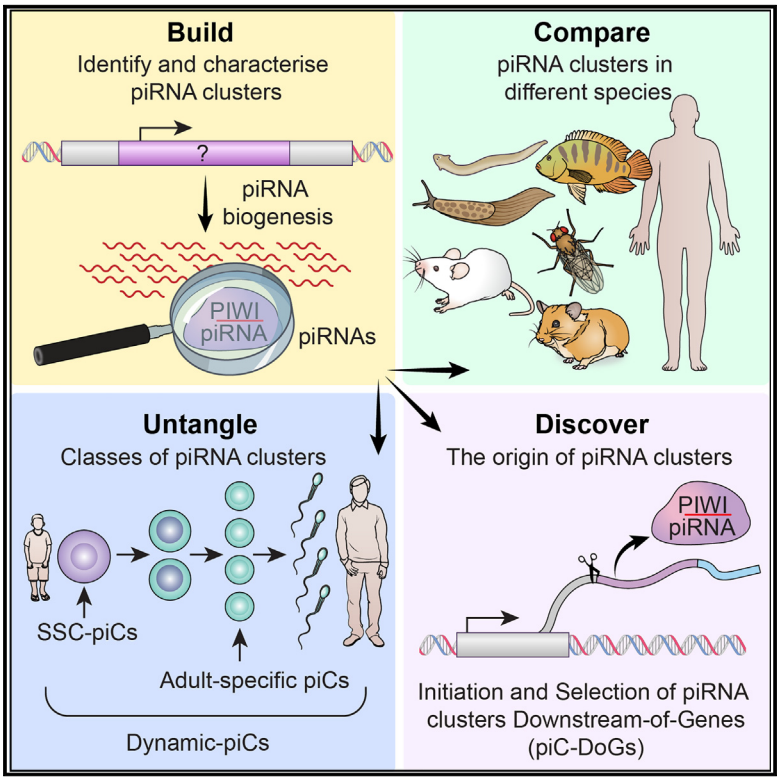
Cell Reports

Article

A comparative roadmap of PIWI-interacting RNAs across seven species reveals insights into *de novo* piRNA-precursor formation in mammals

Graphical abstract



Authors

Parthena Konstantinidou, Zuzana Loubalova, Franziska Ahrend, ..., Eric A. Miska, Josien C. van Wolfswinkel, Astrid D. Haase

Correspondence

astrid.haase@nih.gov

In brief

Konstantinidou et al. build and compare piRNA-generating genomic intervals (piRNA clusters) across seven species. Their comprehensive analysis uncovers a mechanism for the formation of piRNA clusters downstream of genes (piC-DoGs) in mammals and identifies three distinct classes of piRNA clusters in adult human testes.

Highlights

- The piRNA cluster builder (piCB) assembles, ranks, and characterizes piRNA clusters
- Mouse pre-pachytene piRNAs originate from readthrough downstream of genes (DoGs)
- A dynamic class of human piRNA clusters produces pre-pachytene and pachytene piRNAs
- A model for the birth of mammalian piRNA clusters in response to genotoxic stress

Konstantinidou et al., 2024, Cell Reports *43*, 114777
October 22, 2024 Published by Elsevier Inc.
https://doi.org/10.1016/j.celrep.2024.114777

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

A comparative roadmap of PIWI-interacting RNAs across seven species reveals insights into *de novo* piRNA-precursor formation in mammals

Parthena Konstantinidou,[1,14] Zuzana Loubalova,[1,14] Franziska Ahrend,[1,2,14] Aleksandr Friman,[1,3,12,13,14] Miguel Vasconcelos Almeida,[4,5] Axel Poulet,[6,7,8] Filip Horvat,[9,10] Yuejun Wang,[1,2,11] Wolfgang Losert,[12,13] Hernan Lorenzi,[1,11] Petr Svoboda,[9] Eric A. Miska,[4,5] Josien C. van Wolfswinkel,[6,7,8] and Astrid D. Haase[1,15,*]

[1]National Institutes of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA
[2]Oak Ridge Institute for Science and Education, US Department of Energy, Oak Ridge, TN, USA
[3]Biophysics Graduate Program, Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA
[4]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, UK
[5]Wellcome/CRUK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK
[6]Department of Molecular Cellular and Developmental Biology, Yale University, New Haven, CT 06511, USA
[7]Yale Stem Cell Center, Yale School of Medicine, New Haven, CT 06511, USA
[8]Center for RNA Science and Medicine, Yale School of Medicine, New Haven, CT 06511, USA
[9]Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[10]Bioinformatics Group, Division of Molecular Biology, Department of Biology, Faculty of Science, University of Zagreb, 10000 Zagreb, Croatia
[11]TriLab Bioinformatics Group, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA
[12]Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742, USA
[13]Department of Physics, University of Maryland, College Park, MD 20742, USA
[14]These authors contributed equally
[15]Lead contact
*Correspondence: astrid.haase@nih.gov
https://doi.org/10.1016/j.celrep.2024.114777

SUMMARY

PIWI-interacting RNAs (piRNAs) play a crucial role in safeguarding genome integrity by silencing mobile genetic elements. From flies to humans, piRNAs originate from long single-stranded precursors encoded by genomic piRNA clusters. How piRNA clusters form to adapt to genomic invaders and evolve to maintain protection remain key outstanding questions. Here, we generate a roadmap of piRNA clusters across seven species that highlights both similarities and variations. In mammals, we identify transcriptional readthrough as a mechanism to generate piRNAs from transposon insertions (piCs) downstream of genes (DoG). Together with the well-known stress-dependent DoG transcripts, our findings suggest a molecular mechanism for the formation of piRNA clusters in response to retroviral invasion. Finally, we identify a class of dynamic piRNA clusters in humans, underscoring unique features of human germ cell biology. Our results advance the understanding of conserved principles and species-specific variations in piRNA biology and provide tools for future studies.

INTRODUCTION

PIWI-interacting RNAs (piRNAs) and their PIWI protein partners are essential for germ cell health and the survival of species.[1–5] piRNA pathways establish restriction of mobile genetic elements to protect genome integrity.[6,7] Mutations in essential piRNA pathway genes result in sterility in animals and are associated with infertility in humans.[8,9] At the core of piRNA pathways are PIWI-piRNA complexes, comprising a PIWI protein and its guide RNA (piRNA).[10] Within the piRNA-induced silencing complex (piRISC), substrate specificity is dictated by the piRNA through complementary base pairing, while the PIWI protein partner determines the ultimate fate of the target, leading to either transcriptional or post-transcriptional restriction.[11]

piRNA pathways function analogously to adaptive immune systems as they rely on memory, produce mobile guards, and establish sequence-specific restriction of endogenous retroviruses and other transposons.[12] In mammals, zebrafish, and fruit flies, millions of piRNAs originate from thousands of large genomic regions. These "piRNA clusters" function as genetic repositories, retaining a memory of historical transposon mobility.[13–19] Studies in fruit flies and the endogenization of a novel retrovirus into the koala genome have demonstrated that piRNA pathways evolve to gain control over new mobile genetic

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

elements.[14,20–22] However, how new piRNA clusters emerge and adapt to novel transposons remains largely elusive.[23–26]

piRNA clusters are defined by computational assembly of nearby and overlapping piRNAs into larger genomic intervals.[13,14] They infer piRNA-precursor genes. Gene models for these pre-piRNA genes have only been determined for pachytene piRNA precursors, which produce piRNAs in meiotic spermatocytes in mammals.[1] Depletion of transposon sequences and their largely unique sequence space enabled the direct characterization of their long piRNA precursors.[27] The promoters, splicing patterns, and potentially alternative transcripts of other piRNA-precursor genes remain unknown.

Thousands of piRNA clusters have been identified in different organisms, but only a few have been functionally characterized. In *Drosophila*, the *flamenco* piRNA cluster has long been known as an essential transposon control region in ovaries.[28] The Y-linked piRNA cluster *Suppressor-of-Stellate* is essential for male fertility in flies.[29] In mice, a piRNA cluster on chromosome 18 (pi18) plays a central role in gene regulation during spermiogenesis,[30] and loss of pi6 produces sperm with reduced ability to sire viable embryos.[31] In contrast, the *Drosophila* piRNA clusters 42AB, 20A, and 38C, and the murine piRNA clusters *pi2*, *pi7*, *pi9*, and *pi17*, are dispensable for fertility.[31,32] Further studies are required to identify the piRNA clusters responsible for the sterility phenotypes of PIWI mutants in different organisms.[33]

piRNA clusters produce long single-stranded transcripts that are processed into multiple piRNAs.[1,10,33] During primary piRNA biogenesis, the ZUC(Pld6/MitoPld) endonuclease generates piRNAs that preferentially start with uridine (1U-bias).[34] A coordinated piRNA-guided slicing process, known as ping-pong, amplifies piRNA pairs with 10-nt complementarity across their 5′ ends.[14,35,36] These resulting sequence patterns are often used to characterize piRNAs, although the subtle enrichment, typically 2- to 5-fold over expected, is insufficient for piRNA selection.[14,34] Identifying *bona fide* piRNAs by their association with PIWI proteins remains the preferred method.[15]

The concept of piRNA clusters has been widely adopted and greatly facilitates functional studies. Improved piRNA sequencing data and genome assemblies have provided opportunities to reassess piRNA clusters to better understand piRNA precursors and explore additional species. However, methodological differences of piRNA cluster assembly, including varying minimal requirements for piRNA coverage and cluster length, and lack of available source code have complicated comparisons between studies and different organisms.[13–15,18,27,37]

To address these challenges, we developed a versatile toolkit—the piRNA cluster builder (piCB)—to identify, prioritize, and characterize piRNA clusters. We provide guidelines for optimizing assembly parameters based on the number of clusters, their combined genomic space, and the fraction of piRNAs they account for. We suggest attributes for ranking piRNA clusters by their production of piRNAs, drawing on previous studies that showed a positive correlation between piRNA abundance and function.[14,30,31,38,39] Using piCB, we characterized piRNA clusters across seven species, from slug to human. Our results revealed a gene model for mammalian pre-pachytene piRNAs and suggest a mechanism for the origin of transposon-silencing piRNA clusters.

RESULTS

The piCB reveals a flamenco super-cluster in *Drosophila*

We developed a straightforward method for constructing, ranking, and characterizing piRNA clusters. Our piCB refines the original approach of identifying genomic clusters with a high density of mapped piRNA reads by incrementally integrating unique and multimapping piRNAs (Figures 1A and S1A; Data S2 supplemental code and GitHub).[13–15] piCB's clusters are anchored in precise genomic positions through their "seeds" and are expanded into cores and final clusters through the unambiguous, stepwise integration of multimapping piRNAs.

We first applied piCB on the well-characterized Piwi-only piRNA pathway in *Drosophila* ovarian somatic sheath cells (OSCs).[38,40,41] Approximately two-thirds of these Piwi-piRNAs could originate from multiple positions in the genome (multi-mappers) and target endogenous retroviruses for epigenetic restriction. Optimization of piCB parameters ensured the incorporation of a maximum fraction of piRNAs in a minimal genomic space (Figure S1B). We identified more than 7,000 seeds, aggregating into over 6,000 clusters, collectively covering about 10% of the *Drosophila* genome and explaining more than 90% of all piRNAs in OSC (Figure 1B; Data S1).

Our approach, based on the incremental integration of both unique and multimapping piRNAs, resulted in the formation of three distinct types of piRNA clusters: solo-core, extended-core, and multi-core clusters (Figures 1C and 1D). Solo-core clusters assembled unique mapping piRNAs and primary alignments of multimapping reads. Extended-core clusters extended a single core in one or both directions upon integration of all possible multimapping alignments. If the extension reached into the next core and possibly across multiple cores, piCB built a single multi-core cluster. These large multi-core clusters often identified transposon-rich genomic intervals with poor mappability.

While thousands of Piwi-piRNA clusters were scattered throughout the genome, only a few were highly productive (Figures 1E and 1F).[14,38] Ranking piRNA clusters by their cumulative contribution to the piRNA population revealed a steep curve and correlated with reproducibility in biological replicates (Figure 1G). The top 130 clusters generated 90% of all cluster-derived piRNAs (90th percentile), with piC-1 alone accounting for more than half. Neither the length nor the nucleotide composition of piRNA clusters were indicative for piRNA productivity (Figures S1C and S1D).

The top-ranking piRNA cluster (piC-1) identified *flamenco* (*flam*), a well-known piRNA cluster and essential transposon control region.[14,20,21,42] piC-1 extended beyond the previously annotated flam and was closely followed by the second most productive cluster, piC-2 (Figure 1H).[37] Both piCs produced piRNAs targeting the active Ty3/mdg4 family of endogenous retroviruses (formerly known as *gypsy*) and were separated only by a short ~2,200-bp interval containing a single transposable element (Figure 1I). Mapping all piRNAs across this region revealed numerous alignments connecting piC-1 and -2. In contrast, a trailing cluster, piC-67, remained isolated (Figure S1E). Our data suggested that piC-1,2 form an

CellPress
OPEN ACCESS

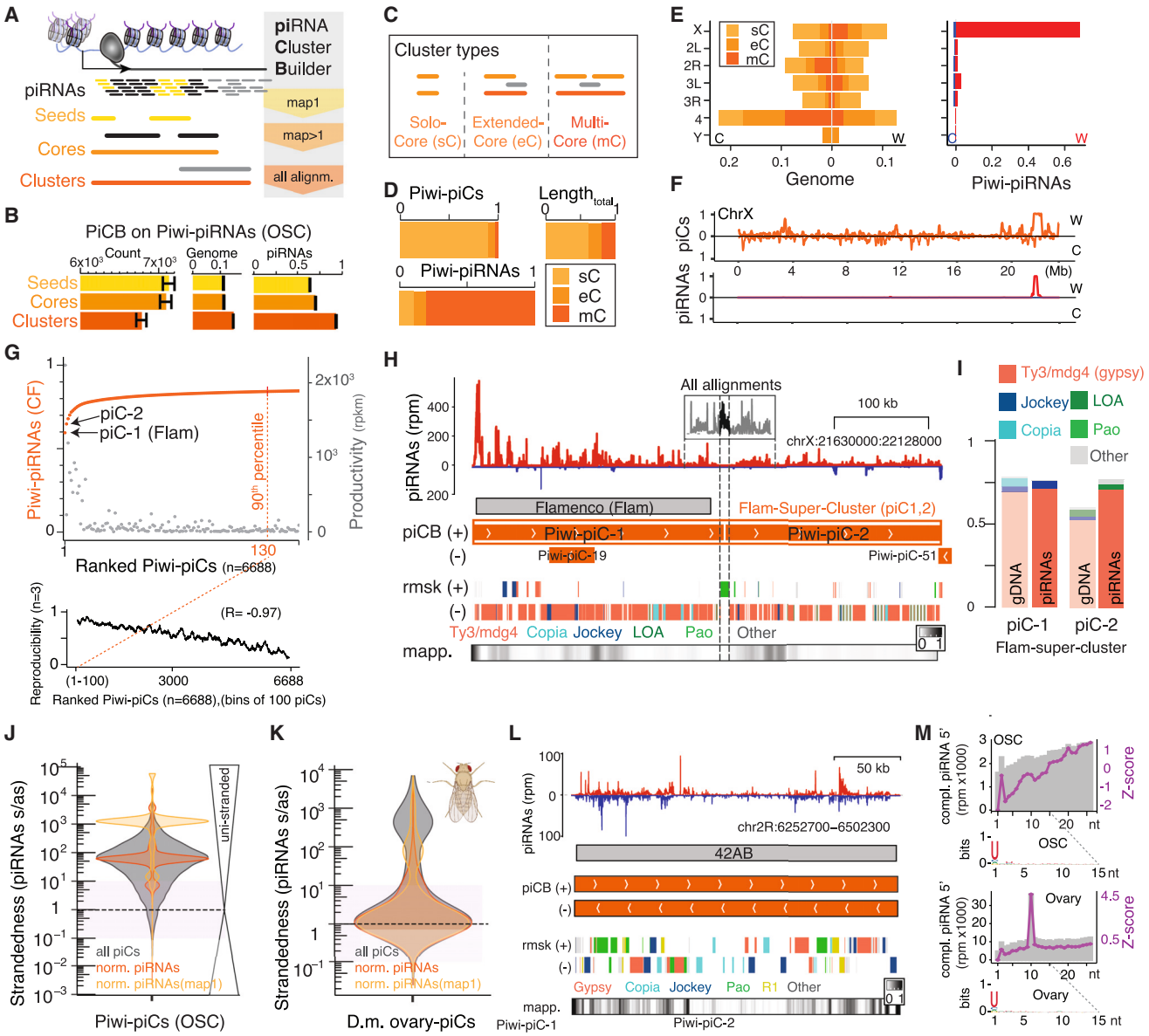


Figure 1. Characterization of *Drosophila* piRNA clusters using the piCB identifies a flam-super-cluster

(A) Stepwise integration of unique mapping piRNAs (map 1), primary alignments of multimapping piRNAs (map > 1), and all possible alignments build seeds, cores, and clusters.

