **480**, 264–267 (2011).
- Di Giacomo, M. et al. Multiple epigenetic mechanisms and the piRNA pathway enforce LINE1 silencing during adult spermatogenesis. *Mol. Cell* **50**, 601–608 (2013).
- Pezić, D., Manakov, S. A., Sachidanandam, R. & Aravin, A. A. piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. *Genes Dev.* **28**, 1410–1428 (2014).
- Manakov, S. A. et al. MIWI2 and MILI have differential effects on piRNA biogenesis and DNA methylation. *Cell Rep.* **12**, 1234–1243 (2015).
- Grimson, A. et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**, 1193–1197 (2008).
- Houwing, S., Berezikov, E. & Ketting, R. F. Zili is required for germ cell differentiation and meiosis in zebrafish. *EMBO J.* **27**, 2702–2011 (2008).
- Das, P. P. et al. 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 (2008).
- Li, X. Z. et al. An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Mol. Cell* **50**, 67–81 (2013).
- Lim, R. S., Anand, A., Nishimiya-Fujisawa, C., Kobayashi, S. & Kai, T. Analysis of Hydra PIWI proteins and piRNAs uncover early evolutionary origins of the piRNA pathway. *Dev. Biol.* **386**, 237–251 (2014).
- Ha, H. et al. A comprehensive analysis of piRNAs from adult human testes and their relationship with genes and mobile elements. *BMC Genomics* **15**, 545 (2014).
- Hirano, T. et al. Small RNA profiling and characterization of piRNA clusters in the adult testes of the common marmoset, a model primate. *RNA* **20**, 1223–1237 (2014).
- Gebert, D., Ketting, R. F., Zischler, H. & Rosenkranz, D. piRNAs from pig testis provide evidence for a conserved role of the piwi pathway in post-transcriptional gene regulation in mammals. *PLoS ONE* **10**, e0124860 (2015).
- Roovers, E. F. et al. Piwi proteins and piRNAs in mammalian oocytes and early embryos. *Cell Rep.* **10**, 2069–2082 (2015).
- Rosenkranz, D., Rudloff, S., Bastuck, K., Ketting, R. F. & Zischler, H. Tupaia small RNAs provide insights into function and evolution of RNAi-based transposon defense in mammals. *RNA* **21**, 911–922 (2015).
- Madison-Villar, M. J., Sun, C., Lau, N. C., Settles, M. L. & Mueller, R. L. Small RNAs from a Big Genome: the piRNA pathway and transposable elements in the salamander species *Desmognathus fuscus*. *J. Mol. Evol.* **83**, 126–136 (2016).
- Praher, D. et al. Characterization of the piRNA pathway during development of the sea anemone *Nematostella vectensis*. *RNA Biol.* **14**, 1727–1741 (2017).
- Lewis, S. H. et al. Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements. *Nat. Ecol. Evol.* **2**, 174–181 (2018).
- Flemer, M. et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* **155**, 807–816 (2013).
- Zhang, P. et al. MIWI and piRNA-mediated cleavage of messenger RNAs in mouse testes. *Cell Res.* **25**, 193–207 (2015).
- Russell, S. et al. Bovine piRNA-like RNAs are associated with both transposable elements and mRNAs. *Reproduction* **153**, 305–318 (2017).
- Rouget, C. et al. Maternal mRNA deadenylation and decay by the piRNA pathway in the early *Drosophila* embryo. *Nature* **467**, 1128–1132 (2010).
- Gou, L.-T. et al. Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Res.* **24**, 680–700 (2014).
- Watanabe, T. & Lin, H. Posttranscriptional regulation of gene expression by Piwi proteins and piRNAs. *Mol. Cell* **56**, 18–27 (2014).
- Barckmann, B. et al. Aubergine iCLIP reveals piRNA-dependent decay of mRNAs involved in germ cell development in germ cell development in the early embryo. *Cell Rep.* **12**, 1205–1216 (2015).
- Rojas-Ríos, P., Chartier, A., Pierson, S. & Simonelig, M. Aubergine and piRNAs promote germline stem cell self-renewal by repressing the proto-oncogene *Cbl*. *EMBO J.* **36**, 3194–3211 (2017).
- Palakodeti, D., Smielewska, M., Lu, Y. C., Yeo, G. W. & Graveley, B. R. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* **14**, 1174–1186 (2008).
- Perrat, P. N. et al. Transposition-driven genomic heterogeneity in the *Drosophila* brain. *Science* **340**, 91–95 (2013).
- Nandi, S. et al. Roles for small noncoding RNAs in silencing of retrotransposons in the mammalian brain. *Proc. Natl Acad. Sci. USA* **113**, 12697–12702 (2016).
- Jones, B. C. et al. A somatic piRNA pathway in the *Drosophila* fat body ensures metabolic homeostasis and normal lifespan. *Nat. Commun.* **7**, 13856 (2016).
- Teixeira, F. K. et al. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature* **552**, 268–272 (2017).
- Ross, R. J., Weiner, M. M. & Lin, H. PIWI proteins and PIWI-interacting RNAs in the soma. *Nature* **505**, 353–359 (2014).
- Juliano, C. E. et al. PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells. *Proc. Natl Acad. Sci. USA* **111**, 337–342 (2013).
- Funayama, N., Nakatsukasa, A. M., Mohri, K., Masuda, Y. & Agata, K. Piwi expression in archeocytes and choanocytes in demosponges: insights into the stem cell system in demosponges. *Evol. Dev.* **12**, 275–287 (2010).
- Miesen, P., Girardi, E. & van Rij, R. P. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in *Aedes aegypti* mosquito cells. *Nucleic Acids Res.* **43**, 6545–6556 (2015).
- Kiuchi, T. et al. A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**, 633–636 (2014).
- Rajasethupathy, P. et al. A role for neuronal piRNAs in the epigenetic control of memory-related synaptic plasticity. *Cell* **149**, 693–707 (2012).
- Rosenberg, G. A new critical estimate of named species-level diversity of the recent mollusca. *Am. Malacol. Bull.* **32**, 308–322 (2014).
- Ma, X. et al. Piwi1 is essential for gametogenesis in mollusk *Chlamys farreri*. *PeerJ* **5**, e3412 (2017).
- Waldron, F. M., Stone, G. N. & Obbard, D. J. Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes. *PLoS Genet.* **14**, e1007533 (2018).
- Zhou, X., Liao, Z., Jia, Q., Cheng, L. & Li, F. Identification and characterization of Piwi subfamily in insects. *Biochem. Biophys. Res. Commun.* **362**, 126–131 (2007).
- Schurko, A. M., Logsdon, J. M. Jr. & Eads, B. D. Meiosis genes in *Daphnia pulex* and the role of parthenogenesis in genome evolution. *BMC Evol. Biol.* **9**, 78 (2009).
- Lewis, S. H., Salmela, H. & Obbard, D. J. Duplication and diversification of dipteran argonaute genes, and the evolutionary divergence of piwi and aubergine. *Genome Biol. Evol.* **8**, 507–518 (2016).
- Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M. & Vervoort, M. Evolution of RNA-binding proteins in animals: insights from genome-wide analysis in the sponge *Amphimedon queenslandica*. *Mol. Biol. Evol.* **28**, 2289–2303 (2011).
- Dowling, D. et al. Phylogenetic origin and diversification of RNAi pathway genes in insects. *Genome Biol. Evol.* **8**, 3784–3793 (2016).
- Sasaki, T., Shiohama, A., Minoshima, S. & Shimizu, N. Identification of eight members of the Argonaute family in the human genome. *Genomics* **82**, 323–330 (2003).
- Murchison, E. P. et al. Conservation of small RNA pathways in platypus. *Genome Res.* **18**, 995–1004 (2008).
- Tong, Y. et al. Transcriptomics analysis of crassostrea hongkongensis for the discovery of reproduction-related genes. *PLoS ONE* **10**, e0134280 (2015).
- Gebert, D., Hewel, C. & Rosenkranz, D. unittas: the universal tool for annotation of small RNAs. *BMC Genomics* **18**, 644 (2017).
- Schorn, A. J., Gutbrod, M. J., LeBlanc, C. & Martienssen, R. LTR-retrotransposon control by tRNA-derived small RNAs. *Cell* **170**, 61–71 (2017).
- Czech, B. & Hannon, G. J. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. *Trends Biochem. Sci.* **41**, 324–337 (2016).
- Rosenkranz, D. & Zischler, H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**, 5 (2012).
- Xu, F. et al. Identification of conserved and novel microRNAs in the Pacific oyster *Crassostrea gigas* by deep sequencing. *PLoS One* **9**, e104371 (2014).
- Llonga, N., Ylla, G., Bau, J., Belles, X., & Piulachs, M. Diversity of piRNA expression patterns during the ontogeny of the German cockroach. *J. Exp. Zool. B Mol. Dev. Evol.* <https://doi.org/10.1002/jez.b.22815> (2018).
- Babicki, S. et al. Heatmapper: web-enabled heat mapping for all. *Nucleic Acids Res.* **44**, W147–W153 (2016).

