Rapid evolution of piRNA clusters in the *Drosophila melanogaster* ovary

Satyam P. Srivastav, Cédric Feschotte and Andrew G. Clark

Genome Res. 2024 34: 711-724 originally published online May 15, 2024
Access the most recent version at doi:10.1101/gr.278062.123

Rapid evolution of piRNA clusters in the *Drosophila melanogaster* ovary

Satyam P. Srivastav, Cédric Feschotte, and Andrew G. Clark

Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA

The piRNA pathway is a highly conserved mechanism to repress transposable element (TE) activity in the animal germline via a specialized class of small RNAs called piwi-interacting RNAs (piRNAs). piRNAs are produced from discrete genomic regions called piRNA clusters (piCs). Although the molecular processes by which piCs function are relatively well understood in *Drosophila melanogaster*, much less is known about the origin and evolution of piCs in this or any other species. To investigate piC origin and evolution, we use a population genomic approach to compare piC activity and sequence composition across eight geographically distant strains of *D. melanogaster* with high-quality long-read genome assemblies. We perform annotations of ovary piCs and genome-wide TE content in each strain. Our analysis uncovers extensive variation in piC activity across strains and signatures of rapid birth and death of piCs. Most TEs inferred to be recently active show an enrichment of insertions into old and large piCs, consistent with the previously proposed "trap" model of piC evolution. In contrast, a small subset of active LTR families is enriched for the formation of new piCs, suggesting that these TEs have higher proclivity to form piCs. Thus, our findings uncover processes leading to the origin of piCs. We propose that piC evolution begins with the emergence of piRNAs from individual insertions of a few select TE families prone to seed new piCs that subsequently expand by accretion of insertions from most other TE families during evolution to form larger "trap" clusters. Our study shows that TEs themselves are the major force driving the rapid evolution of piCs.

[Supplemental material is available for this article.]

Animal genomes are parasitized by a horde of transposable elements (TEs) whose mutagenic activity can lead to sterility (Bingham et al. 1982; Bucheton et al. 1984). In the animal germline, the Piwi-interacting RNA (piRNA) pathway is a conserved small RNA-based mechanism regulating TE activity (Lau et al. 2006; Brennecke et al. 2007; Houwing et al. 2007; Grimson et al. 2008). piRNAs are 23- to 32-nucleotide RNAs produced from discrete loci called piRNA clusters (piCs) that guide effector Piwi proteins to silence TEs (Ozata et al. 2019). The piRNA pathway presents features of an adaptive defense system against TE invasion (Aravin et al. 2007; Brennecke et al. 2008; Khurana et al. 2011; Yu et al. 2019), but little is known about the processes driving its evolution. Many piRNA pathway genes encoding the proteins involved in TE silencing display signatures of adaptive evolution (positive selection) in several species' lineages (Simkin et al. 2013; Yi et al. 2014; Palmer et al. 2018), which may indicate adaptation to a rapidly changing TE sequence pool and new invasions (Cosby et al. 2019). However, little is known about the means by which piRNA-producing loci, piCs, originate and evolve in flies or any other species.

piRNAs are produced from long noncoding RNA precursors that are transcribed from dispersed loci called piRNA clusters (Brennecke et al. 2007; Mohn et al. 2014). piCs occupy 0.1%–3% of the genome in fruit flies, mosquitoes, and mice and are enriched for TEs and other repeats such as satellites, as well as sometimes host gene sequences (Brennecke et al. 2007; Houwing et al. 2007; Chirn et al. 2015; Chen et al. 2021; Ma et al. 2021). The best-characterized function of piRNAs is to repress TEs. Because TE activity and composition vary significantly between and within species, TEs themselves may be important drivers of piC evolution, but this has not been thoroughly tested. TEs show high diversity in their mechanisms of transposition and their genomic distribution (Sultana et al. 2017; Wells and Feschotte 2020), as well as spatial and temporal activity (Pasquesi et al. 2020; Lawlor et al. 2021; Chang et al. 2022). Hence, it is likely that piCs evolve through diverse mechanisms to repress newly introduced TEs.

The organization of piCs is best characterized in *Drosophila melanogaster*. The genome-wide piC landscape in the *D. melanogaster* ovary is composed of tens of large (>10 kb) loci and hundreds of smaller (<10 kb) loci. It is also known that most large clusters (>10 kb) reside in pericentromeric and subtelomeric regions. Larger pericentromeric clusters consist of tens to hundreds of diverse TE insertions, whereas the small clusters (<10 kb) often contain recent TE insertions (Shpiz et al. 2014; Baumgartner et al. 2022; Miller et al. 2023). The architecture and composition of some large clusters suggest a "trap" model for the evolution of piCs, wherein TE insertions within clusters are selectively favored because of their production of piRNAs to repress all copies of the same active family (Bergman et al. 2006; Brennecke et al. 2007; Zanni et al. 2013; Kelleher et al. 2018; Kofler 2019). Over time, this process is predicted to result in enlargement of piCs, which serves as the host's archive of past TE activity. This is consistent with the observation of large piCs producing a bank of diverse piRNAs related to previously encountered TEs. It is important to note that recent studies reported bias in the ovarian piRNA sequence pool and piC loci composition toward younger TE insertions (Saint-Leandre et al. 2020; Gebert et al. 2021; Said et al. 2022). This indicates that evolutionary processes rapidly shape piC sequence composition to target the most recently active TEs.