(B) Characterization of seeds, cores, and clusters assembled from Piwi-piRNAs in ovarian somatic sheath cells (OSCs).

(C) Cluster types: single-core-clusters (sCs), extended-core clusters (eCs), and multi-core-clusters (mCs).

(D) Piwi-piRNA cluster types in OSCs.

(E) Piwi-piC types throughout the *Drosophila* genome. Watson (W) and Crick (C).

(F) Piwi-piCs and piRNAs across the X chromosome.

(G) piCs ranked by the cumulative fraction (CF) of Piwi-piRNAs (orange). Productivity (piRNAs/kb/millions, rpkm; gray). Reproducibility as probability of 98% nucleotide identity between piCs independently assembled from three biological piRNA replicates. Bins of 100. R = Pearson correlation coefficient.

(H) Genome track. Flamenco-super-cluster (Flam-sc). piRNA density, previously annotated flam (Han et al.[37]), piCB piCs, RepeatMasker (rmsk), and mappability (mapp., 20-mer; 1 = best, 0 = worst).

(I) Antisense annotation of genomic DNA (gDNA) and piRNAs.

(J) Strand preference (strandedness). Red/orange, weighed by cluster productivity (norm.).

(K) Strandedness of piCs from *Drosophila melanogaster* (*D.m.*) ovaries.

(L) Genome-track cluster 42AB.

(M) Features of piRNAs. Complementary piRNA 5′ start sites across positions 1–25. An about 2.5-fold enrichment of complementary piRNA start sites (5′) across position 10 indicates ping-pong amplification of piRNA pairs in *Drosophila* ovaries (rmp, reads per million). Sequences logos across piRNA nt 1–15. piCB is available in supplemental code and on GitHub. piRNA cluster coordinates and attributes are available in Data S1.

extensive flamenco super-cluster (flam-sc) spanning 495 kb (chrX:21,631,611-22,127,300).

piRNA clusters are classified as uni- or dual-stranded, depending on their capability to produce piRNAs from one or both genomic strands.[13–15] To systematically assess piRNAs from transposon insertions (piC) strandedness, we computed the ratio of sense and antisense piRNAs originating from a transcript on the same or the opposite genomic strand (Figures 1J and 1K). As anticipated, Piwi-piCs were predominantly unistranded, with a sense-to-antisense piRNA ratio greater than 10 in OSC (Figure 1J). In contrast, the majority of piRNA clusters in *Drosophila* ovaries, assembled from 24- to 35-nt-long small RNAs (Figures S1F–S1H), were dual-stranded (Figure 1K). The two strands of the well-known dual-stranded piC 42AB ranked first and third (Figures 1L and S1H; Data S1).[14,32,34,43] Overall, piCB identified ~75% of the previously annotated piRNA clusters and assembled additional clusters based on improved piRNA sequencing depth and genome assembly (Figure S1I).

To complete our *Drosophila* piRNA analyses, we determined sequence and ping-pong signatures indicative of their biogenesis pathways (Figure 1M).[10,33] As previously reported, only a minor fraction of Piwi-piRNAs (<0.2%) had complementary piRNA partners with no positional preferences across nt 1–25.[14,44] Piwi-piRNAs show a preference for uridine in the first position, indicative of primary piRNAs generated by the ZUC-processor complex.[34] In contrast, about 1% of piRNAs from *Drosophila* ovaries had a complementary piRNA partner, with a ~3-fold enrichment in position 10, indicating ping-pong-generated piRNA pairs in addition to 1U-biased primary piRNAs.[2,14,36]

In summary, our analyses of *Drosophila* piRNA clusters validates piCB as a useful tool for building, ranking, and characterizing these clusters. Optimization curves, the distribution of solo-, extended-, and multi-core clusters, and our straightforward measure for strandedness offer valuable features for assessing piRNA clusters and facilitate comparison of different piRNA pathways.

Prediction of piRNA-precursor transcripts

The pachytene piRNA pathway, specific to mammals, functions in meiotic and post-meiotic spermatocytes and is essential for spermatogenesis.[45] However, the piRNAs responsible for the sterility phenotype of PIWIL1/MIWI mutant mice and their targets remain largely elusive.[30,31,45–47] Pachytene piRNA precursors have been experimentally mapped through an elaborate combination of RNA sequencing (RNA-seq) and chromatin profiling experiments, aided by the low fraction of transposon-derived sequences and thus excellent mappability of these piRNAs.[27,48,49] The analyses revealed both spliced and un-spliced transcripts, often originating from bi-directional promoters.[27,49,50] Here, we discern patterns of spliced and bi-directional precursors that enable us to predict transcript structures based on computed mouse piRNA clusters. This approach aims to facilitate studies in situations where limited material, time, or budget constrain extensive experiments.

We assembled 1,737 and 2,023 predominantly unistranded piRNA clusters from PIWIL1/MIWI-PIWIL2/MILI-piRNA sequencing data of adult mouse testes, respectively (Figures 2A, 2B, S2A, and S2B; Data S1).[51] Ranking revealed a non-uniform distribution, pinpointing 64 top-ranking PIWIL1/MIWI- and 59 PIWIL2/MILI-piCs that accounted for 90% of all cluster-derived piRNAs (90th percentile). The cumulative genomic space of the top-200 pachytene piRNA clusters was approximately 3 Mb with a substantial 79% overlap between PIWIL1/MIWI- and PIWL2/MILI-piCs, and it encompassed all but the shortest experimentally determined piRNA-precursor gene (Figure 2C).[27,49] Similar to observations in flies, piRNA productivity did not correlate with length or base composition (Figures S2C–S2F). Our findings indicate that PIWIL1/MIWI- and PIWIL2/MILI-piRNAs predominantly originate from the same piRNA clusters in primary spermatocytes, which can be readily assembled using piCB.

To identify spliced piRNA precursors and predict their exon-intron structure, we hypothesized that a shorter distance of piRNA clusters (piC-to-piC distance) could indicate exons of the same piC gene (Figure 2D). piC-to-piC distances showed a bimodal distribution with a median distance of 500 kb (Figure 2E). Most piRNA clusters were more than 100 kb apart. However, we also observed a group with piC-to-piC distances shorter than 30 kb. These distances were comparable to the intron lengths of piRNA precursors and protein-coding genes in mouse, which range from a few base pairs to about 100 kb, with a median length of approximately 1.5 kb (Figure 2F). Thus, piRNA clusters with a piC-to-piC length of less than 30 kb might represent exons that belong to the same pre-piRNA gene. Accordingly, we combined 89 PIWIL1/MIWI- and 71 PIWIL2/MILI-piCs into 35 and 27 piC groups that identified 24 and 21 potentially spliced precursors, respectively. A representative genome track illustrates piC groups that identified the experimentally determined piRNA precursor 2-qF1-2536 (Figure 2G).[49] *De novo* assembly of piRNAs into transcripts for individual piC groups further improved our prediction. Our findings indicated that the transcript structure of spliced piRNA precursors can be estimated by considering the distances between piCB-assembled piRNA clusters.

Pachytene piRNA precursors are often transcribed from bi-directional promoters and result in a specific pattern of paired uni-stranded clusters on opposite genomic strands.[13,15,27,49] To identify these bi-directional piC-pairs, we selected piCs on opposite genomic strands that overlapped at their 5′ ends and were uni-stranded throughout their 3′ body (Figure 2H). Using this strategy, we identified 17 piC pairs that included all the reported bi-directional precursors.[27,49] A representative example shows pairs of PIWIL1/MIWI- and PIWIL2/MILI-piCs that correspond to the bi-directional precursors 5-qF-14224 and 5-qF-14508 (Figure 2I).[49]

Among the top 200 PIWIL2/MILI-piCs, 48 exhibited patterns indicative of spliced pre-piRNA genes, 30 formed 16 bi-directional piC-pairs, and more than half (61%) were uni-stranded without signatures of splicing or bidirectionality (Figure 2J). In summary, our simple analyses reliably identified spliced pre-piRNA genes and bi-directional piC-pairs. These methods could prove valuable for inferring piRNA precursors when limitations in material, funds, or time hinder extensive experimental characterization.

piCB-assembled uni- and dual-stranded piRNA clusters in planaria, cichlid fish, and slug

The fundamental role of piRNA pathways in safeguarding genomes from mobile genetic elements places them at the very core of evolution.[12] To understand how piRNA pathways

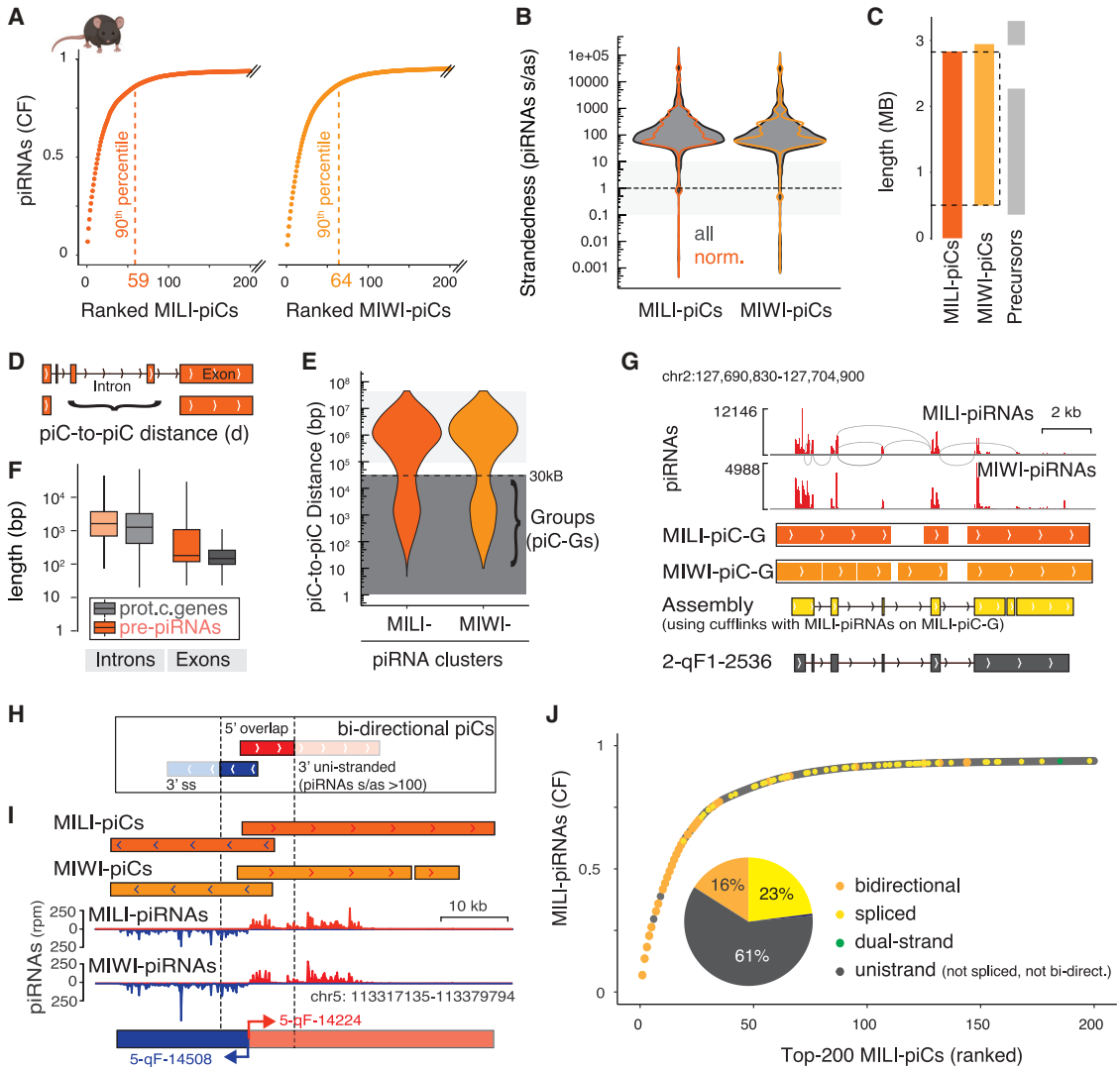


Figure 2. Assembly of mouse pachytene piRNA clusters and prediction of spliced transcripts as cluster-groups (piC group)

(A) piCs ranked by the CF of MILI- and MIWI-piRNAs.

(B) Strandedness of MILI- and MIWI-piCs. Productivity normalized (norm.).

(C) Intersection of MILI- and MIWI-piCs with pachytene piRNA precursors.

(D) Prediction of spliced precursors considering piC-to-piC distance (d).

(E) Distance between MILI- and MIWI-piCs.

(F) Length of introns and exons from protein-coding genes (RefSeq) and spliced precursors in base pairs (bp).

(G) Genome track of a representative piC-group (piC-G) at piRNA precursor (2-qF1-2536). Transcript assembly (Cufflinks) from MILI-piRNAs (yellow).

(H) Systematic identification of bi-directional piCs: 5′ overlap and uni-stranded (piRNA s/as >100) 3′ bodies (3′ 70%).

(I) Representative bi-directional piCs.

(J) Types of MILI-piCs.

confront the challenges of retroviral invasion, cope with genotoxic stress, and ensure the protection of generations, comparative analyses across different species are essential. Here, we explored piRNA clusters in planaria, cichlid fish, and slug (Figure 3). To enrich for piRNAs from total small-RNA sequencing data from planaria, cichlid-fish testes, and slug ovotestes, we removed microRNAs and fragments of abundant cellular RNAs by sequence and then selected for piRNA-sized small RNAs (see STAR Methods) (Figure S3A). The optimization strategies of piCB enabled us to overcome limitations of draft genomes, the simple ratio of sense-to-antisense piRNAs provided an unbiased visualization of strandedness across clusters, and the ranking process identified a shortlist for subsequent in-depth analyses.