57. Fisher, W. S. In *Immunity in Invertebrates* (ed Brehélin, M.) 25–35 (Springer, Berlin, 1986).
58. Lau, Y. T., Sussman, L., Pales Espinosa, E., Katalay, S. & Allam, B. Characterization of hemocytes from different body fluids of the eastern oyster *Crassostrea virginica*. *Fish Shellfish Immunol.* **71**, 372–379 (2017).
59. Senti, K. A., Jurczak, D., Sachidanandam, R. & Brennecke, J. piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. *Genes Dev.* **29**, 1747–1762 (2015).
60. Gainetdinov, I., Skvortsova, Y., Kondratieva, S., Funikov, S. & Azhikina, T. Two modes of targeting transposable elements by piRNA pathway in human testis. *RNA* **23**, 1614–1625 (2017).
61. Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).
62. Aravin, A. A. et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol. Cell* **31**, 785–799 (2008).
63. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761–764 (2007).
64. Kawaoka, S., Izumi, N., Katsuma, S. & Tomari, Y. 3' end formation of PIWI-interacting RNAs in vitro. *Mol. Cell* **43**, 1015–1022 (2011).
65. Zhang, Z. et al. Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. *Mol. Cell* **44**, 572–584 (2011).
66. Vourekas, A. et al. Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nat. Struct. Mol. Biol.* **19**, 773–781 (2012).
67. Cannon, J. T. et al. Xenacoelomorpha is the sister group to Nephrozoa. *Nature* **530**, 89–93 (2016).
68. Rouse, G. W., Wilson, N. G., Carvajal, J. I. & Vrijenhoek, R. C. New deep-sea species of *Xenoturbella* and the position of Xenacoelomorpha. *Nature* **530**, 94–97 (2016).
69. Tosar, J. P., Rovira, C. & Cayota, A. Non-coding RNA fragments account for the majority of annotated piRNAs expressed in somatic non-gonadal tissues. *Commun. Biol.* **1**, 2 (2018).
70. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
71. Marchler-Bauer, A. et al. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–D226 (2015).
72. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).
73. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).
74. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
75. Guindon, S., Delsuc, F., Dufayard, J. F. & Gascuel, O. Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.* **537**, 113–137 (2009).
76. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, 351–358 (2005).
77. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
78. Davison, A. et al. Formin is associated with left-right asymmetry in the pond snail and the frog. *Curr. Biol.* **26**, 654–660 (2016).
79. Cantarel, B. L. et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
80. Morgulis, A., Gertz, E. M., Schäffer, A. A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
81. Rosenkranz, D., Han, C. T., Roovers, E. F., Zischler, H. & Ketting, R. F. Piwi proteins and piRNAs in mammalian oocytes and early embryos: from sample to sequence. *Genom. Data* **5**, 309–313 (2015).
82. Jiang, H. & Wong, W. H. SeqMap: mapping massive amount of oligonucleotides to the genome. *Bioinformatics* **24**, 2395–2396 (2008).
83. Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* **44**, W54–W57 (2016).
84. Lagesen, K. et al. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
85. Johnson, L. S., Eddy, S. R. & Portugaly, E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11**, 431 (2010).
86. Axtell, M. J. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* **19**, 740–751 (2013).
87. Zhou, Z. et al. The identification and characteristics of immune-related microRNAs in haemocytes of oyster *Crassostrea gigas*. *PLoS ONE* **9**, e88397 (2014).
88. Zhao, X., Yu, H., Kong, L., Liu, S. & Li, Q. High throughput sequencing of small RNAs transcriptomes in two *Crassostrea* oysters identifies microRNAs involved in osmotic stress response. *Sci. Rep.* **6**, 22687 (2016).