Corresponding authors: sps257@cornell.edu, cf458@cornell.edu, ac347@cornell.edu
Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.278062.123.

In contrast to the large canonical piCs, individual TE insertions can also function as piCs (Olovnikov et al. 2013; Le Thomas et al. 2014; Akkouche et al. 2017). Recent studies have shown that piRNAs from large "trap"-like piCs are not essential for TE silencing, nor do they contribute the majority of piRNAs out of the total piRNA pool (Gebert et al. 2021; Genzor et al. 2021). Hence, piCs from individual TEs may produce a substantial quantity of piRNAs, which abrogates the need for active TEs to land into existing "trap" clusters to come under the control of the piRNA pathway. Although the "trap" model of piC evolution has received empirical support, its relative contribution to piC evolution in *D. melanogaster* is unclear, and the mechanisms underlying the origin and evolution of piCs remain broadly uncharacterized.

Results

Extensive variation in the genomic landscape of piCs

To quantify piC variation in *D. melanogaster*, we generated a comprehensive annotation of active piCs in eight highly inbred strains. Seven of these strains are derived from natural populations of distinct worldwide origins and have publicly available long-read genome assemblies (Chakraborty et al. 2019). For each of these seven strains, we constructed and sequenced libraries of small RNAs isolated from ovaries of two biological replicates sampled 6 mo apart (Supplemental Fig. S1; Supplemental Table S1). In addition, we analyzed two ovarian small RNA libraries for the reference iso-1 strain, generated as part of two independent studies (Shpiz et al. 2014; Asif-Laidin et al. 2017). Each small RNA library is analyzed separately using the pipeline discussed briefly here and in more detail in the Methods section, which lists three methods to define piCs (Supplemental Fig. S2). The *restrictive* and *proTRAC* methods serve the purpose of discovering moderately to highly expressed piCs using uniquely and multimapping piRNAs, respectively. The *permissive* method is performed mainly to validate low to moderately expressed piCs detected by *proTRAC* using multimapping piRNAs. Both *restrictive* and *proTRAC* methods yield highly reproducible piC coordinates across replicates of each strain with >88% of total piC count shared between two replicates overlapping over >75% of their respective length (Fig. 1A; Supplemental Fig. S3; Supplemental Table S3). However, interstrain pairwise comparisons reveal that pairs of strains share only an average of ~40% of their total piCs counts. The high-confidence set of piCs from *proTRAC*, which either have high expression (>25) or were supported by uniquely mapping piRNAs, along with all piCs from the *restrictive* method for each replicate are combined to create a replicate-specific "master list" of piCs.

Genome-wide visualization of piC annotations in iso-1 coordinates across chromosomes reveals variability in the piC landscape across the eight strains (Fig. 1B; Supplemental Table S4). In aggregate, the total amount of genomic DNA covered by active piCs in each strain ranged from 4.8 Mb to 6.3 Mb (Fig. 1C), encompassing 3.4%–4.8% of their respective genome assemblies (Supplemental Fig. S4C). Although their piC landscape is broadly similar in terms of being denser within peri-centromeric and telomeric heterochromatic regions (characterized by low mappability scores) compared with euchromatic regions, it is readily apparent that many individual clusters are present in only one or a few strains, even within these heterochromatic regions. Smaller piCs, generally located in but not exclusive to euchromatic regions, are characterized by higher mappability scores and appear even more variable across strains (Fig. 1B). Thus, from this broad-scale view, it appears that the total span of the genome occupied by piCs within each strain is largely similar, but the positions of piCs are highly variable across strains.

Abundant strain-specific and strain-biased piCs

To quantify the frequency of piCs across the eight strains, we scored the overlap of piCs predicted independently for each of the eight strains using the master-list coordinates. To account for changes in size of piCs among strains, we required a minimum positional overlap of only 1 bp for piCs to be considered shared between strains. Even when using this relaxed criterion, we found that 568 (*restrictive*) to 906 (*proTRAC*) of piCs are active only in a single or a few strains, confirming that each strain has a unique piC landscape (Fig. 1D). The results are similar whether we used the piC predictions of the *restrictive* and *proTRAC* methods separately or the combined master list (Fig. 1D; Supplemental Table S2). All strains contained 35–60 piCs that are strictly unique to that strain and another approximately 30 piCs (<10%) that could not be lifted-over and therefore are likely to be strain specific also (Supplemental Fig. S4A). Thus, we can conservatively estimate that each strain possesses more than 50 piCs that are not shared by any of the other seven strains examined. In addition, 142 and 69 piCs (*restrictive* pipeline) are shared between two and three strains, respectively. All such piCs, shared by four strains or less, are together termed as "rare" piCs. Rare piCs not only are extremely abundant but also show significant piRNA expression ranging from 10–25 RPM, which is comparable to previously described canonical piCs like *80EF* and the *traffic jam* 3′ UTR (Fig. 2; Supplemental Fig. S5). Additionally, despite their small size, in aggregate, strain-specific piCs contribute a substantial portion of the total genomic span of piCs (average of ~1 Mb) and 15%–20% of the total piC genomic length of each strain (Supplemental Fig. S4B).