Planaria are renowned for their regenerative abilities, relying on piRNA pathways in germ cells and pluripotent stem cells.[52–54] To optimize piCB parameters, we experimented with different settings and selected those capturing 60% of all piRNAs while

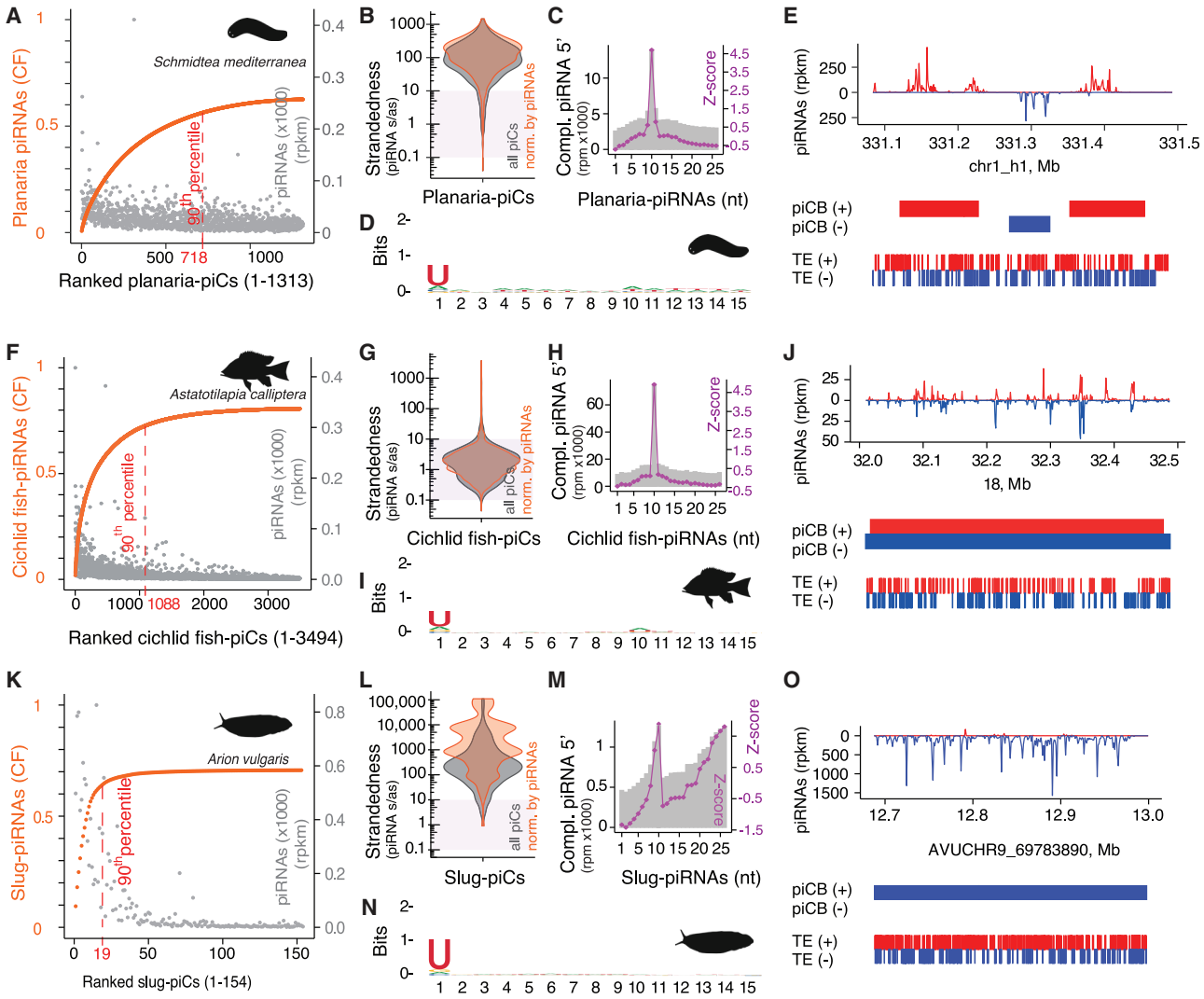


Figure 3. piCB-assembled piRNA clusters in planaria, cichlid fish, and slug ovotestis

(A, F, and K) Ranked piCB-piRNA clusters (piCs) according to the CF of piRNAs.

(B, G, and L) Strandedness of piCs. Productivity normalized (norm.).

(C, H, and M) Complementary (compl.) piRNA start sites (5′) across piRNA nt 1–25. On average, we observed 4,078, 14,080, and 800 complementary piRNA 5′ start sites per million reads (rpm) across nucleotide 1–25 in planaria, cichlid fish, and slug, respectively.

(D, I, and N) Sequence preferences of piRNAs across nucleotides 1–15.

(E, J, and O). Representative genome tracks. Transposable elements (TEs). Plus (red) and minus (blue) strand. (A–E) Planaria-piCs assembled from 28- to 40-nt small RNAs (*Schmidtea mediterranea*); (F–J) Cichlid-fish piCs assembled from testes small RNAs (24–35 nt) (*Astatotilapia calliptera*); (K–O) Spanish slug piCs assembled from ovotestes small RNAs (25–35 nt) (*Arion vulgaris*).

restricting the combined genomic space of the resulting piCs to less than 2% (Figures S3D–S3F). Ranking the 1,313 piCs by their cumulative contribution to the piRNA population identified 718 top-ranking piCs that are responsible for 90% of cluster-derived piRNAs (Figure 3A; Data S1). Planarian piRNA clusters were diverse in length, with no observable correlation between length or nucleotide composition and piRNA productivity (Figures S3G–S3I). Most piRNA clusters appeared uni-stranded, with a sense-to-antisense piRNA ratio exceeding 100 (Figure 3B). About 0.4% of all piRNAs overlapped with a complementary piRNA with an about 3-fold enrichment for complementary piRNAs in position 10, which is indicative for ping-pong pairs (Figure 3C). Planaria piRNAs showed a preference for uridine in the 5′-most position (Figure 1D). Similar to flies and mice, planarian piCs were distributed throughout the genomes, yet their uneven productivity created genomic hotspots for piRNA production (Figures S3J and S3K). Figure 3E provides a representative example of piRNA density across piRNA clusters.

East African cichlid fishes are known for their extreme phenotypic diversity despite high genomic sequence similarity and provide a unique model for studying the epigenetic contributions to species diversification.[55–59] In this context, we characterized

CellPress
OPEN ACCESS

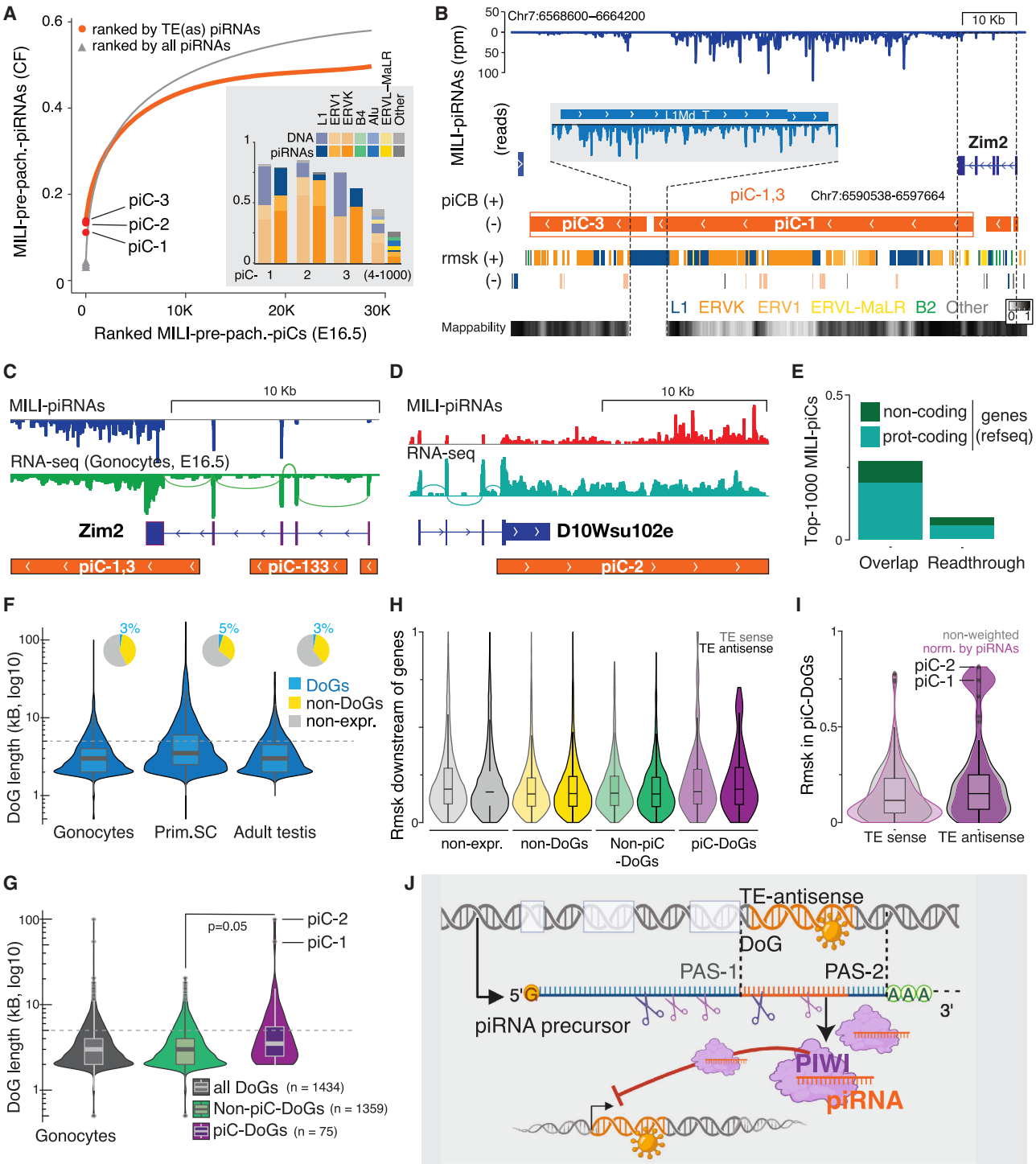


Figure 4. piC-DoGs suggest a model for the formation of pre-pachytene piRNA clusters in response to retroviral invasion in mice

(A) piCs ranked by the CF of TEas (orange) or all (gray) MILI-pre-pachytene piRNAs. Insert: fraction of TEas sequences at genomic (DNA, faded colors) and piRNA (strong colors) level.

(B) piC1,3 super-cluster associates with an upstream gene (*Zim2*). piC-1 and -3 are separated by a L1MdT element covered by multimapping (>1,000 alignments) piRNAs. RepeatMasker (rmsk), mappability (mapp.; 20mers, 1 = best), and RefSeq genes.

(C) piCs, piRNAs, and transcripts associated with *Zim2* in gonocytes.

(D) piC-2 overlaps with *D10Wsu102e*. piRNAs originate from all exons and from the 3′ extended region.

(legend continued on next page)

piRNA clusters in testes of the eastern happy (*Astatotilapia calliptera*). Utilizing piCB, we identified 3,494 piRNA clusters explaining more than 80% of all piRNAs (Figure 3F). Most of these clusters produced piRNAs from both genomic strands, appearing dual-stranded with a sense-to-antisense piRNA ratio close to one (Figure 3G). These piRNAs showed a ping-pong signature and a preference to harbor a uridine in the first position (Figures 3H and 3I). A representative example of a cichlid-fish dual-stranded piRNA cluster and its piRNAs is illustrated in Figure 3J. A follow-up study has revealed dynamic co-evolution of piRNA pathways and mobile genetic elements furthering our understanding of genomic diversity and the need to maintain genome integrity.[60]

Next, we compiled piRNA clusters from ovotestis of the Spanish slug (*Arion vulgaris*), a major European pest known for its dense mucus.[61,62] Like piRNAs in flies, mice, planaria, and fish, slug piRNAs formed extensive genomic clusters with varying piRNA productivity (Figure 3K). Ranking the slug piCs revealed a steep cumulative curve with highly productive clusters that contributed disproportionately to the observed piRNA population. Slug piCs were predominantly uni-stranded, with a sense-to-antisense piRNA ratio exceeding 100 (Figure 3L). Less than 1% of slug piRNAs can be matched with complementary piRNA partners and we do not observe a clear preference for piRNA pairs with 10-nt overlap (Figure 3M). Instead, slug piRNAs show a preference for uridine in the first position, which is indicative of primary piRNAs (Figure 3N). Illustrating a representative top-ranked piRNA cluster, we observed piRNA production from the minus strand capturing information from densely packed transposon fragments on the plus strand (Figure 3O). The resulting transposon-antisense (TEas) piRNAs could be vital to the fertility of slugs and might provide information on how to control this invasive species.

In conclusion, our investigation into piRNA clusters across diverse organisms revealed both commonalities and distinctions, underscoring the diversity of piRNA-generating genomic regions and the versatility of piCB in exploring various piRNA pathways in an impartial and adaptable manner.

piCB identified pre-pachytene piC downstream of genes, suggesting a model for the formation of transposon-silencing piRNA clusters in mammals

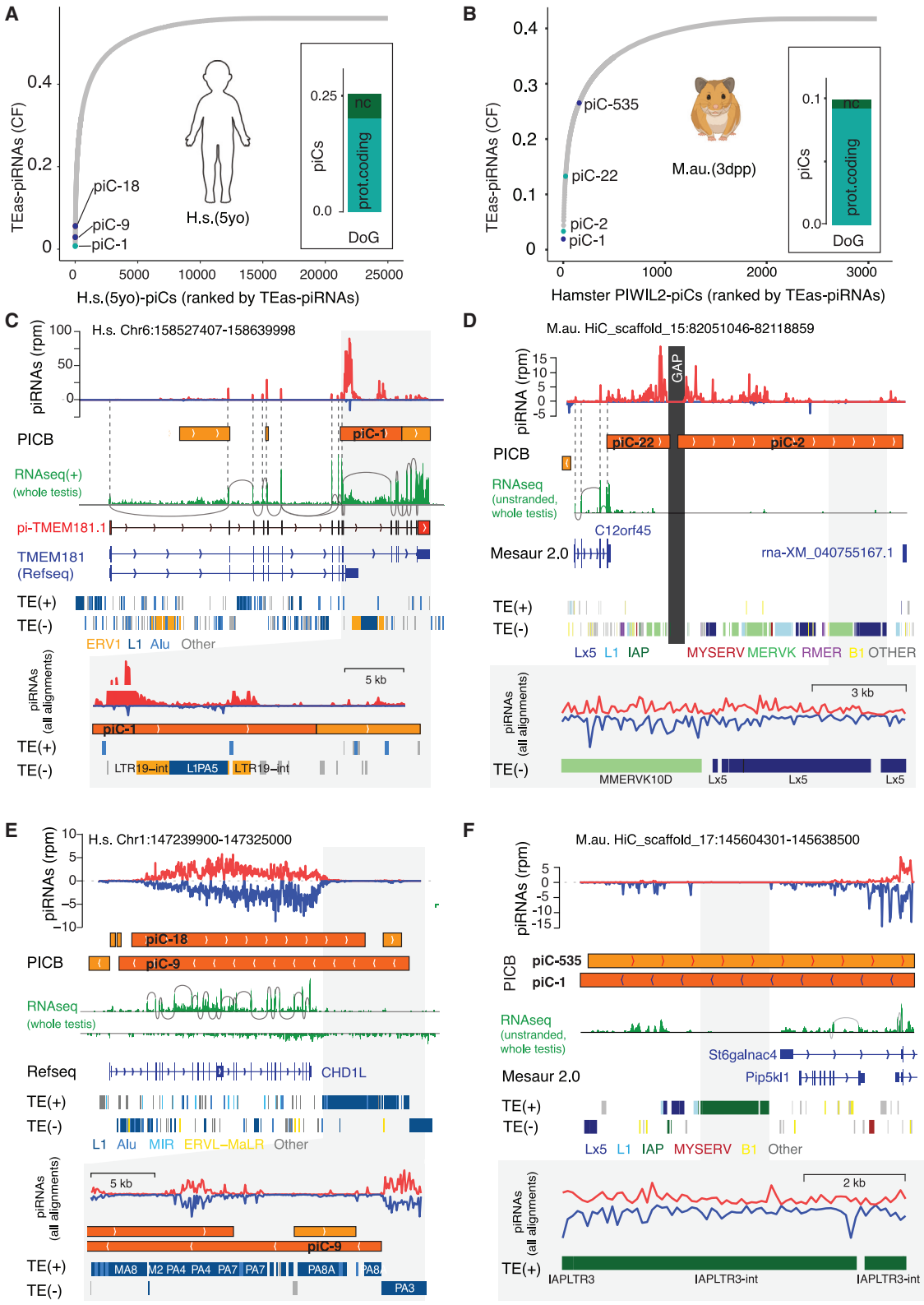
Pre-pachytene piRNAs establish lasting epigenetic restriction of transposons in gonocytes (prospermatogonia), which are derived from primordial germ cells and give rise to spermatogonial stem cells (SSCs) in mammals.[63–67] While the importance of these piRNAs for genome integrity and fertility is well established, little is known about their precursors and how they adapt to control novel genomic invaders throughout evolution. Using available PIWIL2(MILI-) and PIWIL4(MIWI2-)piRNA data from newborn mouse testes,[68] we assembled 28,626 and 11,461 piRNA clusters, respectively (Figures S4A–S4D). Based on their function in transposon control, we prioritized piRNA clusters by their cumulative production of TEas piRNAs and identified three top-ranking clusters that collectively accounted for more than 15% of all transposon-targeting piRNAs (Figure 4A). piC-1–3 were enriched in sequences of the same three retrotransposon families: the non-long terminal repeat (LTR) elements Line-1 (L1), the endogenous retroviruses 1 (ERV1), and ERVK, and over 60% of their piRNAs were directed against these elements (Figure 4A insert).

