**Supplemental Methods**

**Processing of small RNA libraries**

Reads were first trimmed of the adapter sequences and quality filtered using cutadapt (v3.4)(Martin 2011). Read length distribution, sequence composition and duplication level were obtained from FastQC(Andrews 2010). Reads mapping to annotated miRNAs (Kozomara et al. 2019), other non-coding RNAs like rRNA, snRNA, tRNA, and snoRNA sequences(Hoskins et al. 2015) were removed using bowtie (-v 2 -k 1 -y –un -S) and a combined custom reference of non-coding small RNAs. The remaining 23-29 nt genome-mapping reads were retained as piRNA reads. Sequencing depth, major rRNAs quantified, and putative piRNA read depth are provided in **Supplemental Table S1.**

**Small RNA library QC analysis**

piRNA length profile of each library is plotted (**Supplemental Fig. S1A)** from FastQC-0.11.8 runs on fastq files after in-silico removal of reads mapping to rRNA, miRNA, tRNA, snoRNA and snRNAs. rRNA-derived reads in trimmed libraries were quantified by mapping reads representative rRNA sequences extracted from iso-1 reference assembly using RefSeq annotations downloaded from FlyBase (**Supplemental Fig. S1B)** (Larkin et al. 2021). Additionally, spearman correlation of trimmed reads mapped to TE consensus library are plotted for uniquely and multi-mapping reads from each library (**Supplemental Fig. S1 C,D**) to determine the reproducibility of the small RNA pool. Next, to determine the saturation in small RNA sequencing across libraries, reads were collapsed to retain unique sequences by TBr2\_collapse of NGS-toolbox (Rosenkranz et al. 2015) and each sequences’ abundance quantified in non-collapsed library (Genzor et al. 2021). Abundance of unique putative (23-29nt) piRNA sequences are plotted by their cumulative contribution to the library depth for two replicates each of strains A1 and B6 (**Supplemental Fig. S3 B,C**).

**piC annotation method**

Annotation pipeline steps are shown in **Supplemental Fig. S2A**. Genome mapping 23-29 reads, filtered from annotated non-piRNA small RNA genes, were mapped to respective genome assemblies using bowtie -n 1 -l 12 -a -m 1 -y -S and resulting unique alignments were separated from unmapped reads using samtools in bam files (Langmead et al. 2009; Li et al. 2009). bedtools *makewindows* (Quinlan and Hall 2010) was used to create 500 bp bookended windows from 7 DSPR genome assemblies(Chakraborty et al. 2019) and iso-1 reference assembly(Hoskins et al. 2015) followed by bedtools *coverage* to calculate uniquely mapped piRNA reads per million (RPM) per window from bam files. Windows with piRNA expression of 2 RPM or more were merged if located within 10 kb of each other into piRNA expression domains. RPKM values for piRNA expression were calculated for such domains (ranged from 500 bp to 330 kb).

piC annotation from merged domains of uniquely mapping piRNAs was conducted in two modes – ‘*permissive*’ and ‘*restrictive*’. First piRNA reads that uniquely map to selected domains were recovered and separated for each domain using samtools to quantify unique piRNA sequences per domain. Additionally, theoretical mappability scores of 0-1 for 25 nt reads was computed using GEM for bookended 500 bp windows for each genome assembly were recovered and separated for each domain using samtools to quantify unique piRNA sequences per domain(Derrien et al. 2012). To recover complete piCs using uniquely-mapping piRNAs only, low-mappability regions flanked by such domains were merged into a cluster and analyzed for 1U-bias and the density of unique piRNA sequences to remove false positives (Mohn et al. 2014). The *restrictive* piCs were then curated with stringent cutoffs in expression (>5 RPKM), density (8 hits/500 bp), and mappability of piRNA domains. While this method considers differential genomic mappability, it only uses uniquely mapped piRNAs which only comprise ~16-24% of total predicted piRNA reads in all libraries. Hence, piCs were also annotated using a second method, *proTRAC*, which utilizes all piRNA reads and normalizes piRNA expression by its mappability to uncover piCs even when there are no uniquely mapping piRNAs (Rosenkranz and Zischler 2012). A third method, we call “*permissive*”, is carried out similarly to *restrictive* but with lower cutoffs (2 RPKM and 2 hits/500 bp) to recover clusters from extremely low mappability regions and few uniquely mapping piRNAs.