Next, we examined the relationship between the size of piCs and their level of sharing across strains. First, we note that the piC length predicted from each library is similar, with a median from 5.7 kb to 7.5 kb (Supplemental Fig. S4C). We find that piC size is positively correlated with the level of sharing across strains, and this correlation holds true for all prediction methods (Pearson's $r = 0.56$ for proTRAC, 0.58 for restrictive, and 0.76 for master-list, $P\text{-value} < 2.2 \times 10^{-16}$) (Fig. 1E). We also verified this correlation with differing piC overlap parameters to compute sharing across strains (Supplemental Fig. S11). In other words, piCs detected in a single or a minority of the strains (rare piCs) tend to be smaller (2–10 kb) than those shared by most of the strains (common piCs). If we posit that rare piCs compared with common piCs represent evolutionarily younger piCs, this relationship suggests that piCs are born relatively small and increase in size as they get older. Alternatively, larger piCs may be more evolutionarily stable than smaller ones. We note, however, that even large piCs can still be variable in activity across strains. For example, large well-known piCs like *42AB* and *38C* are still only active in six or seven of the eight strains (see below). Taken together, these results suggest that ovarian piCs are extremely labile and poorly conserved in activity across *D. melanogaster* strains.

Extensive variability in piRNA expression of piCs

To illustrate the differences in activity of piCs among the eight strains, we examined the piRNA coverage profiles for *42AB* and *38C*, two large canonical piCs, and three small piCs with varying activity across strains in *76C* to *76E* cytogenetic regions (Fig. 2).