The most productive piC (piC-1) overlapped with the last exon of a known gene, zinc-finger imprinted 2 (*Zim2*) and was trailed by piC-3 (Figure 4B). Similar to flam-sc in *Drosophila*, piC-1 and -3 were connected by a young LINE-1 (L1MdT) insertion, which became visible when plotting piRNAs with more than 100 mappable positions in the genome (Figure 4B insert). Examining *Zim2* gene expression in embryonic mouse testes (E16.5) revealed transcriptional readthrough, resulting in transcripts that extended beyond the annotated polyadenylation site (PAS) and across piC1–3 (Figure 4C).[69] These readthrough transcripts potentially exceeded 100 kb in length and produced piRNAs throughout the transcript. Our data corroborate previous suggestions that the Zim2 promoter generates a piRNA precursor and reveal a readthrough transcript.[39]

The second top-ranking cluster (piC-2) intersected with a known protein-coding gene (*D10Wsu102e*). Similar to piC-1, we observed a readthrough transcript across piC-2 and detected piRNAs from all exons. This suggested that the piC-2 piRNA precursor is transcribed from the host gene's promoter as a spliced transcript that extends beyond the annotated PAS (Figure 4D). Motivated by these instances of transcriptional readthrough leading to pre-pachytene piRNA precursors, we conducted a systematic analysis to explore the connection between piRNA clusters and transcriptional readthrough. Our results unveiled that 269 out of the top 1,000 piRNA clusters were positioned downstream of known genes and overlapped with their last exon (Figure 4E). A focused examination of transcriptional readthrough using RNA-seq data from sorted embryonic gonocytes (E16.5) identified *bona fide* readthrough transcripts at 76 piC-associated genes, highlighting the significance of transcriptional readthrough in generating pre-pachytene piRNA precursors (Figure 4E).[69,70]

Next, we sought to understand the reason behind the production of these readthrough transcripts in gonocytes. Previous research in somatic cell lines has associated transcriptional

(E) 27% of the top-1000 MILI-piCs overlap with an upstream gene (RefSeq, protein(prot)-coding and non-coding).
(F) Length of DoG transcripts in gonocytes, primary spermatocytes, and adult testis. Pie charts: DoGs (blue) as percentage of all annotated genes (RefSeq). Yellow, expressed in gonocytes; gray, non-expressed.
(G) DoGs in gonocytes: piC-DoGs are generally longer than non-piC-DoGs.
(H) Genomic transposon (TE) content does not correlate with DoGs.
(I) The most productive piC-DoGs are enriched in TE antisense.
(J) A gene model for transposon-silencing piRNA precursors in murine gonocytes: promoters of protein-coding and non-coding genes generate long piRNA-precursor transcripts that extend beyond their original polyadenylation site (PAS-1). piC-DoGs capture a transposon-rich region downstream of the gene before terminating at a new polyadenylation site (PAS-2). The resulting transcript generates piRNAs with antisense complementarity to transposons and the ability to silence the corresponding elements *in trans*.



(legend on next page)

readthrough with various stress conditions, including viral infection, and identified transcripts that contained sequences downstream of genes (DoG).[71–78] However, the physiological relevance of these stress DoGs remains elusive. Our discovery that pre-pachytene piRNAs originate from DoG transcripts raised the possibility that gonocytes might suffer from acute stress. Alternatively, piC-DoGs could be relicts of historical stress events, potentially linked to the endogenization of a novel retrovirus or transposon activity.

To test these hypotheses, we measured the overall extent of readthrough transcription in gonocytes.[69,70] For comparison, we re-analyzed confirmed stress DoGs in murine fibroblasts (Figure S4E) and examined readthrough transcripts in whole testes and primary spermatocytes, where piRNAs do not originate from piC-DoGs (Figure 4F).[27,51,70,79] Somatic stress DoGs were previously reported to be lengthy, with a mean readthrough extension of 5 kb or more, and some exceeding 50 kb (Figure S4E).[73] In contrast, DoG transcripts in gonocytes, primary spermatocytes, and whole testes were similar to un-stressed somatic cells, indicating the absence of acute stress (Figure 4F). However, the top-ranking piRNA clusters stood out by the length of their DoG transcripts (Figure 4G). DoG transcripts associated with piC-1 and -2 exceeded 50 kb and were comparable to the longest stress DoGs in somatic cells (Figures 4G and S4F). Our results suggested that, although prominent piC-DoGs share length similarities with stress DoGs, they exist in the absence of acute stress in gonocytes.

Top-ranking piC-DoGs contained densely packed transposon fragments downstream of the gene. Thus, we tested whether transposon insertions DoG more generally correlate with readthrough transcription. We calculated the fraction of transposon sequences across 10 kb downstream of all annotated PASs (Figure 4H). Because our top-ranking piC-DoGs showed a remarkable bias for TEas sequences, we stratified sense and antisense transposon content with respect to the transcriptional orientation of the upstream gene. We observed a similar median (~15%) content of transposon sense and antisense sequences within 10 kb DoG that were not expressed (non-expr.), those that were expressed but did not show signs of transcriptional readthrough (non-DoGs), genes that produced DoG transcripts but no piRNAs (non-piC-DoGs), and piRNA clusters with readthrough transcription (piC-DoGs). Weighting piC-DoGs by their piRNA production revealed an enrichment of TEas sequences in highly productive piC-DoGs. Our data showed that there is no general association between transposon content downstream of a gene and transcriptional readthrough in murine gonocytes.

Our results reveal a gene model for pre-pachytene piRNA precursors (Figure 4J). These piC-DoGs harness promoters of protein-coding and non-coding genes and generate 3′ extended transcripts while maintaining splicing patterns and preserving coding sequences of their host genes. The *Zim2* promoter has previously been shown to be essential for piRNA production from a downstream piRNA cluster.[39] Because *Zim2* is not protein coding, this individual instance could have been interpreted as mis-annotation at the time. However, our data show that *Zim2* is not an isolated example but reveal a general mechanism for pre-pachytene piRNA production. The striking amount of TEas content in piC-DoGs could be the results of purifying selection to optimize these alternative transcripts to produce transposon-silencing piRNAs.

piC-DoGs produce transposon-silencing piRNAs in human and hamster

Next, we investigated whether the phenomenon of piC-DoGs is conserved in other mammals. We identified 25,004 human and 3,083 hamster pre-pachytene piRNA clusters based on publicly available piRNA sequencing data from a 5-year-old boy and a 3-day-old hamster (Figures 5A and 5B).[50,80] Similar to our observations in mouse, approximately 25% of the human and 10% of the hamster piCs were located DoG and overlapped with the genes' last exons (Figures 5A and 5B inserts).

Ranking human pre-pachytene piCs by their cumulative production of transposon-silencing piRNAs (TEas-piRNAs) revealed a lead cluster (piC-1) located downstream of *TMEM181*, encoding a transmembrane protein. piC-1(TMEM181) pre-piRNA transcripts are alternative variants of the piRNA precursor pi-TMEM181.1 and include intronic space that harbors major transposon insertions (LINE-1 and ERV1) on the opposite genomic strand.[50] The piC-1(TMEM181) transcripts capture transposon fragments downstream of *TMEM181* in antisense orientation (Figure 5C) and generate ample piRNAs against LINE-1 and ERV1 (Figure 5C insert). Another piC-DoG, piC-4, extends downstream of the KANTR gene, which encodes an integral membrane protein. The piC-4(*KANTR*)-DoG transcript has recently been mapped as a pre-pachytene piRNA precursor (pi-KANTR.1) (Figure S5A).[50] It captures antisense information of ERVL and ERVK insertions downstream of the annotated KANTR mRNA while maintaining the upstream transcript structure. In summary, our data suggest that, similar to mice, human pre-pachytene piRNAs originate from piC-DoGs and their precursors are readthrough transcripts of piC-associated genes.

In golden hamster, another well-established model for mammalian piRNA biology, we identified two lead piRNA clusters, piC-2 and -22, separated by a gap in the genome assembly chaperon (Figure 5D).[81–83] piC-(2,22) extended more than 10 kb downstream of the *C12orf45* gene and captured a transposon-dense genomic interval representing MMERVK, Lx5, and other young transposon families, producing piRNAs with mostly TEas information.[82] Integrating gene expression and piRNA sequencing data from testes of a 3-day-old hamster, we observed piRNA production from the exons but not the introns of *C12orf45*, and from the region downstream of the annotated PAS, consistent with piRNA production from a readthrough transcript.[80] Another potential piC-DoG associated with the *Cbl* proto-oncogene, which encodes an E3 ubiquitin ligase on the

Figure 5. Pre-pachytene piC-DoGs and dual-stranded piRNA clusters in human and hamster

(A and B) piCs ranked by the CF of TEas piRNAs in human (H.s.) (A) and golden hamster (M.au.) (B). Insert: piC-DoGs, 25% of human and 10% of hamster pre-pachytene piCs overlap with upstream genes.

(C–F) Representative examples of top-ranking uni-stranded (C and D) and dual-stranded (E and F) piCs.

CellPress
OPEN ACCESS

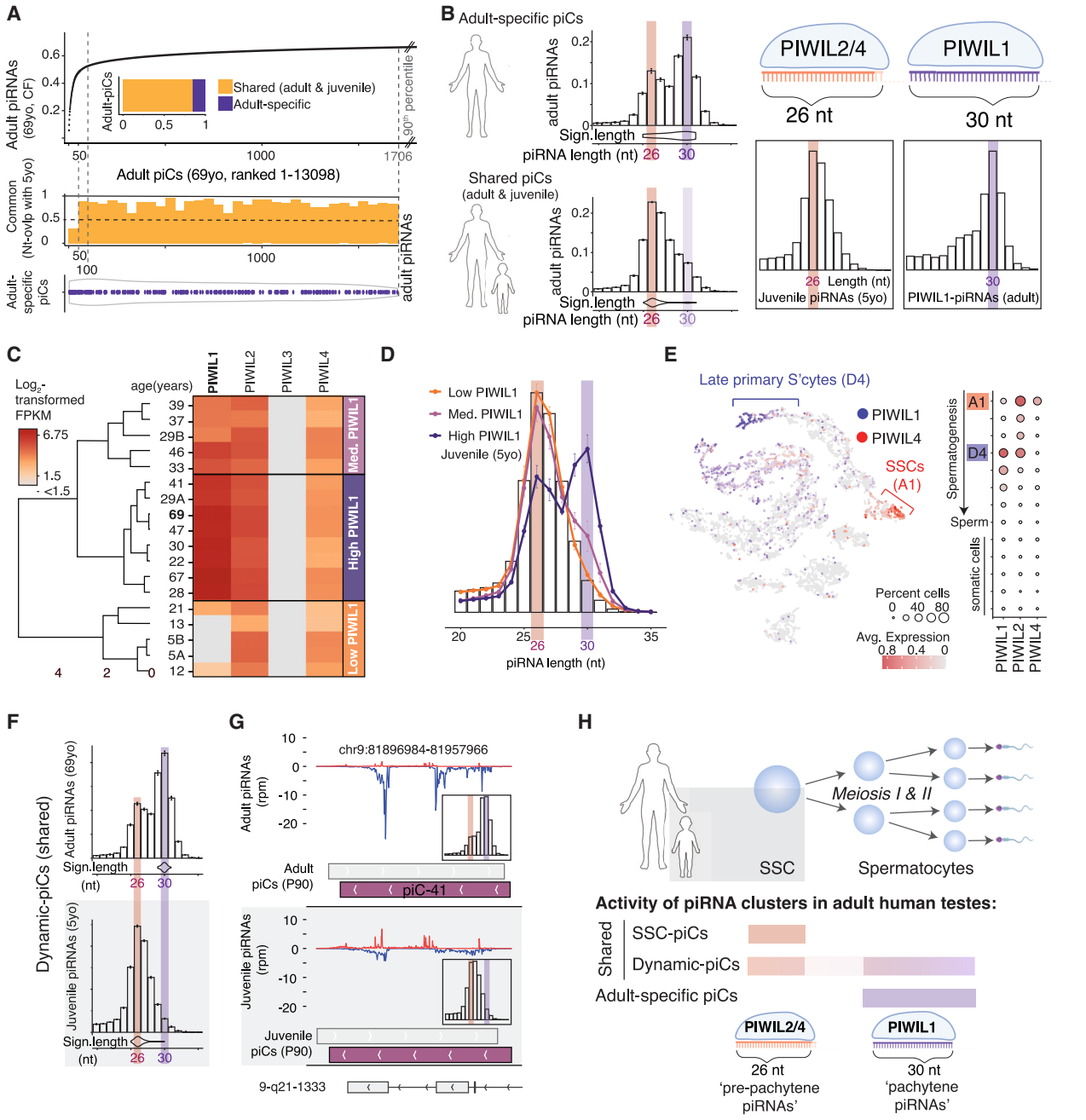


Figure 6. Disentangling piRNA populations in human testes reveals the persistence of pre-pachytene piRNAs in adult SSCs and a class of developmentally dynamic piCs

(A) Productivity does not distinguish adult-specific from shared (common between adult and juvenile) piRNA clusters in human testes. Ranked piCs from adult testes (69 years old). Nucleotide overlap (Nt-ovlp) with juvenile piCs (5 years old). Insert: piRNAs in adult testes originate from adult-specific and shared (adult and juvenile) piCs.

(B) Adult piRNA length distributions per piC for adult-specific piCs (top) and shared piCs (bottom). Signature piRNA length per piC (Sign. length). The length profile of juvenile piRNAs agrees with PIWIL2/4-associated piRNAs (peak-length 26 nt). In contrast, PIWIL1-associated piRNAs peak at 30 nt.

(C) Expression of PIWIL1-4 in testes RNA-seq from individuals of varying age identifies three groups based on PIWIL1 expression (indicated).

(D) piRNA length distribution in PIWIL1-high, -medium, and -low groups, and juvenile testes.

(E) Differential expression of PIWIL1 and PIWIL4 according to the Young Adult Human Testis Atlas. Spermatogonial stem cells (SSCs, A1), late primary spermatocytes (late primary S'cytes, D4).

(legend continued on next page)

minus strand of HiC-scaffold-8 (Figure S5B). Taken together, our findings in mouse, human, and hamster support a conserved role for piC-DoGs and their readthrough transcripts in generating transposon-silencing pre-pachytene piRNAs.

In addition to piC-DoGs, we identified dual-stranded piRNA clusters among the top-ranking transposon-silencing piCs in human and hamster with no signs of potential readthrough from upstream genes. For instance, human piC-9 captures antisense information from densely packed LINE-1 insertions on the plus strand and other L1 insertions on the minus strand (Figure 5E). Despite overlapping with a protein-coding gene, chromodomain helicase DNA binding protein 1-like (*CHD1L*), we observed piRNAs from introns and exons of *CHD1L*, suggesting that a piC-9 uses an independent transcriptional unit. In hamster, piC-1 is part of a dual-stranded pair with the lower-ranking piC-535 and spans a large internal LTR element of the young potentially active IAP/LTR3 subfamily (Figure 5F).[82] Two additional examples of dual-stranded piRNA clusters in human and hamster are illustrated in Figures S5C and S5D. Overall, our systematic analysis of pre-pachytene piRNA clusters in human and hamster identified patterns of piC-DoGs and dual-stranded clusters, emphasizing conserved patterns and diversity among mammalian piRNA precursors.

A new class of developmentally dynamic piRNA clusters produces pre-pachytene as well as pachytene piRNAs in human testes

Finally, we used piCB to characterize piRNA clusters in adult human testes. Our analyses, integrating publicly available single-cell, bulk, and piRNA sequencing data, unveiled distinct piRNA pathways in adult SSCs and meiotic spermatocytes and discovered a new category of dynamic piRNA clusters that produce stem cell (pre-pachytene) piRNAs in juvenile testes and meiotic (pachytene) piRNAs in the adult.[50,84]

Among the top-ranking adult piRNA clusters in the testes of a 69-year-old, we observed a substantial overlap with juvenile piRNA clusters (Figures 6A and S6A).[50] Of the top 1,706 adult piCs, accounting for 90% of all cluster-derived piRNAs, 85% coincided with juvenile piCs from a 5-year-old (Figure 6A insert). Sorting piRNA clusters based on their cumulative contribution to adult piRNAs did not separate these shared from the adult-specific piCs, indicating that piRNA production is not a distinguishing feature. However, we noticed a significant difference in the length profiles of piRNAs generated from shared or adult-specific clusters. Shared piCs predominantly produced 26-nt piRNAs, reminiscent of juvenile piRNAs associated with PIWIL2 and PIWIL4 (PIWIL2/4). Conversely, adult-specific piCs produced mainly 30-nt piRNAs, aligning with the profile of PIWIL1-piRNAs (Figure 6B). The length of piRNAs is dictated by the footprint of their associated PIWI proteins.[8,13–15,63,85,86] Our findings, revealing distinct length profiles of piRNAs produced by shared or adult-specific piRNA clusters, suggest their association with different PIWI proteins.