For the *restrictive* mode of annotation - domains with at least 8 unique piRNAs per kb per million total piRNA reads were selected and then merged if interrupted by low mappability region of 10 kb or less. Similarly, for the permissive mode, domains with at least 2 piRNAs per kb per million were selected and then merged if interrupted by low mappability region of 15 kb or less. While the permissive mode had very relaxed parameters and likely produced many false positive predictions, the primary function of this mode was simply to provide unique piRNA support for the piRNAs detected from proTRAC method (see below), which utilizes multi-mapping piRNAs and predicts the majority of piRNAs in extremely low-mappability regions.

Next, we used the master list of each strain to compare their piC landscape. To do so, we lifted over the piC coordinates from their own genome assembly to the iso-1 reference genome using the NCBI remapping tool(NCBI). We found that even with relaxed mapping criteria (0.33X to 3X coverage and >70% identity) to the reference genome, only ~85-90% of all piCs from any given strain could be lifted over to the reference genome. Further inspection of piCs that failed lift-over revealed that they were relatively small clusters (500 - 5000 bp) and entirely absent in the reference genome or were large clusters (25-200 kb) that had undergone extensive structural rearrangements and therefore could not be lifted over to the reference genome using the NCBI remapping tool. To recover piCs that were >25 kb in size, apparently active in multiple strains but highly structurally variable (like *20A, 42AB* etc.), we manually identified and curated their coordinates by searching for the nearest flanking genes in their respective genome assemblies (**Supplemental** **Fig. 2B**). We combined the results of the two prediction methods from two replicates to produce a collapsed master-list of the piCs for each strain.

**Alternate *de novo* annotation of piCs by proTRAC**

The same processed small RNA libraries as described above were used for alternative piC annotation using proTRAC-v2.4.4 (Rosenkranz and Zischler 2012). Specifically, for proTRAC analysis, each library was collapsed to include only unique piRNA sequences using TBr2\_collapse of NGS-toolbox(Rosenkranz et al. 2015) and mapped to respective genomes using bowtie -n 1 -l 12 -a --best --strata --quiet -y -chunkmbs 1024. proTRAC 2.4.4 was run on sam files generated from bowtie with the following parameters -swsize 500 -swincr 100 -clsize 500 -1Tor10A 0.6 -pimin 23 -clhitsn 10 -pdens 0.2 -pti followed by removal of clusters with normalized multi-mapped piRNA coverage of 25 or less.proTRAC piC annotation was extracted from resulting clusters.gtf files.

**liftOver (remap) of piCs and genes for DSPR genome assemblies.**

piCs annotated from the above methodsfor each strain were lifted-over to the iso-1 reference genome (Release 6) using NCBI remap (NCBI) Briefly piC coordinates from the custom *restrictive* pipeline and *proTRAC* were lifted-over from each strain to iso-1. Remap parameters chosen for identification of all piCs with minimum alignment coverage of 0.3 and maximum expansion or contraction of 3X, allowing for clusters with strain-specific structural variation to be detected. Similarly, gene annotation from the reference genome was also lifted over to the DSPR genomes but with higher mapping stringency of minimum alignment coverage of 0.9.

**Manual curation and collapsed replicate annotations**

Recovery of complete piC regions primarily depends on expression and density of uniquely mapped piRNAs along the length of the cluster. Clusters identified independently from two different strains may differ in length due to natural variation in piRNA expression or density along the length of the cluster. To identify homologous clusters across strains despite changed boundaries due to structural variation in piCs, most heterochromatic piCs were examined in the IGV browser. Clusters resulting from merged bins across annotated protein-coding genes were unmerged into separate clusters. Any cluster partitioned into multiple smaller clusters due to lack of any uniquely mapping piRNAs for more than 20 kb were merged to recover the complete piC locus. All strain-genome assembly matched piC annotations from *restrictive*, *proTRAC* and *master-list* are provided in **Supplemental Table S3,** whereas curated and replicate collapsedannotations with population frequency are provided in **Supplemental Table S4.**