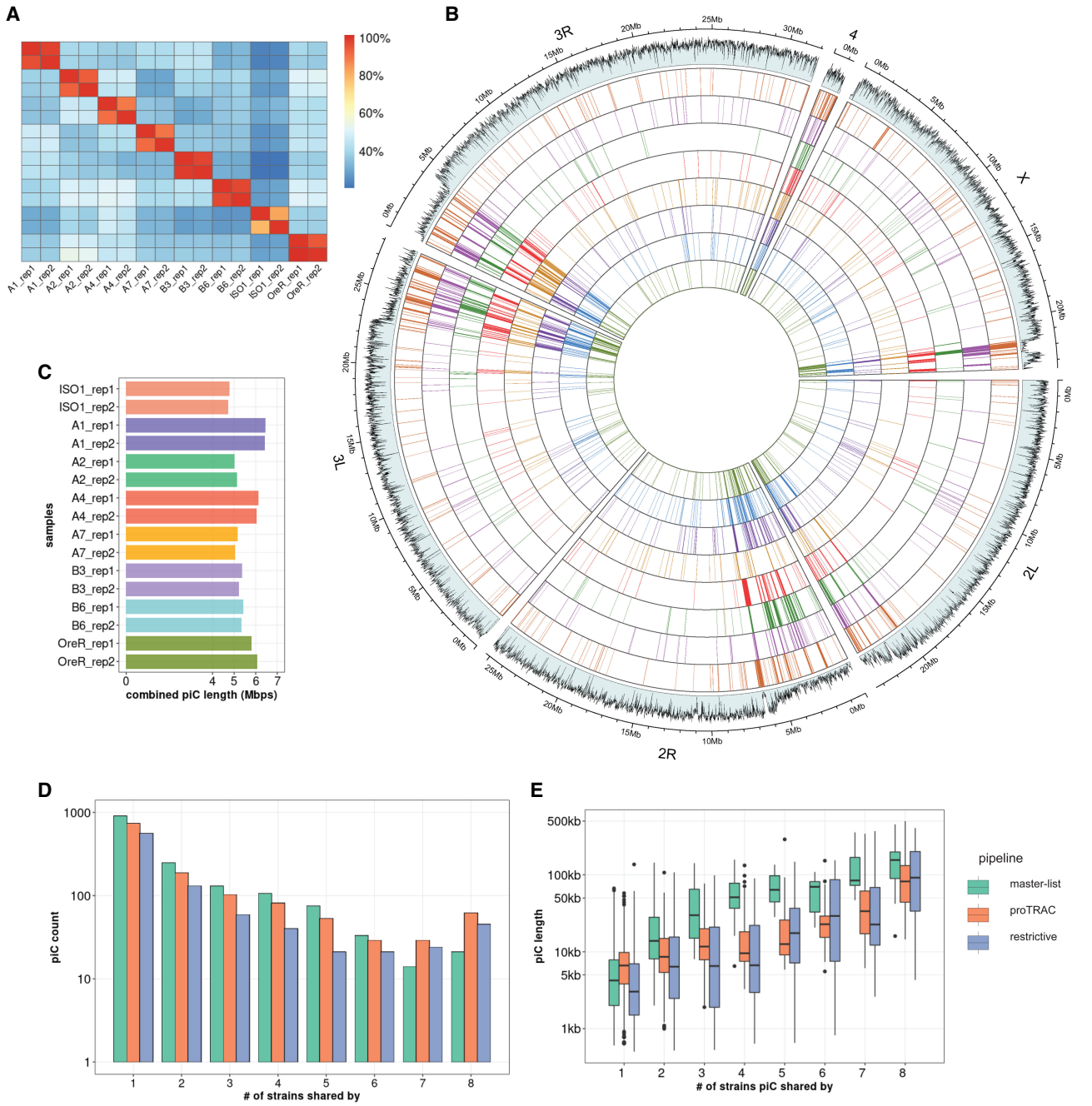

Figure 1. Interstrain variability of piCs in *D. melanogaster* strains. (*A*) Cross-strain and replicate overlap (percentage of total piCs count) of independently predicted piCs for each small RNA library using the restrictive method. Interstrain comparisons are performed after piCs were remapped to iso-1 coordinates, and overlap >1 bp between two piCs is deemed as shared. Intrastrain comparisons are performed similarly but in native genome assembly. (*B*) Genome-wide distribution of lifted-over piCs in seven DSPR strains and reference iso-1 strain. Bars along the circumference represent the presence of piCs in 10-kb bins for each chromosome. The *outermost* bar plot is piRNA mappability scores, followed by iso-1 piCs and then by piCs of seven DSPR strains. (*C*) Combined size of predicted piCs genome-wide from each replicate small RNA library independently in respective genome assemblies. (*D*) Population frequency of piCs quantified after liftover to the reference iso-1 genome. (*E*) piC length distribution by population frequency in kilobase pairs quantified after liftover to the iso-1 genome.

The *42AB* piC has been extensively documented for its high piRNA expression (Brennecke et al. 2007; Klattenhoff et al. 2009). We present normalized coverage of uniquely mapping piRNAs to per million miRNAs (Fig. 2A) for the respective *42AB* assemblies from both small RNA library replicates of four strains. Additionally, to examine differences in read coverage owing to mappability, theoretical mappability scores are visualized along the length of the cluster in 100-bp bins. Strains A1 and A7 have severely reduced (more than 20-fold) piRNA expression levels throughout *42AB* compared with the other strains, whereas iso-1 and B6 show *42AB* piRNA abundance similar to levels reported in other studies (Brennecke et al. 2007; Klattenhoff et al. 2009). Similarly, *38C*, a highly active