The human genome encodes four PIWI proteins (PIWIL1–4), three of which are expressed during spermatogenesis.[81,84,87,88] PIWIL1, expressed post puberty, distinguished juvenile from adult testes across 17 males aged 5–69 years (Figure 6C). As anticipated, PIWIL2 was expressed in all samples, and the oocyte-specific PIWIL3 was absent. Intriguingly, PIWIL4, typical in gonocytes of embryonic and juvenile mouse testes, displayed robust expression in adult human testes, aligning with the persistence of undifferentiated SSCs beyond puberty in humans (S6B).[89–93] According to the differential expression of PIWIL1 and -4, we observed varying piRNA length profiles in testes of low-, medium-, and high-PIWIL1 individuals (Figure 6C).[50] Integration of single-cell gene expression data from the Human Testis Atlas revealed the distinct expression of PIWIL4 and PIWIL1 in SSCs and primary spermatocytes, respectively (Figure 6E).[92] Co-expression of PIWIL2 and other essential piRNA biogenesis factors in both cell types supported the presence of two distinct piRNA pathways in adult human testes (Figure S6C). Taken together, our data suggested that the unique length profiles of shared and adult-specific piRNAs indicate their function in distinct piRNA pathways and at specific stages of spermatogenesis.

To estimate the activity of individual piRNA clusters during different stages of spermatogenesis, we calculated the preferred piRNA length for each cluster. As expected, adult-specific piCs showed a signature length of 30 nt, consistent with PIWIL1-piRNAs in spermatocytes, and identified pachytene piRNA precursors (Figure S6D). Surprisingly, shared piCs separated into two groups based on their signature piRNA length. Most shared piRNA clusters (90%) produced piRNAs with a preferred length of 26 nt, indicating PIWIL2/4-associated piRNAs in adult SSCs (Figure S6E). The signature piRNA length for these SSC-piCs was the same for adult and juvenile testes and identified the top-ranking pre-pachytene piRNA cluster piC-1 (TMEM181) (Figures S6F and 5B). Ten percent of shared piCs exhibited a dynamic signature length, producing piRNAs with a preferred length of 30 nt in adult and 26 nt in juvenile testes (Figures 6F and 6G). These "dynamic piCs" constitute a novel class of piRNA clusters that generate pre-pachytene piRNAs in germline stem cells and pachytene piRNAs in meiotic and post-meiotic spermatocytes during different stages of life.

In summary, considering piRNA length, indicative of associated PIWI proteins, and spatiotemporal expression of PIWIL1–4 in testes, we untangled different piRNA pathways in adult human testes and identified three classes of piRNA clusters (Figure 6H). Adult-specific piRNA clusters produce PIWIL1-piRNAs in meiotic and post-meiotic germ cells. Shared piRNA clusters are active in juvenile and adult testes and comprise two types: SSC-piCs specifically produce pre-pachytene piRNAs in undifferentiated spermatogonial stem cells, while dynamic piCs also produced PIWIL1-piRNAs in meiotic and post-meiotic germ cells in the adult. piRNA clusters of all three classes contributed significantly to adult piRNA population and could not be deconvoluted by their

(F and G) (F) Dynamic piCs produce piRNAs with different length profiles in juvenile and adult testes. Representative example piC (G).
(H) Three classes of piRNA clusters in human testes. Shared SSC-piCs produce piRNAs in SSCs with a PIWIL2/4 length profile in juvenile and adult. Dynamic piCs produce piRNAs with a PIWIL2/4 length profile in juvenile and with a PIWIL1 length profile in adult. Adult-specific piCs are only present in adult testes and associate with PIWIL1.

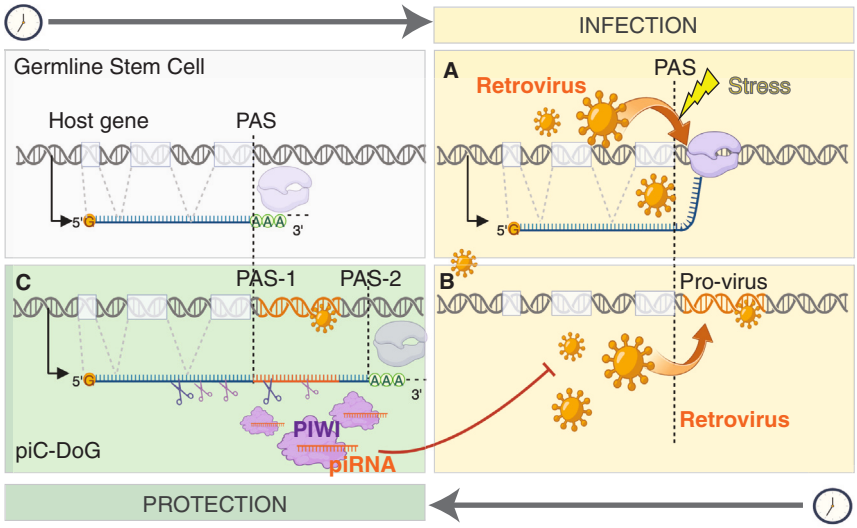


Figure 7. A model for the origin of piRNA clusters in response to retroviral invasion of germline stem cells

Retroviral infection of germline stem cells induces pervasive readthrough transcription (stress DoGs) (A) and inserts proviruses into the genome (B). A readthrough transcript that captures viral antisense information produces protective piRNAs, allowing the germ cell to survive (C). Surviving germ cells contribute to new generations and maintain readthrough transcription across the protective piC-DoG long after acute stress has been resolved.

piRNA production (Figures 6A and S6G). Results from our analyses revealed that the complexity of piRNAs in adult testes results from diverse piRNA pathways in different cell types that represent distinct stages of spermatogenesis.

DISCUSSION

Variations in piRNA cluster patterns and the uneven distribution of piRNA production

With 10% or more of the entire genome, the piRNA-producing genomic space exceeds that of protein-coding genes.[7,94] However, most piRNAs originate from only a few clusters.[14] Prioritizing piRNA clusters is essential to characterize their pathways and identify outstanding piCs for in-depth analysis. Based on their essential function in transposon silencing, ranking mammalian pre-pachytene piRNA clusters by their production of TEas piRNAs identified unique piRNA clusters that—like flamenco in flies—captured sequences of densely packed transposon fragments across tens of kilobases. Unlike flam, these piRNA clusters often resided DoG with piRNA precursors originating from readthrough transcription. These piC-DoGs illustrate common patterns (flam-like super clusters) and species-specific variations (readthrough transcription and spliced piRNA precursors) of piRNA clusters. With piCB as a versatile tool to characterize piRNA clusters in different organisms, future studies are bound to reveal more patterns and variations across species.

Little is known about the promoters of transposon-silencing piRNA clusters and the gene models of their piRNA precursors

Promoters of pachytene piRNA clusters, which are not linked to transposon regulation, have been identified.[27,30,31,50,95,96] However, the promoters serving transposon-silencing piRNA clusters remain largely elusive. *Drosophila* flamenco is thought to be transcribed from a single promoter, as suggested by the elimination of flam piRNAs through a single P element insertion downstream of the DIP-1 gene.[14,20,21,42] However, with advances in genome assembly revealing flam's potential length of about 500 kb; future analyses might consider multiple promoters, potentially coordinated by essential enhancer elements in the DIP-1 downstream region.[42]

Transposon promoters were suggested to contribute to the generation of piRNA precursors in mice, and retroviral genomes are directly processed into piRNAs in koala.[22,63] Here, we identified a mechanism that harnesses host gene promoters to transcribe piC-DoGs in mammals. We observe dense transposon fragments in these piC-DoGs immediately downstream of their associated host gene and extending across kilobases. The impeccable precision of these insertions to avoid damage to the host gene's coding sequence and splicing pattern suggests the powerful action of purifying selection. While piC-DoGs seem to produce mostly the longer pre-piRNA transcript in gonocytes, the same genes produce the annotated shorter host transcripts in other cell types. Future studies are geared at identifying the molecular mechanisms underlying the regulation of transcriptional readthrough at these specific genes.

A plausible model for the formation of protective piRNA clusters in response to retroviral infection in germline stem cells

Stress-induced transcriptional readthrough has been observed in somatic cells *ex vivo*, although its function remains unclear.[72,78] One possible reason for the cells' ability to produce stress DoGs might be found in germline stem cells. Here, we propose a plausible model for the formation of piC-DoGs in response to retroviral stress (Figure 7). Retroviral infection or the activity of mobile genetic elements could induce stress DoGs in germline stem cells (Figure 7A). The insertion of new proviruses and the associated DNA damage might contribute the stress-induced transcriptional readthrough. At the same time, proviral insertions add sequence information about the invader into the host genome (Figure 7B). On rare occasions, that readthrough transcription that captures viral sequences generates piRNAs that can silence the invading virus, allowing the germ cell to survive (Figure 7C). These surviving germ cells give rise to future generations. To ensure the persistence of protective piRNAs, purifying selection maintains readthrough transcription at protective piC-DoGs long after the initial stress has been resolved. Over time, genomic recombination and other mechanisms could add and remove sequences to form the massive piRNA clusters observed today.[26,97]

piC-DoGs might have originated as stress DoGs but have evolved into "watchdogs." Their observable remnants reveal ancient battles between genome invaders and the piRNA-guided defense, battles that the piRNA pathway has ultimately won.

Limitations of the study

In mammals, piRNA-guided epigenetic silencing of transposons occurs in a rare population of transient germline stem cells and is best studied in murine gonocytes at embryonic day 12.5.[1,10,33] The rarity and inaccessibility of these embryonic germ cells for *ex vivo* manipulation hampers in-depth experimental analysis. While readthrough transcripts across piC-DoGs are readily observed in RNA-seq data of sorted gonocytes (Figure 4), our extended evolutionary model for the origin of piRNA precursors from stress DoGs remains speculative (Figure 7). To date, there is no experimental model for the endogenization of a retrovirus into a germline genome in mammals, and stress DoGs have only been studied *ex vivo* in somatic cell culture.[72,78] If the physiological relevance of stress DoGs occurs during embryonic development, appropriate *in vivo* models need to be developed.

The current germline invasion of a retrovirus into the koala genome offers a unique glimpse into mammalian evolution. A study observing the piRNA pathway's acute response has started to illuminate the molecular events during this critical evolutionary event.[22] Novel piRNA precursors and other retroviral insertions will permanently alter the genomes of these animals. The success of the piRNA pathway will determine the survival and genetic make-up of future generations.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Astrid D. Haase (astrid.haase@nih.gov).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Small-RNA sequencing data generated in this study have been deposited at GEO (GSE259230) and are publicly available as of the date of publication. This manuscript analyzes existing, publicly available data. DOIs and accession numbers are listed in the key resources table.
- Computational analyses are documented in Data S2 (supplemental code) and are available on GitHub (https://github.com/HaaseLab/PICB and https://zenodo.org/doi/10.5281/zenodo.13376884).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We would like to thank all members of the Haase, van Wolfswinkel, Miska, and Svoboda laboratories for helpful discussions. We are particularly grateful to Pavol Genzor and Daniel Stoyko for their work on an earlier version of piCB, Valeria Buccheri and Simin Sakaki for help with the preparation of slug piRNA sequencing data, Florencia Pratto for valuable advice on mouse and human germ cell biology, the Hafner and Macfarlan groups for discussions, Qingcai Meng and Angel Jaimes for critical comments on the manuscript, and the piRNA community for their support. The graphical abstract was generated with help from Erina He and the NIH medical arts department. BioRender was used for illustrations. A.D.H.'s research group is supported by the intramural research program of the National Institute of Diabetes and Digestive and Kidney Diseases (ZIA DK075111-07). P.S.'s research group was supported by the Ministry of Education, Youth, and Sports (MEYS) of the Czech Republic project NPU1 LO1419. M.V.A. is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement no. 101027241. E.A.M. is supported by the following grants: Wellcome Trust Senior Investigator Award (219475/Z/19/Z) and CRUK award (C13474/A27826). The authors also acknowledge core funding to the Gurdon Institute from Wellcome (092096/Z/10/Z and 203144/Z/16/Z) and CRUK (C6946/A24843). J.C.v.W. is supported by NIH grants R35GM128619 and R01AG078926 and the Vallee Foundation.

AUTHOR CONTRIBUTIONS

P.K., Z.L., F.A., A.F., and A.D.H. conceived and led the project. P.K., Z.L., F.A., and A.F. conducted experiments that are presented in Figures 1, 2, 4, 5, and 6. M.V.A., A.P., and F.H. generated data and performed analyses for planaria, cichlid fish, and slug under the mentorship of E.A.M., J.C.v.W., and P.S. (Figure 3). Y.W. and H.L. assisted with computational analyses and cod optimization. A.F. is co-mentored by W.L. A.D.H. wrote the manuscript with input from P.K., Z.L., F.A., and A.F. All authors contributed to the discussions and commented on the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - *Arion vulgaris* (Spanish slug)
- METHOD DETAILS
  - RNA libraries
  - Small RNA library preparation
  - Processing small RNA sequencing data
  - Mapping small RNA libraries to the reference genome
  - Analyses of ping-pong signatures
  - Mapping total RNA-seq libraries to the reference genome
  - Published clusters/precursors used for comparison
  - Prediction of piRNA clusters by PICB
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Optimization of PICB parameters per sample
  - Unambiguous attribution of piRNA reads to piCB-clusters
  - Ranking of piCB clusters
  - Estimation of cluster reproducibility
  - Evaluation of cluster strand preferences (strandedness)
  - Evaluation of cluster nucleotide biases and their effect on productivity
  - Estimating mappability of genomic loci
  - Identifying potentially spliced mouse pachytene piRNA clusters
  - Identifying bidirectional piRNA clusters and precursors
  - Identifying clusters linked to upstream annotated genes
  - Transcriptional readthrough analysis (down-stream-of-genes transcripts, DoGs)
  - Human testis disentanglement of piRNA clusters
  - Human PIWI expression analysis
  - Human piRNA length distribution analysis for PIWI expression groups

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2024.114777.