**Structural variant calling**

After INDEL calling from separate callers – cuteSV, svim, and Sniffles2, only insertions and deletions of >30 bp were retained. Next, for each caller, biallelic SVs with precise mapping were selected and merged from 8 samples using Survivor (Jeffares et al. 2017). While no outgroup was used to polarize insertions and deletions until the later stage, only simple (non-complex and unambiguous) were used. Genotyping for merged SVs for each caller was performed by cuteSV-1.0.13 with min\_support set as 3 reads. Genotyped calls that were supported by two SV callers or more, for which intra-strain allele frequency (AF) could be determined were then filtered and their SV length, AF, and read support averaged from results of SV callers using bedtools *merge* function. Summary of filtered and raw SV calls are in **Supplemental Table S5**. Additionally, overlapping SVs of same type with length difference by >20% or 500 bp were treated as independent events, otherwise collapsed as a same SV event. 71% of all simple SVs filtered were detected by at least two callers and 44.5% detected by all three callers. Next, *D. melanogaster* SVs were polarized by comparison to their absolute presence or absence in *D. simulans* and *D. sechelia*. Any SVs with conflicted calls between these two sister species were ignored. Filtered and polarized SV calls are reported in **Supplemental Table S6**.

**Structural variant parsing and enrichment analyses**

Filtered SVs of insertions and deletions class were used for parsing for further analysis. Most SVs are expected to have extremely low frequency of <0.1, which reflects the general deleterious nature of SVs. Since raw long read data used from published studies were from pooled sequencing of ~200 flies for DSPR strains and 60-80 flies for ISO1, only SVs with intra-strain AF of 0.2 or more were considered, to enrich for germline SVs. SV enrichment analysis was carried out using *poverlap* (<https://github.com/brentp/poverlap>) with 1000 bootstraps of random shuffling of SV calls. Mean expected overlap counts against piC coordinates were compared to expected overlap.

**Consensus TE library curation**

Putative novel TE fragments in the compiled DSPR-library were then clustered using cd-hit c 0.8 -aL 0.05 -aS 0.8 -A 80 -sc 1 -T 0 -d 1000 -g 1 -M 0 to create a non-redundant library of putative novel TEs. Post-clustering, 176 novel TE fragments remained, 52 of which were then manually curated into TE consensus sequences that could be classified into TE subclass and superfamilies. However, all 176 fragments were utilized for RepeatMasker run presented in **Fig. 4B.** Manual curation of novel TE fragments consisted of three steps. First, as step 1, fragments were searched against the Dfam (Storer et al. 2021) database of *D. melanogaster*, with e-value increased to 0.1. If >50% of the sequence could not be matched to any Dfam entry, then in step 2, conserved domains were searched for such fragment consensus in the NCBI Conserved Domain v3.19 database with e-value increased to 0.1 (Lu et al. 2020). If a conserved domain or domains was found, for example – RNaseH or Integrase, the consensus was retained in the library. In step 3, the fragment consensus name was changed to reflect the closest related family or superfamily identified for that consensus. Details of the comprehensive TE library classification are provided in **Supplemental Table S7,** andTElibrary consensus and RepeatMasker outputs are provided at https://github.com/kerogens101/Dmel\_piCs.git.

**Phylogenetic tree construction**

Maximum likelihood (ML) trees were constructed for all defragmented iso-1 TE insertions for TE families reported in **Figure 6** using the method described below. First, TE insertion sequences were extracted from the reference genome. Sequences in the range of 200-500 bp were manually examined for defragmentation if nearby insertions of the same family were present with non-overlapping sequences when compared to consensus. Sequences were then subjected to multiple sequence alignment using mafft v7.453 with the E-INS-i strategy and following parameters --adjustdirectionaccurately --maxiterate 1000. Next, ML trees were constructed using raxML-HPC with GTRGAMMA model (Stamatakis 2014). ML trees were then uploaded in R and terminal branch lengths calculated and visualized using the tidytree and ggtree R packages (Yu 2022; Yu et al. 2018).