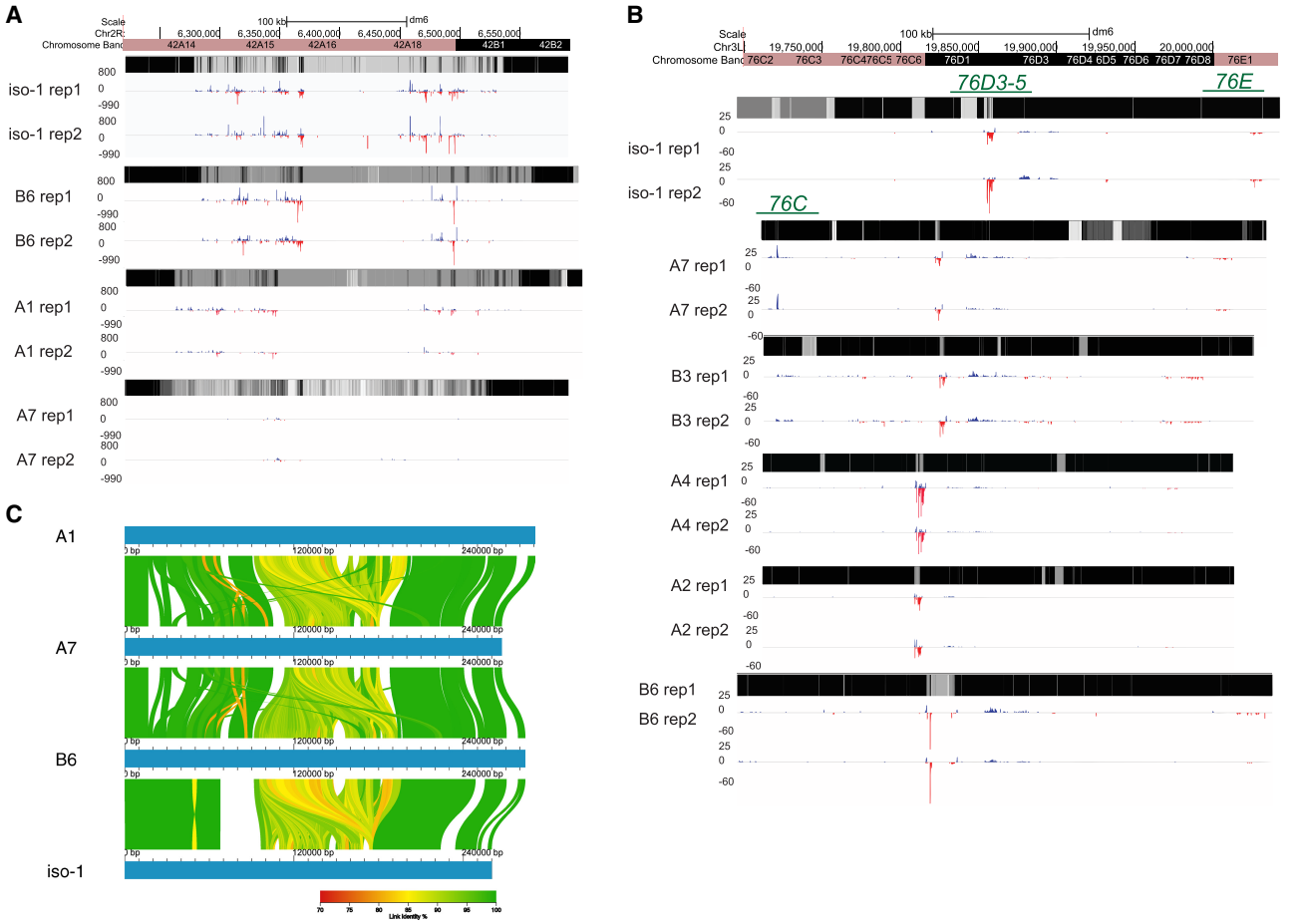

Figure 2. Natural variation in expression of uniquely mapping piRNAs from *42AB*, *76C*, and *76D*. (*A*) Uniquely mapping piRNA abundance profiles of *42AB* piCs for the four trains with two small RNA library replicates. The *y*-axis values are piRNA reads for 100-bp bins per million miRNA reads. Mappability scores (0–1) are shown for 100-bp bins of each respective *42AB* genomic assembly in the heatmap. (*B*) Uniquely mapping piRNA expression profile of *76C*, *76D*, and *76E* piCs for the six strains with two small RNA library replicates. The *y*-axis values are piRNA reads for 100-bp bins per million miRNA reads. Mappability scores (0–1) are shown for 100-bp bins of each respective *76C*, *76D*, and *76E* genomic assembly in the heatmap. (*C*) Ribbon plot of multiple-sequence alignment for *42AB* assembly in four strains. Only alignments of >80% identity and >1 kb in length are shown.

dual-stranded piC in iso-1, shows significant variability in uniquely mapping piRNAs across strains (Supplemental Fig. S5). Similar interstrain difference in normalized piRNA abundance was observed when piRNA coverage was normalized by sequencing depth, as exemplified by *38C* (Supplemental Fig. S5) and additional small piCs for strain B6 (Supplemental Fig. S6). Because *42AB* and *38C* are active in six out of eight strains, it is most parsimonious to conclude that these piCs are relatively old but have lost activity in a subset of strains.

Two piCs in the *76C*, *76D*, and *76E* cytogenetic regions are shown because each shows different levels of sharing across strains. *76C* is a ~15-kb piC, likely unistranded, exclusively active in strains A7 and B3, whereas *76D* is a dual-stranded piC ranging from 6 kb to 20 kb in size but active in all the six strains shown (Fig. 2B). *76E*, a unistranded piC with relatively low piRNA abundance, is considered active in all strains except A2. Comparison of such syntenic piCs between strains in their native genome assemblies provides validation of the variable activity of piCs across strains presented earlier from annotation pipelines (Fig. 1B). In addition, we also note that although genomic assemblies of *42AB* from different strains are similar in size, there are structural rearrangements, which may contribute to differences in sequence composition of this piC across strains (Fig. 2C; Supplemental Figs. S8, S9).

Structural variation in piCs supports aspects of the "trap" model

To understand the mutational processes underlying the changes in structure and sequences of piCs among strains, we examined the contribution of interstrain structural variants (SVs), namely, insertions and deletions (indels). We detected indels for each of the strains relative to the iso-1 reference strain (Solares et al. 2018; Chakraborty et al. 2019). We mapped raw long sequencing reads for each strain to the iso-1 genome and called indels using three independent SV callers (see Methods). Indels from all strains were then collapsed to construct a list of unique SVs that consisted of 2273 putative insertions and 4409 deletions. Indels were then polarized into "true" insertions and deletions by comparison of each variant to the *Drosophila simulans* and *Drosophila sechellia* reference strains, which enabled inference of the ancestral state (see Methods) (Supplemental Fig. S10A,B). Polarization led to loss of ~55% of indels as the ancestral or derived state of the loci could

not be determined owing to conflicts in calls between the two outgroup species. After this filtering, 1183 insertions and 1873 deletions were retained for analysis (Supplemental Table S6).

We examined the size distribution of indels overlapping piC and non-piC regions of the genome to assess indels associated with piC variation. First, genome-wide, insertions range from 30 bp to 91 kb with a median of 612 bp, whereas deletions range from 30 bp to 7.6 kb with a lower median of 208 bp compared with insertions (Fig. 3A), which is consistent with previous SV profiling of *D. melanogaster* strains (Dopman and Hartl 2007; Zichner et al. 2013; Huang et al. 2014). However, insertions overlapping piCs have a significantly larger median length (2.2 kb vs. 0.512 kb) than insertions nonoverlapping piCs (Kruskal–Wallis test, $\chi^2 = 10.812$, df = 1, *P*-value = 0.001). Meanwhile, deletions overlapping piCs are not significantly different in length than those nonoverlapping piCs (Kruskal–Wallis test, $\chi^2 = 0.72404$, df = 1, *P*-value = 0.39) (Fig. 3B). We also compared the length distribution of indels in piCs grouped by the number of strains with which they are shared. We find that strain-specific or rare piCs (shared by less than four strains) are associated with relatively large insertions (median length of 5.2 kb), whereas common piCs (shared by more than half of the strains) have a median insertion length of <1 kb (Fig. 3C). In sum, rare piCs are associated with relatively large insertions, which is consistent with the idea that these piCs emerged from recent TE insertions.