Received: May 2, 2024
Revised: August 9, 2024
Accepted: September 4, 2024
Published: September 19, 2024

REFERENCES

1. Ozata, D.M., Gainetdinov, I., Zoch, A., O'Carroll, D., and Zamore, P.D. (2019). PIWI-interacting RNAs: small RNAs with big functions. Nat. Rev. Genet. *20*, 89–108. https://doi.org/10.1038/s41576-018-0073-3.
2. Aravin, A.A., Hannon, G.J., and Brennecke, J. (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science *318*, 761–764. https://doi.org/10.1126/science.1146484.
3. Onishi, R., Yamanaka, S., and Siomi, M.C. (2021). piRNA- and siRNA-mediated transcriptional repression in Drosophila, mice, and yeast: new insights and biodiversity. EMBO Rep. *22*, e53062. https://doi.org/10.15252/embr.202153062.
4. Yamashiro, H., and Siomi, M.C. (2018). PIWI-Interacting RNA in Drosophila: Biogenesis, Transposon Regulation, and Beyond. Chem. Rev. *118*, 4404–4421. https://doi.org/10.1021/acs.chemrev.7b00393.
5. Cox, D.N., Chao, A., Baker, J., Chang, L., Qiao, D., and Lin, H. (1998). A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. Genes Dev. *12*, 3715–3727. https://doi.org/10.1101/gad.12.23.3715.
6. Wang, C., and Lin, H. (2021). Roles of piRNAs in transposon and pseudogene regulation of germline mRNAs and lncRNAs. Genome Biol. *22*, 27. https://doi.org/10.1186/s13059-020-02221-x.
7. Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N. Engl. J. Med. *377*, 361–370. https://doi.org/10.1056/NEJMra1510092.
8. Nagirnaja, L., Mørup, N., Nielsen, J.E., Stakaitis, R., Golubickaite, I., Oud, M.S., Winge, S.B., Carvalho, F., Aston, K.I., Khani, F., et al. (2021). Variant PNLDC1, Defective piRNA Processing, and Azoospermia. N. Engl. J. Med. *385*, 707–719. https://doi.org/10.1056/NEJMoa2028973.
9. Wang, X., Ramat, A., Simonelig, M., and Liu, M.F. (2023). Emerging roles and functional mechanisms of PIWI-interacting RNAs. Nat. Rev. Mol. Cell Biol. *24*, 123–141. https://doi.org/10.1038/s41580-022-00528-0.
10. Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E.L., Fabry, M.H., Kneuss, E., and Hannon, G.J. (2018). piRNA-Guided Genome Defense: From Biogenesis to Silencing. Annu. Rev. Genet. *52*, 131–157. https://doi.org/10.1146/annurev-genet-120417-031441.
11. Haase, A.D. (2022). An introduction to PIWI-interacting RNAs (piRNAs) in the context of metazoan small RNA silencing pathways. RNA Biol. *19*, 1094–1102. https://doi.org/10.1080/15476286.2022.2132359.
12. Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat. Rev. Genet. *12*, 615–627. https://doi.org/10.1038/nrg3030.
13. Aravin, A., Gaidatzis, D., Pfeffer, S., Lagos-Quintana, M., Landgraf, P., Iovino, N., Morris, P., Brownstein, M.J., Kuramochi-Miyagawa, S., Nakano, T., et al. (2006). A novel class of small RNAs bind to MILI protein in mouse testes. Nature *442*, 203–207. https://doi.org/10.1038/nature04916.
14. Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G.J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell *128*, 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043.
15. Girard, A., Sachidanandam, R., Hannon, G.J., and Carmell, M.A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature *442*, 199–202. https://doi.org/10.1038/nature04917.
16. Houwing, S., Kamminga, L.M., Berezikov, E., Cronembold, D., Girard, A., van den Elst, H., Filippov, D.V., Blaser, H., Raz, E., Moens, C.B., et al. (2007). A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. Cell *129*, 69–82. https://doi.org/10.1016/j.cell.2007.03.026.
17. Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. (2006). A distinct small RNA pathway silences selfish genetic elements in the germline. Science *313*, 320–324. https://doi.org/10.1126/science.1129333.
18. Grivna, S.T., Beyret, E., Wang, Z., and Lin, H. (2006). A novel class of small RNAs in mouse spermatogenic cells. Genes Dev. *20*, 1709–1714. https://doi.org/10.1101/gad.1434406.
19. Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367. https://doi.org/10.1126/science.1130164.
20. Desset, S., Meignin, C., Dastugue, B., and Vaury, C. (2003). COM, a heterochromatic locus governing the control of independent endogenous retroviruses from Drosophila melanogaster. Genetics *164*, 501–509. https://doi.org/10.1093/genetics/164.2.501.
21. Pelisson, A., Song, S.U., Prud'homme, N., Smith, P.A., Bucheton, A., and Corces, V.G. (1994). Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. EMBO J. *13*, 4401–4411. https://doi.org/10.1002/j.1460-2075.1994.tb06760.x.
22. Yu, T., Koppetsch, B.S., Pagliarani, S., Johnston, S., Silverstein, N.J., Luban, J., Chappell, K., Weng, Z., and Theurkauf, W.E. (2019). The piRNA Response to Retroviral Invasion of the Koala Genome. Cell *179*, 632–643.e12. https://doi.org/10.1016/j.cell.2019.09.002.
23. Srivastav, S., Feschotte, C., and Clark, A.G. (2023). Rapid evolution of piRNA clusters in the Drosophila melanogaster ovary. Preprint at bioRxiv. https://doi.org/10.1101/2023.05.08.539910.
24. Yamanaka, S., Siomi, M.C., and Siomi, H. (2014). piRNA clusters and open chromatin structure. Mob. DNA *5*, 22. https://doi.org/10.1186/1759-8753-5-22.
25. Wierzbicki, F., and Kofler, R. (2023). The composition of piRNA clusters in Drosophila melanogaster deviates from expectations under the trap model. BMC Biol. *21*, 224. https://doi.org/10.1186/s12915-023-01727-7.
26. Assis, R., and Kondrashov, A.S. (2009). Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. Proc. Natl. Acad. Sci. USA *106*, 7079–7082. https://doi.org/10.1073/pnas.0900523106.
27. Li, X.Z., Roy, C.K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B.W., Xu, J., Moore, M.J., Schimenti, J.C., Weng, Z., and Zamore, P.D. (2013). An ancient transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. Mol. Cell *50*, 67–81. https://doi.org/10.1016/j.molcel.2013.02.016.
28. Sarot, E., Payen-Groschêne, G., Bucheton, A., and Pélisson, A. (2004). Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. Genetics *166*, 1313–1321. https://doi.org/10.1534/genetics.166.3.1313.
29. Adashev, V.E., Kotov, A.A., Bazylev, S.S., Shatskikh, A.S., Aravin, A.A., and Olenina, L.V. (2020). Stellate Genes and the piRNA Pathway in Speciation and Reproductive Isolation of Drosophila melanogaster. Front. Genet. *11*, 610665. https://doi.org/10.3389/fgene.2020.610665.
30. Choi, H., Wang, Z., and Dean, J. (2021). Sperm acrosome overgrowth and infertility in mice lacking chromosome 18 pachytene piRNA. PLoS Genet. *17*, e1009485. https://doi.org/10.1371/journal.pgen.1009485.
31. Wu, P.H., Fu, Y., Cecchini, K., Özata, D.M., Arif, A., Yu, T., Colpan, C., Gainetdinov, I., Weng, Z., and Zamore, P.D. (2020). The evolutionarily conserved piRNA-producing locus pi6 is required for male mouse fertility. Nat. Genet. *52*, 728–739. https://doi.org/10.1038/s41588-020-0657-7.
32. Gebert, D., Neubert, L.K., Lloyd, C., Gui, J., Lehmann, R., and Teixeira, F.K. (2021). Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. Mol. Cell *81*, 3965–3978.e5. https://doi.org/10.1016/j.molcel.2021.07.011.

33. Loubalova, Z., Konstantinidou, P., and Haase, A.D. (2023). Themes and variations on piRNA-guided transposon control. Mob. DNA *14*, 10. https://doi.org/10.1186/s13100-023-00298-2.

34. Stein, C.B., Genzor, P., Mitra, S., Elchert, A.R., Ipsaro, J.J., Benner, L., Sobti, S., Su, Y., Hammell, M., Joshua-Tor, L., and Haase, A.D. (2019). Decoding the 5' nucleotide bias of PIWI-interacting RNAs. Nat. Commun. *10*, 828. https://doi.org/10.1038/s41467-019-08803-z.

35. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G.J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. Science *316*, 744–747. https://doi.org/10.1126/science.1142612.

36. Gunawardane, L.S., Saito, K., Nishida, K.M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M.C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. Science *315*, 1587–1590. https://doi.org/10.1126/science.1140494.

37. Han, B.W., Wang, W., Zamore, P.D., and Weng, Z. (2015). piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. Bioinformatics *31*, 593–595. https://doi.org/10.1093/bioinformatics/btu647.

38. Genzor, P., Konstantinidou, P., Stoyko, D., Manzourolajdad, A., Marlin Andrews, C., Elchert, A.R., Stathopoulos, C., and Haase, A.D. (2021). Cellular abundance shapes function in piRNA-guided genome defense. Genome Res. *31*, 2058–2068. https://doi.org/10.1101/gr.275478.121.

39. Watanabe, T., Cui, X., Yuan, Z., Qi, H., and Lin, H. (2018). MIWI2 targets RNAs transcribed from piRNA-dependent regions to drive DNA methylation in mouse prospermatogonia. EMBO J. *37*, e95329. https://doi.org/10.15252/embj.201695329.

40. Niki, Y., Yamaguchi, T., and Mahowald, A.P. (2006). Establishment of stable cell lines of Drosophila germ-line stem cells. Proc. Natl. Acad. Sci. USA *103*, 16325–16330. https://doi.org/10.1073/pnas.0607435103.

41. Saito, K., Inagaki, S., Mituyama, T., Kawamura, Y., Ono, Y., Sakota, E., Kotani, H., Asai, K., Siomi, H., and Siomi, M.C. (2009). A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. Nature *461*, 1296–1299. https://doi.org/10.1038/nature08501.

42. van Lopik, J., Alizada, A., Trapotsi, M.A., Hannon, G.J., Bornelöv, S., and Czech Nicholson, B. (2023). Unistrand piRNA clusters are an evolutionarily conserved mechanism to suppress endogenous retroviruses across the Drosophila genus. Nat. Commun. *14*, 7337. https://doi.org/10.1038/s41467-023-42787-1.

43. Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell *137*, 522–535. https://doi.org/10.1016/j.cell.2009.03.040.

44. Nishida, K.M., Okada, T.N., Kawamura, T., Mituyama, T., Kawamura, Y., Inagaki, S., Huang, H., Chen, D., Kodama, T., Siomi, H., and Siomi, M.C. (2009). Functional involvement of Tudor and dPRMT5 in the piRNA processing pathway in Drosophila germlines. EMBO J. *28*, 3820–3831. https://doi.org/10.1038/emboj.2009.365.

45. Deng, W., and Lin, H. (2002). miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. Dev. Cell *2*, 819–830. https://doi.org/10.1016/s1534-5807(02)00165-x.

46. Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., Antony, C., Sachidanandam, R., and Pillai, R.S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. Nature *480*, 264–267. https://doi.org/10.1038/nature10672.

47. Vourekas, A., Zheng, Q., Alexiou, P., Maragkakis, M., Kirino, Y., Gregory, B.D., and Mourelatos, Z. (2012). Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. Nat. Struct. Mol. Biol. *19*, 773–781. https://doi.org/10.1038/nsmb.2347.

48. Li, X.Z., Roy, C.K., Moore, M.J., and Zamore, P.D. (2013). Defining piRNA primary transcripts. Cell Cycle *12*, 1657–1658. https://doi.org/10.4161/cc.24989.

49. Yu, T., Fan, K., Özata, D.M., Zhang, G., Fu, Y., Theurkauf, W.E., Zamore, P.D., and Weng, Z. (2021). Long first exons and epigenetic marks distinguish conserved pachytene piRNA clusters from other mammalian genes. Nat. Commun. *12*, 73. https://doi.org/10.1038/s41467-020-20345-3.

50. Ozata, D.M., Yu, T., Mou, H., Gainetdinov, I., Colpan, C., Cecchini, K., Kaymaz, Y., Wu, P.H., Fan, K., Kucukural, A., et al. (2020). Evolutionarily conserved pachytene piRNA loci are highly divergent among modern humans. Nat Ecol Evol *4*, 156–168. https://doi.org/10.1038/s41559-019-1065-1.

51. Gainetdinov, I., Colpan, C., Arif, A., Cecchini, K., and Zamore, P.D. (2018). A Single Mechanism of Biogenesis, Initiated and Directed by PIWI Proteins, Explains piRNA Production in Most Animals. Mol. Cell *71*, 775–790.e5. https://doi.org/10.1016/j.molcel.2018.08.007.

52. Palakodeti, D., Smielewska, M., Lu, Y.C., Yeo, G.W., and Graveley, B.R. (2008). The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. RNA *14*, 1174–1186. https://doi.org/10.1261/rna.1085008.

53. Reddien, P.W., Oviedo, N.J., Jennings, J.R., Jenkin, J.C., and Sánchez Alvarado, A. (2005). SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. Science *310*, 1327–1330. https://doi.org/10.1126/science.1116110.

54. Li, D., Taylor, D.H., and van Wolfswinkel, J.C. (2021). PIWI-mediated control of tissue-specific transposons is essential for somatic cell differentiation. Cell Rep. *37*, 109776. https://doi.org/10.1016/j.celrep.2021.109776.

55. Salzburger, W. (2018). Understanding explosive diversification through cichlid fish genomics. Nat. Rev. Genet. *19*, 705–717. https://doi.org/10.1038/s41576-018-0043-9.

56. Svardal, H., Salzburger, W., and Malinsky, M. (2021). Genetic Variation and Hybridization in Evolutionary Radiations of Cichlid Fishes. Annu. Rev. Anim. Biosci. *9*, 55–79. https://doi.org/10.1146/annurev-animal-061220-023129.

57. Vernaz, G., Hudson, A.G., Santos, M.E., Fischer, B., Carruthers, M., Shechonge, A.H., Gabagambi, N.P., Tyers, A.M., Ngatunga, B.P., Malinsky, M., et al. (2022). Epigenetic divergence during early stages of speciation in an African crater lake cichlid fish. Nat. Ecol. Evol. *6*, 1940–1951. https://doi.org/10.1038/s41559-022-01894-w.

58. Vernaz, G., Malinsky, M., Svardal, H., Du, M., Tyers, A.M., Santos, M.E., Durbin, R., Genner, M.J., Turner, G.F., and Miska, E.A. (2021). Mapping epigenetic divergence in the massive radiation of Lake Malawi cichlid fishes. Nat. Commun. *12*, 5870. https://doi.org/10.1038/s41467-021-26166-2.

59. Brawand, D., Wagner, C.E., Li, Y.I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A.Y., Lim, Z.W., Bezault, E., et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. Nature *513*, 375–381. https://doi.org/10.1038/nature13726.

60. Almeida, M.V., Blumer, M., Yuan, C.U., Sierra, P., Price, J.L., Quah, F.X., Friman, A., Dallaire, A., Vernaz, G., Putman, A.L.K., et al. (2024). Dynamic co-evolution of transposable elements and the piRNA pathway in African cichlid fishes. Preprint at bioRxiv. https://doi.org/10.1101/2024.04.01.587621.

61. Dorler, D., Kropf, M., Laaha, G., and Zaller, J.G. (2018). Occurrence of the invasive Spanish slug in gardens: can a citizen science approach help deciphering underlying factors? BMC Ecol. *18*, 23. https://doi.org/10.1186/s12898-018-0179-7.

62. Liegertova, M., Semeradtova, A., Kocholata, M., Prusova, M., Nemcova, L., Stofik, M., Krizenecka, S., Maly, J., and Janouskova, O. (2022). Mucus-derived exosome-like vesicles from the Spanish slug (Arion vulgaris): taking advantage of invasive pest species in biotechnology. Sci. Rep. *12*, 21768. https://doi.org/10.1038/s41598-022-26335-3.

63. Aravin, A.A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K.F., Bestor, T., and Hannon, G.J. (2008). A piRNA pathway primed

by individual transposons is linked to de novo DNA methylation in mice. Mol. Cell *31*, 785–799. https://doi.org/10.1016/j.molcel.2008.09.003.

64. Kluin, P.M., and de Rooij, D.G. (1981). A comparison between the morphology and cell kinetics of gonocytes and adult type undifferentiated spermatogonia in the mouse. Int. J. Androl. *4*, 475–493. https://doi.org/10.1111/j.1365-2605.1981.tb00732.x.

65. Kuramochi-Miyagawa, S., Watanabe, T., Gotoh, K., Totoki, Y., Toyoda, A., Ikawa, M., Asada, N., Kojima, K., Yamaguchi, Y., Ijiri, T.W., et al. (2008). DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. Genes Dev. *22*, 908–917. https://doi.org/10.1101/gad.1640708.

66. Pezic, D., Manakov, S.A., Sachidanandam, R., and Aravin, A.A. (2014). piRNA pathway targets active LINE1 elements to establish the repressive H3K9me3 mark in germ cells. Genes Dev. *28*, 1410–1428. https://doi.org/10.1101/gad.240895.114.

67. Yoshida, S., Sukeno, M., Nakagawa, T., Ohbo, K., Nagamatsu, G., Suda, T., and Nabeshima, Y.i. (2006). The first round of mouse spermatogenesis is a distinctive program that lacks the self-renewing spermatogonia stage. Development *133*, 1495–1505. https://doi.org/10.1242/dev.02316.

68. Yang, F., Lan, Y., Pandey, R.R., Homolka, D., Berger, S.L., Pillai, R.S., Bartolomei, M.S., and Wang, P.J. (2020). TEX15 associates with MILI and silences transposable elements in male germ cells. Genes Dev. *34*, 745–750. https://doi.org/10.1101/gad.335489.119.