Acknowledgements

We thank Sacha Heerschoop, Julia Schumacher, Isabel Fast, and Hans Zischler for helpful comments and discussion. We also thank Mark Helm group for kindly providing chemicals. This work was supported by the International PhD Program (IPP) coordinated by the Institute of Molecular Biology IMB, Mainz, Germany, funded by the Boehringer Ingelheim Foundation.

Author contributions

J.J. and J.S.T.K. performed total RNA extraction and qPCR experiments. J.J. analyzed qPCR data and prepared the corresponding figures. D.G. identified PIWI paralogs in sequenced but unannotated molluscan genomes. D.G. performed PIWI gene tree reconstruction and prepared the corresponding figure. F.P. was responsible for farming and dissection of *L. stagnalis* animals. F.P. and S.S. performed RNA extraction for subsequent sRNA sequencing. C.H., F.P., and S.S. performed de novo miRNA annotation based on the obtained sRNA data. D.G. performed bioinformatics analysis of piRNA clusters and prepared the corresponding figures. D.R. analyzed sRNA data and developed Perl scripts for data analysis. D.G., J.J., and D.R. wrote the manuscript. J.S.T.K., F.P., and C.H. provided valuable input for corrections and improvements of the manuscript.

Additional information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s42003-018-0141-4>.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as 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 images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018