Next, we tested whether piCs are enriched for indels relative to the rest of the genome. To do this, we compared the indel counts overlapping piCs for each of the piC frequency categories with those expected based on 1000 sets of randomly shuffled indels. We find that deletions are significantly enriched in common piCs, but not in rare piCs (Fig. 3D). Insertions are strongly enriched both in rare and common piCs (Fig. 3E). These results may be confounded by the location of many piCs within constitutive heterochromatin, for which the rate of SVs is generally high (Montgomery et al. 1991; Chakraborty et al. 2021). However, we find that only ~28% of all piCs lie within constitutive heterochromatin boundaries of the reference genome assembly, and indels are significantly enriched in piCs even when we compare them

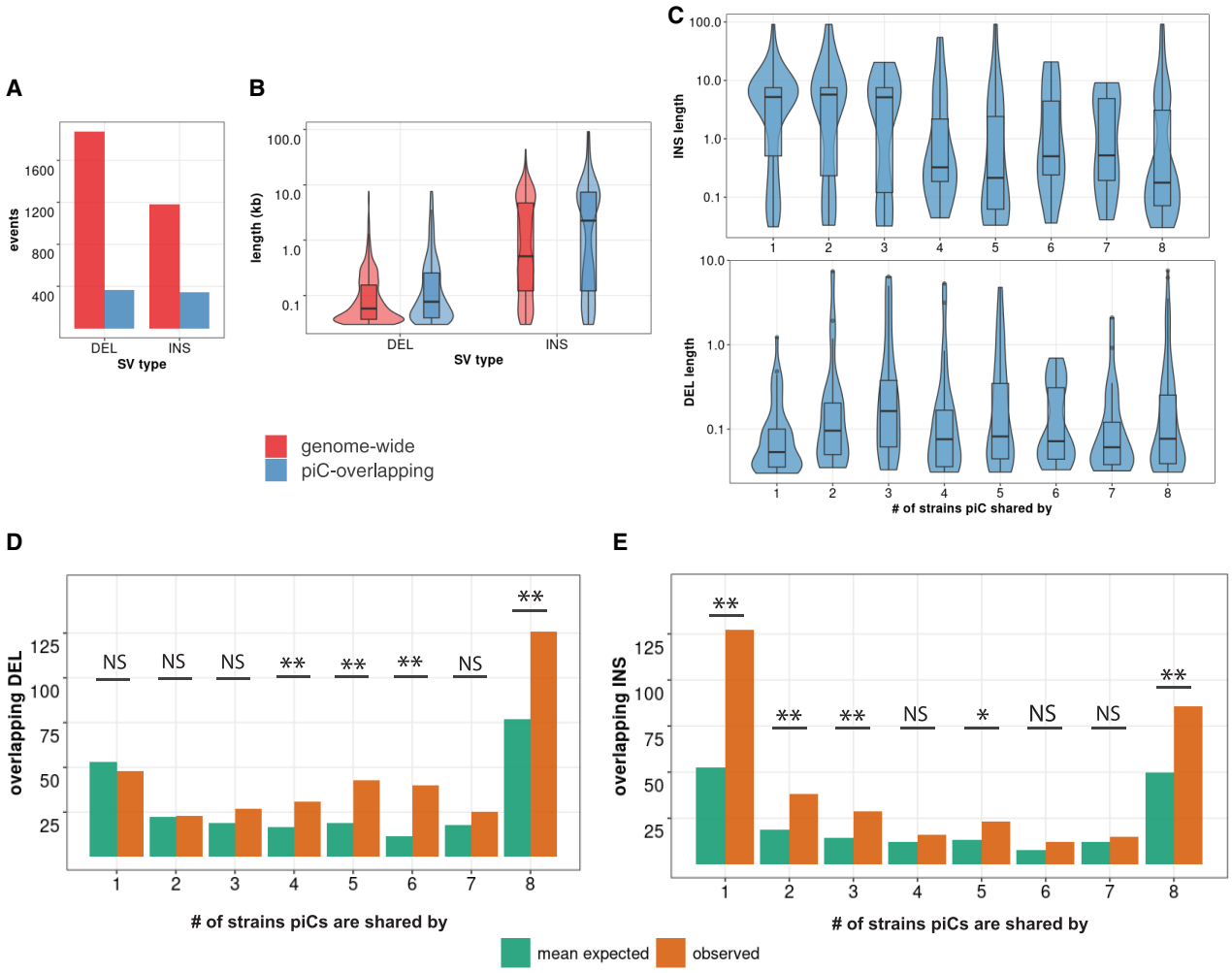

Figure 3. Common piCs show "trap" like sequence turnover. (*A*) Observed counts of indels genome-wide and overlapping with piCs. (*B*) Length distribution of indels genome-wide and overlapping with piCs. Significant differences are shown from Kruskal–Wallis test comparisons. (*C*) Length distribution of indels overlapping piCs grouped by the number of strains they are shared by (i.e., population frequency). (*D*,*E*) Enrichment analyses of deletion (DEL) and insertions (INS) variants in piCs performed. Variants overlapping with piCs of differing population frequencies are compared with the expected mean overlap counts genome-wide. Variant calls were shuffled 1000 times. (*) $P < 0.05$, (**) $P < 0.005$.

to heterochromatic regions (Supplemental Fig. S7; Riddle et al. 2011). Thus, the SV enrichment we observe within the common piCs is unlikely to be solely driven by their location within constitutive heterochromatin. Overall, we conclude that piCs are subject to a high rate of structural genomic change relative to the rest of the genome, which likely contributes to their rapid evolutionary turnover. Additionally, we find that common piCs are enriched for both insertions and deletions, which is consistent with these clusters evolving as "traps" that expand and contract. In contrast, rare piCs are only enriched for insertions, which supports the notion that these are generally young clusters born from presumably recent TE insertions.