69. Schopp, T., Zoch, A., Berrens, R.V., Auchynnikava, T., Kabayama, Y., Vasiliauskaite, L., Rappsilber, J., Allshire, R.C., and O'Carroll, D. (2020). TEX15 is an essential executor of MIWI2-directed transposon DNA methylation and silencing. Nat. Commun. *11*, 3739. https://doi.org/10.1038/s41467-020-17372-5.

70. Roth, S.J., Heinz, S., and Benner, C. (2020). ARTDeco: automatic readthrough transcription detection. BMC Bioinf. *21*, 214. https://doi.org/10.1186/s12859-020-03551-0.

71. Heinz, S., Texari, L., Hayes, M.G.B., Urbanowski, M., Chang, M.W., Givarkes, N., Rialdi, A., White, K.M., Albrecht, R.A., Pache, L., et al. (2018). Transcription Elongation Can Affect Genome 3D Structure. Cell *174*, 1522–1536.e22. https://doi.org/10.1016/j.cell.2018.07.047.

72. Rosa-Mercado, N.A., and Steitz, J.A. (2022). Who let the DoGs out? - biogenesis of stress-induced readthrough transcripts. Trends Biochem. Sci. *47*, 206–217. https://doi.org/10.1016/j.tibs.2021.08.003.

73. Vilborg, A., Passarelli, M.C., Yario, T.A., Tycowski, K.T., and Steitz, J.A. (2015). Widespread Inducible Transcription Downstream of Human Genes. Mol. Cell *59*, 449–461. https://doi.org/10.1016/j.molcel.2015.06.016.

74. Grosso, A.R., Leite, A.P., Carvalho, S., Matos, M.R., Martins, F.B., Vítor, A.C., Desterro, J.M.P., Carmo-Fonseca, M., and de Almeida, S.F. (2015). Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. Elife *4*, e09214. https://doi.org/10.7554/eLife.09214.

75. Rutkowski, A.J., Erhard, F., L'Hernault, A., Bonfert, T., Schilhabel, M., Crump, C., Rosenstiel, P., Efstathiou, S., Zimmer, R., Friedel, C.C., and Dölken, L. (2015). Widespread disruption of host transcription termination in HSV-1 infection. Nat. Commun. *6*, 7126. https://doi.org/10.1038/ncomms8126.

76. Bauer, D.L.V., Tellier, M., Martinez-Alonso, M., Nojima, T., Proudfoot, N.J., Murphy, S., and Fodor, E. (2018). Influenza Virus Mounts a Two-Pronged Attack on Host RNA Polymerase II Transcription. Cell Rep. *23*, 2119–2129.e2113. https://doi.org/10.1016/j.celrep.2018.04.047.

77. Tian, B., and Manley, J.L. (2017). Alternative polyadenylation of mRNA precursors. Nat. Rev. Mol. Cell Biol. *18*, 18–30. https://doi.org/10.1038/nrm.2016.116.

78. Morgan, M., Shiekhattar, R., Shilatifard, A., and Lauberth, S.M. (2022). It's a DoG-eat-DoG world-altered transcriptional mechanisms drive downstream-of-gene (DoG) transcript production. Mol. Cell *82*, 1981–1991. https://doi.org/10.1016/j.molcel.2022.04.008.

79. Vilborg, A., Sabath, N., Wiesel, Y., Nathans, J., Levy-Adam, F., Yario, T.A., Steitz, J.A., and Shalgi, R. (2017). Comparative analysis reveals genomic features of stress-induced transcriptional readthrough. Proc. Natl. Acad. Sci. USA *114*, E8362–E8371. https://doi.org/10.1073/pnas.1711120114.

80. Lv, X., Xiao, W., Lai, Y., Zhang, Z., Zhang, H., Qiu, C., Hou, L., Chen, Q., Wang, D., Gao, Y., et al. (2023). The non-redundant functions of PIWI family proteins in gametogenesis in golden hamsters. Nat. Commun. *14*, 5267. https://doi.org/10.1038/s41467-023-40650-x.

81. Ishino, K., Hasuwa, H., Yoshimura, J., Iwasaki, Y.W., Nishihara, H., Seki, N.M., Hirano, T., Tsuchiya, M., Ishizaki, H., Masuda, H., et al. (2021). Hamster PIWI proteins bind to piRNAs with stage-specific size variations during oocyte maturation. Nucleic Acids Res. *49*, 2700–2720. https://doi.org/10.1093/nar/gkab059.

82. Loubalova, Z., Fulka, H., Horvat, F., Pasulka, J., Malik, R., Hirose, M., Ogura, A., and Svoboda, P. (2021). Formation of spermatogonia and fertile oocytes in golden hamsters requires piRNAs. Nat. Cell Biol. *23*, 992–1001. https://doi.org/10.1038/s41556-021-00746-2.

83. Zhang, H., Zhang, F., Chen, Q., Li, M., Lv, X., Xiao, Y., Zhang, Z., Hou, L., Lai, Y., Zhang, Y., et al. (2021). The piRNA pathway is essential for generating functional oocytes in golden hamsters. Nat. Cell Biol. *23*, 1013–1022. https://doi.org/10.1038/s41556-021-00750-6.

84. Yang, Q., Li, R., Lyu, Q., Hou, L., Liu, Z., Sun, Q., Liu, M., Shi, H., Xu, B., Yin, M., et al. (2019). Single-cell CAS-seq reveals a class of short PIWI-interacting RNAs in human oocytes. Nat. Commun. *10*, 3389. https://doi.org/10.1038/s41467-019-11312-8.

85. Bronkhorst, A.W., and Ketting, R.F. (2018). Trimming it short: PNLDC1 is required for piRNA maturation during mouse spermatogenesis. EMBO Rep. *19*, e45824. https://doi.org/10.15252/embr.201845824.

86. Stoyko, D., Genzor, P., and Haase, A.D. (2022). Hierarchical length and sequence preferences establish a single major piRNA 3'-end. iScience *25*, 104427. https://doi.org/10.1016/j.isci.2022.104427.

87. Roovers, E.F., Rosenkranz, D., Mahdipour, M., Han, C.T., He, N., Chuva de Sousa Lopes, S.M., van der Westerlaken, L.A.J., Zischler, H., Butter, F., Roelen, B.A.J., and Ketting, R.F. (2015). Piwi proteins and piRNAs in mammalian oocytes and early embryos. Cell Rep. *10*, 2069–2082. https://doi.org/10.1016/j.celrep.2015.02.062.

88. Sasaki, T., Shiohama, A., Minoshima, S., and Shimizu, N. (2003). Identification of eight members of the Argonaute family in the human genome. Genomics *82*, 323–330. https://doi.org/10.1016/s0888-7543(03)00129-0.

89. Paniagua, R., Codesal, J., Nistal, M., Rodríguez, M.C., and Santamaría, L. (1987). Quantification of cell types throughout the cycle of the human seminiferous epithelium and their DNA content. A new approach to the spermatogonial stem cell in man. Anat. Embryol. *176*, 225–230. https://doi.org/10.1007/BF00310055.

90. Tan, K., and Wilkinson, M.F. (2019). Human Spermatogonial Stem Cells Scrutinized under the Single-Cell Magnifying Glass. Cell Stem Cell *24*, 201–203. https://doi.org/10.1016/j.stem.2019.01.010.

91. Hermann, B.P., Cheng, K., Singh, A., Roa-De La Cruz, L., Mutoji, K.N., Chen, I.C., Gildersleeve, H., Lehle, J.D., Mayo, M., Westernströer, B., et al. (2018). The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. Cell Rep. *25*, 1650–1667.e8. https://doi.org/10.1016/j.celrep.2018.10.026.

92. Guo, J., Grow, E.J., Mlcochova, H., Maher, G.J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., et al. (2018). The adult human testis transcriptional cell atlas. Cell Res. *28*, 1141–1157. https://doi.org/10.1038/s41422-018-0099-2.

93. Wang, M., Liu, X., Chang, G., Chen, Y., An, G., Yan, L., Gao, S., Xu, Y., Cui, Y., Dong, J., et al. (2018). Single-Cell RNA Sequencing Analysis Reveals Sequential Cell Fate Transition during Human Spermatogenesis.

Cell Stem Cell *23*, 599–614.e4. https://doi.org/10.1016/j.stem.2018.08.007.

94. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921. https://doi.org/10.1038/35057062.
95. Yu, T., Biasini, A., Cecchini, K., Saflund, M., Mou, H., Arif, A., Eghbali, A., de Rooij, D., Weng, Z., Zamore, P.D., and Ozata, D.M. (2022). A-MYB/TCFL5 regulatory architecture ensures the production of pachytene piRNAs in placental mammals. RNA *29*, 30–43. https://doi.org/10.1261/rna.079472.122.
96. Zhou, L., Canagarajah, B., Zhao, Y., Baibakov, B., Tokuhiro, K., Maric, D., and Dean, J. (2017). BTBD18 Regulates a Subset of piRNA-Generating Loci through Transcription Elongation in Mice. Dev. Cell *40*, 453–466.e5. https://doi.org/10.1016/j.devcel.2017.02.007.
97. Srivastav, S.P., Feschotte, C., and Clark, A.G. (2024). Rapid evolution of piRNA clusters in the Drosophila melanogaster ovary. Genome Res. *34*, 711–724. https://doi.org/10.1101/gr.278062.123.
98. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.
99. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
100. Martin, M. (2011). Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads. EMBnet. J. *17*, 10–12. https://doi.org/10.14806/ej.17.1.200.
101. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
102. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.
103. Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. *28*, 511–515. https://doi.org/10.1038/nbt.1621.
104. Antoniewski, C. (2014). Computing siRNA and piRNA overlap signatures. Methods Mol. Biol. *1173*, 135–146. https://doi.org/10.1007/978-1-4939-0931-5_12.
105. Ivankovic, M., Brand, J.N., Pandolfini, L., Brown, T., Pippel, M., Rozanski, A., Schubert, T., Grohme, M.A., Winkler, S., Robledillo, L., et al. (2023). A comparative analysis of planarian genomes reveals regulatory conservation in the face of rapid structural divergence. Preprint at bioRxiv. https://doi.org/10.1101/2023.12.22.572568.
106. Chen, Z., Doğan, Ö., Guiglielmoni, N., Guichard, A., and Schrödl, M. (2022). Pulmonate slug evolution is reflected in the de novo genome of Arion vulgaris Moquin-Tandon, 1855. Sci. Rep. *12*, 14226. https://doi.org/10.1038/s41598-022-18099-7.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Arion vulgaris ovotestis	This study	N/A
Chemicals, peptides, and recombinant proteins		
Qiazol	Biotech (Qiagen)	CAT#79306
Critical commercial assays		
NEXTFLEX® Small RNA-Seq Kit v3	PerkinElmer	NOVA-5132-06
Deposited data		
Sequencing data generated in this study	This manuscript	GEO: GSE259230
Resource website for piCB	This manuscript	https://github.com/HaaseLab/PICB
Ovarian somatic sheet cells (*Drosophila melanogaster*)_Piwi-IP small RNA	Genzor et al.[38]	GEO: GSE156058
Drosophila melanogaster ovary_small RNA	Stein et al.[34]	GEO: GSE115839
*Schmidtea mediterranea*_small RNA	Li et al.[54]	PRJNA756531
Astatotilapia calliptera testes_small RNA	Almeida et al.[60]	GEO: GSE252804
Mouse testes (P0)_PIWI-IPs small RNA	Yang et al.[68]	GEO: GSE107832
Mouse primary spermatoyctes_RNA and PIWI-IPs small RNA	Gainetdinov et al.[51]	PRJNA421205
Mouse gonocytes (E16.5)_RNA	Schopp et al.[69]	GEO: GSE150350
Mouse testes (P42)_RNA	Li et al.[27]	GEO: GSE44654
NIH 3T3-stress induced RNA	Vilborg et al.[79]	GEO: GSE98906
Golden hamster testes (P3)_RNA and PIWIL2-IP small RNA	Lv, Xiao et al.[80]	GEO: GSE217621
Human testes_HIWI-IP small RNA	Yang et al.[84]	GEO: GSE95218
Human testes_RNA and oxidized small RNA	Ozata et al.[50]	PRJNA506245
Experimental models: Organisms/strains		
Arion vulgaris (wild strain, central Bohemia, Czech Republic)	This study	N/A
Software and algorithms		
STAR version v2.7.10b	Dobin et al.[98]	https://github.com/alexdobin/STAR/releases
Samtools v1.17	Li et al.[99]	https://www.htslib.org/
FastQC v0.11.3	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Cutadapt v4.4	Martin et al.[100]	https://github.com/marcelm/cutadapt.git
ARTDeco v0.4	Roth et al.[70]	https://github.com/sjroth/ARTDeco
FeatureCounts v2.0.3	Liao et al.[101]	https://subread.sourceforge.net/featureCounts.html
R ≥ v4.2.1	https://www.r-project.org/	https://www.r-project.org/
Human Testis Atlas Browser	Guo et al.[92]	https://github.com/yueqiw/shiny_cell_browser
IGV v2.11.9	Robinson et al.[102]	https://software.broadinstitute.org/software/igv/
Cufflinks v 2.2.1	Trapnell et al.[103]	https://cole-trapnell-lab.github.io/cufflinks/
piCB	This study	https://github.com/HaaseLab/PICB and supplemental code
Cluster Analyses	This study	Supplemental code
ARTbio	Antoniewski[104]	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Arion vulgaris (Spanish slug)

Wild *Arion vulgaris* animals collected from central Bohemia (Czech Republic) were kept in plastic boxes containing coconut fiber substrate at room temperature. Slugs were sacrificed in 30% ethanol and ovotestis from adult hermaphrodite animals were immediately collected for RNA isolation.

METHOD DETAILS

RNA libraries

Publicly available and original libraries used in this paper are listed in the supplemental data (Data S1). We prepared a small RNA library from slug *Arion vulgaris* ovotestis and used publicly available small RNA and RNA libraries from *Drosophila melanogaster*, *Mus musculus*, *Mesocricetus auratus*, *Homo sapiens*, *Schmidtea mediterranea* and *Astatotilapia calliptera*.

Small RNA library preparation

For the small RNA-seq analysis of *Arion vulgaris* ovotestis, total RNA was isolated using Qiazol and small RNA libraries were prepared using the NextFlex Small-RNA-seq v3 kit (PerkinElmer). Libraries were prepared according to the manufacturer's instructions with ligation of the 3′ adapter overnight at 16°C. Libraries were separated on a 2.5% agarose gel using 1x lithium borate buffer and visualized with ethidium bromide. The 140–170 bp fraction was cut from the gel and the MinElute Gel Extraction Kit (Qiagen) was used to retrieve DNA. Final libraries were sequenced with single-end reads of 75 nucleotides using the Illumina NextSeq500/550 platform. Raw data were deposited in the Gene Expression Omnibus (GEO: GSE259230).

Processing small RNA sequencing data

Each small RNA library was processed according to the protocol used or the information provided in the respective publication. The *Drosophila* OSS Piwi-IP libraries for example, had the following structure: 5-FivePrimeAdapter–NNNNNNNN–smallRNA–NN–ThreePrimeAdapter-3. First, the constant region of the adapter sequences was trimmed from the small RNA libraries using cutadapt.[100] For libraries such as the example above, that contain unique molecular identifiers (UMIs, *N*s), trimmed reads were collapsed by sequence to eliminate PCR duplicates, and UMIs were removed using cutadapt. When analyzing total small RNA libraries from piRNA-relevant tissues (e.g., animal gonads), size-filtering may be necessary to remove abundant cellular contaminants. These may include siRNAs and miRNAs (typically <24 nucleotides) and, in some cases, longer RNA species or fragments if long reads were acquired during sequencing. A size distribution analysis of the sequencing reads is recommended in order to determine appropriate lower and upper size limits. These size filters can be applied during the initial reads processing (e.g., using cutadapt with parameters: -m 24 -M 34) or while loading bam files into piCB (e.g., using piCBload with parameter: READ.SIZE.RANGE = c(24,34)) for non-PIWI-IP samples. Alternatively, miRNAs can be removed by annotation if a reliable miRNA annotation is available for the corresponding model organism. For PIWI-IP or pre-size-filtered total small RNA libraries, we suggest using the default piCBload parameters, which import reads between 18 and 50 nucleotides long. Applied size filters for all the small-RNA libraries used in our study are reported in the supplemental data (Data S1). A comprehensive list of adjustable parameters and output columns is available in the piCB GitHub repository (https://github.com/HaaseLab/PICB).