**Normalized piRNA coverage plots**

Uniquely mapping piRNA coverage was computed from sorted and indexed bam files from genome assembly alignments of putative piRNA reads using *bamCoverage* of deepTools-3.5.1 (Ramírez et al. 2014). Normalization was performed using per million mapped miRNA reads and per million piRNA reads. Normalized coverage bedGraph files were loaded in IGV 2.9 browser and browser snapshots are presented in Fig. 2 and Supplemental Fig. S5,S6.

**TE age and coverage heatmap**

Mean percent divergence of TE insertions in piCs and flanking genomic non-piC regions was calculated using *computeMatrix* at 50bp intervals from RepeatMasker output of iso-1 genome assembly and master-list piC annotations. Profile plots was created using *plotHeatmap* of deepTools-3.5.1 (Ramírez et al. 2014). Similarly, Mean TE coverage (if TE insertion(s) present) was calculated using *computeMatrix* and heatmap and profile plot made using *plotHeatmap*.

**Multiple sequence alignment and syntenic visualizations of piCs**

*42AB* and *38C* piC sequences were extracted from respective strain assemblies, aligned and plotted using AliTV (Ankenbrand et al. 2017). Only alignments with >75% identity and >1kb in length are displayed in **Fig 2C** and **Supplemental Fig S8**.

**Supplemental references**

Andrews S, others. 2010. FastQC: a quality control tool for high throughput sequence data.

Ankenbrand MJ, Hohlfeld S, Hackl T, Förster F. 2017. AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput Sci* **3**: e116.

Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691–1692. https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr174

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10**: 1–11. https://doi.org/10.1038/s41467-019-12884-1

Derrien T, Estellé J, Sola SM, Knowles DG, Raineri E, Guigó R, Ribeca P. 2012. Fast Computation and Applications of Genome Mappability. *PLoS One* **7**: e30377. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0030377

Genzor P, Konstantinidou P, Stoyko D, Manzourolajdad A, Marlin Andrews C, Elchert AR, Stathopoulos C, Haase AD. 2021. Cellular abundance shapes function in piRNA-guided genome defense. *Genome Res* **31**: 2058–2068.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, Booth BW, Pfeiffer BD, George RA, Svirskas R, et al. 2015. The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**: 445.

Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C, Bähler J, Sedlazeck FJ. 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**: 14061.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. MiRBase: From microRNA sequences to function. *Nucleic Acids Res* **47**: D155–D162. https://academic.oup.com/nar/article/47/D1/D155/5179337 (Accessed May 25, 2021).

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25. http://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-3-r25

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, Garapati P V, Goodman JL, Gramates LS, Millburn G, Strelets VB, et al. 2021. FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res* **49**: D899–D907. https://academic.oup.com/nar/article/49/D1/D899/5997437

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079. https://pubmed.ncbi.nlm.nih.gov/19505943/

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res* **48**: D265–D268.

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

Mohn F, Sienski G, Handler D, Brennecke J. 2014. The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell* **157**: 1364–1379.

NCBI. NCBI Genome Remapping Service. *https://www.ncbi.nlm.nih.gov/genome/tools/remap*.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://academic.oup.com/bioinformatics/article/26/6/841/244688

Ramírez F, Dündar F, Diehl S, Grüning BA, Manke T. 2014. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42**: W187–W191. https://academic.oup.com/nar/article/42/W1/W187/2435511

Rosenkranz D, Han C-T, Roovers EF, Zischler H, Ketting RF. 2015. Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. *Genom Data* **5**: 309. /pmc/articles/PMC4583685/

Rosenkranz D, Zischler H. 2012. proTRAC--a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**: 5. http://www.ncbi.nlm.nih.gov/pubmed/22233380.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**: 1312–3.

Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mob DNA* **12**: 2.

Yu G. 2022. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. CRC Press.

Yu G, Lam TT-Y, Zhu H, Guan Y. 2018. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree ed. F.U. Battistuzzi. *Mol Biol Evol* **35**: 3041–3043. https://academic.oup.com/mbe/article/35/12/3041/5142656.