TE reannotation of each strain uncovers ~3 Mb of unannotated TE DNA

Although our analysis supports that SV events are important drivers of piC evolution, it does not directly address the role of TE activity. To assess the contribution of TEs to the composition and changes in the activity of piCs across strains, we performed de novo annotation of TEs in each of the strain genome assemblies. This was necessary because many TE consensus sequences present in the reference TE library for *D. melanogaster* (FlyBase release 2019_05) were discovered and curated more than two decades ago using primarily the iso-1 and Oregon-R strains (Bowen and McDonald 2001; Bartolomé et al. 2002; Kaminker et al. 2002). However, recent advances in long-read sequencing technology have provided a means to obtain a more unbiased view of the repetitive landscape of *Drosophila* genomes, revealing novel TE families (Ellison and Cao 2020; Han et al. 2022; Rech et al. 2022). We developed a TE annotation pipeline based on RepeatModeler2 (for discovery) (Smit 1999; Flynn et al. 2020), RepeatMasker (for annotation), and additional tools to distinguish novel TEs from known TEs and curate a comprehensive TE library for the eight strains used in this study (see Methods) (Fig. 4A).

TE family sequence assemblies by RepeatModeler2 for each strain were aligned to the reference TEs using the 80-80-80 rule (Wicker et al. 2007). RepeatModeler2 sequences already present in the reference TE library were removed. Next, the remaining RepeatModeler2 sequences were used to remask the genomes to examine them for novel TE families. More than 5000 insertions (>500 bp in size, median of 676 bp), very similar (<5% divergence) to their respective RepeatModeler2 family consensus, were discovered for each strain (for an example of strain B6, see Fig. 4B). These novel insertions resulted in masking of an additional 2.5 Mb to 4 Mb in each genome assembly that would have been missed or misannotated as highly diverged insertions by masking with only the reference TE library (Fig. 4C). In summary, a refined and comprehensive TE library was created with a combination of 129 reference TE consensus sequences and 47 uncharacterized consensus (Supplemental Table S7) sequences that capture all TE insertions genome-wide and reflect their relative age.

Young LTR retrotransposons are enriched within rare piCs

To understand the coevolution of TE and piCs, we correlated TE insertion age and piC sharing across strains (as proxy for piC age). Using the new TE library described above, we sought to compare the age and composition of TEs within piCs to those in the rest of the genome. To infer the age of each family, we used the median sequence divergence of each insertion to their family consensus. To examine TE composition, we grouped TEs into the major subclasses and superfamilies represented in *Drosophila*: non-LTR retrotransposons (LINEs), LTR retrotransposons (*Ty1*/*copia*; *Ty3*/*mdg4*; *BEL*/*Pao* superfamilies), Rolling Circle (RC) transposons, and cut-and-paste DNA transposons. We found that TE copies from all three LTR superfamilies are significantly younger in piCs than non-piC regions (Fig. 4D). Conversely, TE copies from the LINE, RC, and DNA subclasses are not significantly different in age in piCs than in non-piC regions. To test these results using an independent method to date insertions, we built phylogenetic trees from all copies for one LTR superfamily (*Ty1*/*copia*) and one DNA transposon superfamily (*Tc1*/*mariner*) and used terminal branch lengths to estimate their relative age (Carr et al. 2012). We chose these superfamilies because they are of moderate abundance and therefore manageable for multiple sequence alignments and phylogenetic analyses. The results of these analyses yielded the same trend observed genome-wide using sequence divergence from consensus sequences: The *Ty1*/*copia* LTR retrotransposons ($n = 135$) overlapping piCs are significantly younger than nonoverlapping ones, whereas *Tc1*/*mariner* ($n = 89$) DNA transposons show no such bias (Fig. 4E).

To examine whether these trends hold at the level of individual TE families, we selected one family with moderate copy number from the LTR, LINE, and DNA subclasses and compared the age of piC overlapping and nonoverlapping copies within each family. We analyzed as a representative *Ty3*/*mdg4* LTR superfamily, *blood*, a family with 63 copies in the iso-1 strain that is known to be transpositionally active (Bingham and Chapman 1986; Kofler et al. 2015). Consistent with the trend observed at the level of the LTR superfamily, we found that 43 out of 63 *blood* insertions are associated with piCs. Most of these are very recent insertions with a median terminal branch length of <0.002, which is significantly shorter than that of insertions not overlapping with piCs (Wilcoxon rank-sum test, P-value = 0.014) (Fig. 4F). In other words, piC overlapping *blood* insertions are significantly younger than the nonoverlapping ones. We analyzed, as a representative of the *Tc1/mariner* superfamily of DNA transposons, *Tc1-2*, a family with 35 copies in the iso-1 genome. Consistent with the trend observed at the level of the entire superfamily, the age of the *Tc1-2* copies overlapping piC is not significantly different than that of non-piC overlapping copies (Wilcoxon rank-sum test, P-value = 0.903) (Fig. 4G). Analyzing the *G*-element LINE family, which counts 35 copies in iso-1 and is still active (Di Nocera et al. 1986), we found that the age of piC overlapping copies is not significantly different from nonoverlapping copies (Wilcoxon rank-sum test, P-value 0.855), and the youngest *G*-element insertions according to terminal branch length do not overlap piCs (Fig. 4H). Taken together, these results suggest that young LTR retrotransposon insertions tend to be enriched in piCs, but this trend is not observed for other TE subclasses and superfamilies.