Mapping small RNA libraries to the reference genome

Prior to genome mapping, we removed abundant cellular RNAs (rRNAs, tRNAs, snRNAs, snoRNAs, and miRNAs) by annotation where possible for Drosophila, mouse, and human samples. We then aligned sequences to the reference genomes using STAR[98] typically allowing up to 1 mismatch and 100 alignments per read. Reference genomes used were: Drosophila (Dm6), mouse (mm10), human (hg38), hamster (PRJDB10770,[81]; planaria (schMed3,[105]; slug[106] and cichlid fish (fAstCal1.2). Species-specific exceptions to standard parameters included: 1000 alignments per read for golden hamster, up to 2 mismatches for planaria and cichlid fish, and 0 mismatches with 5000 alignments per read for slug. To suppress spliced alignments, we generated genome indexes without annotation GTF files and used the STAR option –alignIntronMax 1.

Analyses of ping-pong signatures

The ping-pong signature analysis was performed using the "Small RNA signatures" toolkit from the ARTbio project (https://github.com/ARTbio/tools-artbio/tree/main/tools/small_rna_signatures).[104] BAM files were first filtered to retain only primary alignments using 'samtools view -F 256'. These filtered files were then processed with the signature.py script from the toolkit, applying the following parameters: –minquery 23 –maxquery 34 –mintarget 23 –maxtarget (23–34 nucleotides). To normalize the observed overlaps to the library size and assess the prevalence of ping-pong signatures, we calculated Reads Per Million (RPM) by multiplying the number of pairs in each position by 2 and dividing by the number of primary alignments in millions. This adjustment was necessary because 'Small RNA signatures' reports the number of pairs rather than individual reads. The resulting z-scores and RPM values per overlap position were visualized using the ggplot2 package in R. This analysis allowed us to quantify and visualize the characteristic 10-nucleotide overlap indicative of the ping-pong amplification cycle. In addition to the overlap analysis, we generated sequence

logos to visualize nucleotide preferences characteristic of ping-pong signatures, particularly the presence of an adenine (A) at position 10 of piRNAs. To generate the sequence logos, we loaded the originally sequenced piRNA sequences (not reference) in R using the PICBload function with additional parameters: GET.ORIGINAL.SEQUENCE = TRUE and IS.SECONDARY.ALIGNMENT = FALSE. We then trimmed the sequences down to their first 15 nucleotides and generated sequence logos using the R package ggseqlogo (v0.2).

Mapping total RNA-seq libraries to the reference genome

Bulk RNA-seq of mouse and human samples were mapped directly to their reference genomes (mm10, hg38) allowing up to 100 alignments with a maximum mismatch ratio of 0.05. RNA-seq data from 3-days old golden hamster testes were mapped to the reference genome (PRJDB10770,[81]) allowing up to 20 alignments with –outFilterMismatchNMax 999 option (STAR,[98]).

Published clusters/precursors used for comparison

To compare with our predicted piRNA clusters, we retrieved previously published piRNA clusters and precursors from *Drosophila*,[14] mouse ([27], as refined in[49]) and human.[50] Where necessary, we converted the published genomic coordinates to current reference genome builds (Dm6 for *Drosophila*, mm10 for mouse, and hg38 for human) using UCSC LiftOver tool.

Prediction of piRNA clusters by PICB

PiCB is an R library that enables seamless integration of piRNA clustering into bioinformatics pipelines using standard genomic R packages, such as GenomicRanges and can be easily installed on any computer running R. A piRNA dataset in the form of a coordinate-sorted bam file is the required input for PICB. The user is additionally required to provide the reference genome in one of the following formats: BSgenome, Seqinfo from GenomeInfoDb (chromosome names and lengths) or fasta. PICB dissects the reference genome into sliding windows with adjustable width and assembles piRNA clusters through a stepwise integration of the imported uniquely mapping alignments (which make up the seeds), primary multimapping and secondary alignments. Seeds and primary multimapping windows overlapping seeds merge into cores, while cores and secondary alignment windows overlapping cores merge into clusters. Standalone seeds and cores are also considered cores and clusters, respectively. In addition to the coordinates of the piRNA clusters, PICB output includes cluster information such as various productivity measures. PICB allows a wide-range of parameter adjustments to adapt to incomplete reference genomes of non-standard model organisms and to specific limitations of the input data set (such as reduced coverage and microRNA contamination). All clusters assembled by piCB in this study can be found in the supplemental table (Data S1). PiCB source code, user manual, and tables of adjustable settings are available in the supplemental code as well as on GitHub (https://github.com/HaaseLab/PICB). Future versions will be deposited in the same GitHub repository.

QUANTIFICATION AND STATISTICAL ANALYSIS

Optimization of PICB parameters per sample

To determine the optimal number of uniquely mapping reads per window (threshold) required for seed discovery, we tested thresholds ranging from 1 unique read up to 10 fragments per kilobase of transcript per million mapped reads (FPKM). We then assessed performance by plotting the fraction of total library reads explained and the genomic space occupied by the resulting clusters. Through this optimization we defined default PICBbuild parameters with sample-specific thresholds (Figures S1B, S1F, S3A, S3B, S4C, and S4D). The default threshold of 2 FPKM was used for all the samples with the following exceptions: hamster (5.9 FPKM), planaria (17.7 FPKM), mouse MIWI2 (2.8 FPKM), cichlid fish (5.88 FPKM) and slug (12.36 FPKM).

Unambiguous attribution of piRNA reads to piCB-clusters

To assign multi-mapping reads to a single predicted piRNA cluster, we developed a priority attribution system. First, we ranked all predicted clusters by the number of uniquely mapping reads overlapping each cluster. We then assigned all sequences with at least one reported alignment (primary or secondary) within the boundaries of the top-ranked cluster to that cluster. For the remaining unassigned sequences, we repeated this attribution to the next highest ranked cluster. This iterative process enabled the unambiguous allocation of multi-mapping piRNAs to our predicted clusters based on a ranked priority system.

Ranking of piCB clusters

Ranking of piRNA clusters is essential for prioritizing the most impactful piRNA sources and uncovering key characteristics that may be conserved. The ranking strategies to be used should be aligned with the objectives of the downstream investigation(s). In this study, we used several different strategies tailored to specific biological context and questions. The sum of all piRNAs per cluster was most commonly used to rank clusters. For piRNA samples enriched in multimapping sequences, we derived the sum of piRNAs per cluster after the unambiguous allocation of the multimapping reads as described earlier (Figures 1G, 3A, 3D, 3G, 4A, 6A, S1G, S6A, and S6G). For the repeat-depleted pachytene mouse piRNAs, we used the primary alignments per cluster (Figure 2A). To identify clusters deriving from longer, possibly spliced precursor transcripts, we normalized the primary alignments to cluster length and extracted the top 200 clusters (Figures 2E, 2F, and 2J). This allowed us to group short neighboring cluster-exons of similar

productivity regardless of their length. Alternatively, ranking of mouse, human and hamster pre-pachytene piRNA clusters by the sum of their transposon antisense (TEas) piRNAs, allowed us to prioritize loci that are likely to serve crucial genome defense roles (Figures 4A, 5A, and 5B).

Estimation of cluster reproducibility

To assess the reproducibility of piCB-assembled piRNA clusters, we analyzed three biological replicates of Drosophila Ovarian Somatic Cells (OSC). Using the PICBcombine function, we identified the intersection of cluster coordinates across all replicates. For each cluster in replicate 1, we calculated the percentage of its genomic space that overlaps with this three-replicate intersection. We then sorted replicate 1 clusters by productivity, defined by the sum of piRNAs per cluster after the unambiguous allocation of multimapping reads as described earlier. Using a sliding window approach (window size: 100 clusters, step size: 1 cluster), we grouped the sorted clusters and calculated the fraction of clusters within each group that had $\geq$95% of their genomic space reproduced across all three replicates. This fraction served as a probability measure of reproducibility. We used Pearson correlation to analyze the relationship between this reproducibility probability and the productivity-based group order.

Evaluation of cluster strand preferences (strandedness)

To evaluate strand biases within predicted piRNA clusters, we calculated a ratio of sense to antisense piRNAs uniquely mapping on each strand. For every cluster, we divided the number of unique sense piRNAs by the numer of unique antisense piRNAs. We depicted the distribution of ratio values across all predicted clusters in violin plots. Ratios between 0.1 and 10 were considered to represent putative dual-strand clusters while ratios above 10 were interpreted as clusters with a strong unidirectional strand preference. Weighing the ratios by cluster productivity during plotting prevented poorly expressed clusters from disproportionately impacting perceived overall strand preferences (Figures 1J, 1K, 2B, 3B, 3E, and 3H).

Evaluation of cluster nucleotide biases and their effect on productivity

To investigate potential implications of the nucleotide composition of clusters on piRNA abundance, we examined the nucleotide frequencies within the predicted cluster genomic sequences. For each predicted cluster, unique base counts for A, T, C, G, A or T (A/T) and C or G (C/G) were normalized to total cluster length. We tested for correlations between specific nucleotide frequencies and cluster productivity using the Pearson correlation test (Figures S1D, S2E, S2F, and S3F).

Estimating mappability of genomic loci

To estimate the mappability within a genomic region, we divided the selected locus into 20 nucleotide long sliding windows (sliding step = 1). We then aligned their sequences to the reference genome, allowing up to one mismatch and up to 1000 alignments per sequence. The sum of unique 20-mers was then averaged within larger 1000 nucleotide long sliding windows. The uniqueness of sliding 20-mers across the region was then visualized with a heatmap across the genomic region of interest (1 = good mappability/uniqueness, 0 = poor mappability/repetitiveness) (Figures 1H, 1L, and 4B).

Identifying potentially spliced mouse pachytene piRNA clusters

To identify pachytene piRNA clusters that may originate from the same spliced transcripts in mice, we first examined exon and intron lengths of annotated mouse protein-coding genes (RefSeq) and previously published pachytene piRNA precursors ([27], as refined in[49]). We then filtered for the top 200 most productive pachytene clusters by fragments per kilobase of transcript per million mapped reads (FPKM) and grouped clusters on the same strand located less than 30 kilobases apart (Figure 2E). This 30 kilobase cutoff marked the bimodal distribution bottleneck of inter-cluster distances and covered the average intron sizes of both protein-coding genes and published pachytene precursors (Figure 2F). To visualize and validate potential splicing within the grouped pachytene piRNA clusters (Figure 2G), we used the Cufflinks software[103] which assembles aligned RNA-seq reads into transcripts and can report both annotated and novel splice variants. For this analysis, we first remapped MILI-piRNAs to the mouse genome, this time allowing the reporting of spliced reads. Then we used all mapped reads as input for cufflinks which we ran without a reference annotation to enable the detection of novel transcripts. The resulting transcript assembly (in GTF format), is displayed in Figure 2G (yellow track). While total RNA or mRNA-sequencing data are typically used as input for Cufflinks, we ran the software with piRNA-sequencing data since our aim was to extract maximum information from existing datasets without the need for additional experiments. We used this approach as a cost-effective and resource-efficient method to test our cluster grouping rationale. The identified spliced clusters can be prioritized for downstream experimental validation if needed.

Identifying bidirectional piRNA clusters and precursors

We leveraged the broad boundaries of our computationally predicted clusters which resulted in overlaps for clusters derived from bidirectional promoters. To extract such clusters, we first identified cluster pairs transcribed from opposite strands with overlapping 5′ ends but non-overlapping 3′ ends. We then filtered out clusters where the ratio of sense to antisense piRNAs mapping to the 3′-most 70% of either cluster was less than 100. For published piRNA precursors, which lack extended ends, we first computationally extended the precursors by 2,000 nucleotides upstream of their annotated transcription start site before searching for 5′ end overlaps (Figure 2J).

Identifying clusters linked to upstream annotated genes

To determine if predicted piRNA clusters may be transcriptionally linked to annotated upstream genes, we looked for overlaps between gene 3' ends and cluster 5' ends. Specifically, we extended the 3' UTRs of annotated genes upstream by 1,000 nucleotides, or for genes without an annotated 3' UTR, we took the 3'-most 1/3 of the gene body. We then checked whether these expanded gene regions overlapped with the 5'-most 25% of each predicted piRNA cluster. Clusters with such overlap were considered to have putative transcriptional links (Figures 4E and 5B). As an adjustment for the hamster data analysis where long predicted clusters overlapped short annotated genes, we shortened the cluster regions checked for overlap to just the 5'-most 5 nucleotides of each hamster cluster (Figure 5A).

Transcriptional readthrough analysis (down-stream-of-genes transcripts, DoGs)

To systematically analyze and characterize transcriptional readthrough, we used the published ARTDeco pipeline (v0.4,[70]). Total RNA-seq from sorted mouse spermatogonia from embryonic day 16.5 (SRR11916388,[69]), primary spermatocytes (SRR7760359,[51]), and day 42 postpartum testes (SRR765631,[27]) were used as input for the ARTDeco pipeline requesting a minimum downstream of Gene (DoG) length of 2 kb and a minimum DoG coverage of 0.1 FPKM (Figures 4E–4G and S4F). We differentiated between non-expressed genes (non-expr.), expressed genes with no predicted DoGs (non-DoGs), and expressed genes with predicted DoGs (DoGs). Non-expressed genes with predicted DoGs were assigned in the 'non-expr.' group.

To evaluate whether clusters (piC) that overlap with expressed DoGs (piC-DoGs) derive from an acute stress state, we compared them to DoGs from published RNA-seq from NIH 3T3 fibroblasts treated with heat shock, osmotic stress or oxidative stress as well as untreated (GSE98906,[79]) (Figure S4E). The Welch's t test was used to assess the significance between lengths of piC-DoGs and Non-piC-DoGs (Figure 4G).

To assess the enrichment of Transposable Elements (TE) downstream of non-expressed (non-expr.), non-DoG-, non-piC-DoG-, and piC-DoG-genes, we calculated the genomic fraction covered by transposable elements (RepeatMasker) in both sense and antisense orientation relative to the transcription of the upstream gene. The enrichment was assessed across a 10-kilobase segment downstream of the 3′-end of the gene's most distal annotated isoform (Figure 4H). In addition, we assessed the TE enrichment within the assembled cluster regions of piC-DoGs. We weighted the TE-covered genomic fraction by the respective piRNA cluster productivity (primary alignments RPKM) (Figure 4I).

Human testis disentanglement of piRNA clusters

To untwine piCs predicted from published human testis small RNA-seq (SRP185903,[50]), we selected the top-ranking piCs accommodating 90% of the explained reads from one representative adult (SRR8575350, 1706 clusters) and one juvenile (SRR8575410, 8701 clusters) sample (Figures 6A, S6A, and S6G). piCs of the adult sample overlapping with those of juvenile samples were identified and termed 'shared (adult & juvenile)' and divided into two groups based on the piC's most dominant piRNA length ('shared SSC-piCs' $\leq$ 28 nt and 'shared dynamic-piCs' > 28 nt), while non-overlapping piCs were segregated and termed as 'adult-specific'.

Human PIWI expression analysis

Published unsorted human testis total RNA-seq from 18 samples (14 adult, 3 juveniles including duplicates from one juvenile, SRP185903,[50]) were mapped to the reference genome (hg38) as previously described. Read counts were generated with featureCounts (v2.0.3,[101]) and converted to Log2-transformed FPKMs. Values below 1.5 were set to 0. Euclidean distances were calculated (stats-package) using the normalized expression levels of PIWIL1 and PIWIL4 to perform complete linkage clustering (Figures 6C and 6D). This created three major PIWI expression groups. To show the prevalence of PIWI-genes and associated gene markers in specific testis developmental stages, figures and data values were extracted from the Young Adult in the Atlas Dataset in Human Testis Atlas Browser.[92]

Human piRNA length distribution analysis for PIWI expression groups

Within each expression group established in the PIWI expression analysis with total RNA-seq, the mean value and standard error for each length of the piRNA length distribution were calculated (Figure 6D). Two samples in the mid PIWIL1 group (SRR8575391, SRR8575348) and one sample in the high PIWIL1 group (SRR8575387) were excluded due to lower read count and quality (Figure 6D).