A small subset of active LTR retrotransposon families give rise to young piCs

To test the central prediction of the "trap," in which recent transposition events from active TEs must be enriched in piCs, we established a set of nonreference TE insertions in each DSPR strain using the raw long-read data available for each. Briefly, we applied TLDR (a long-read TE insertion detection tool) (Ewing et al. 2020) with a cutoff of at least two supporting reads per 10× genome coverage to remove false positives and enrich for germline insertions (see Methods). Using these parameters, we identified 285 to 857 nonreference TE insertions in each of the seven DSPR lines but only 75 insertions for iso-1, which is expected because the reference

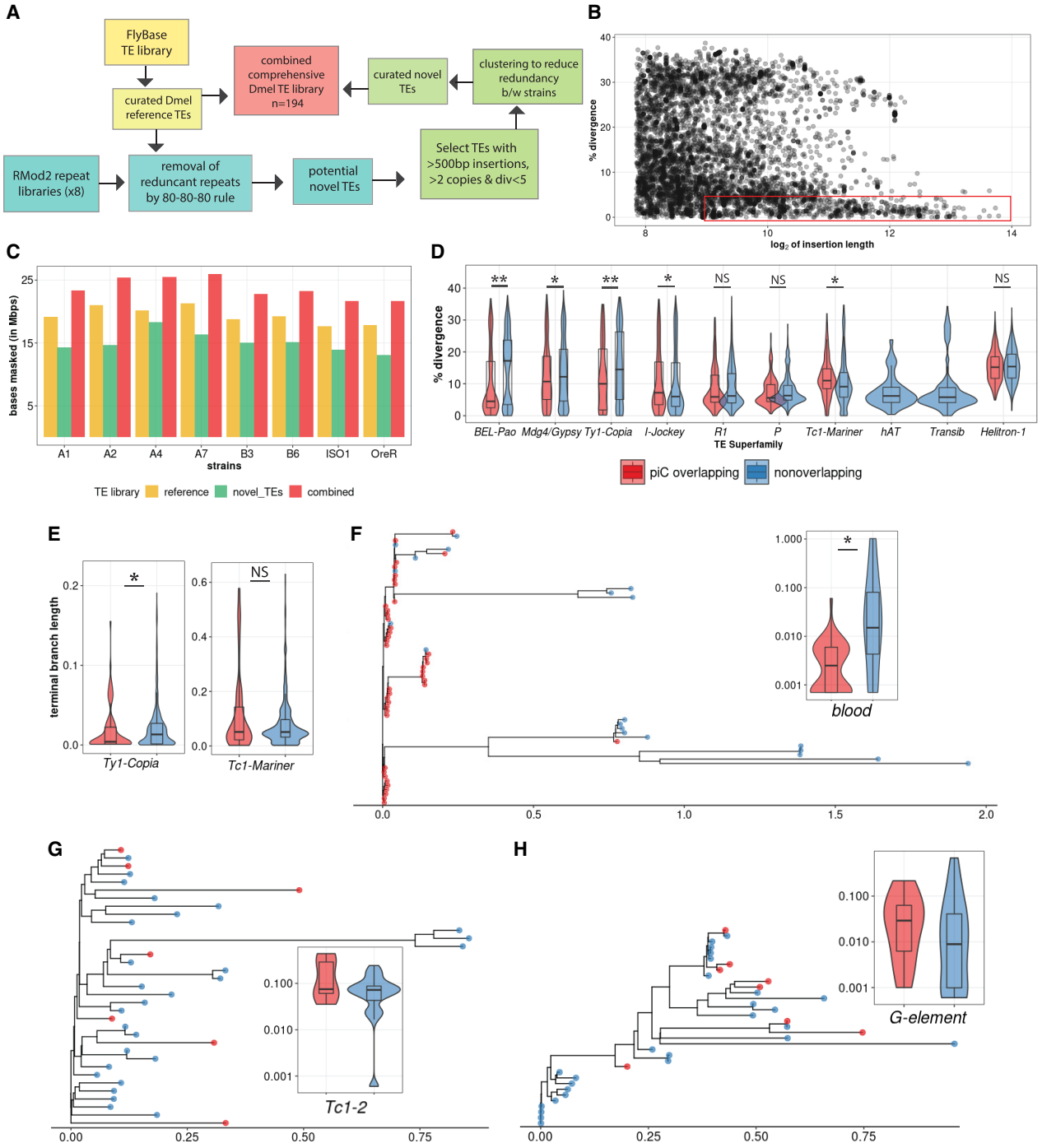

Figure 4. De novo TE annotation uncovers ~3 Mb of hidden TEs and reveals strong associations of young LTR TEs with piCs more than any other TE subclasses. (*A*) TE annotation pipeline using RepeatModeler2 and RepeatMasker to create the comprehensive TE library. (*B*) Abundance of extremely similar and long TE insertions from RepeatMasker output of strain B6 using novel TE consensus library. (*C*) Differences in million base-pairs (Mbp) masked in RepeatMasker results using novel-only, reference-only, and combined TE library. (*D*) Divergence estimates for all defragmented iso-1 insertions (>200 bp) from RepeatMasker output. Insertions with >1-bp overlap with master-list iso-1 piCs are considered piC overlapping. Difference between groups is tested by Wilcoxon rank-sum test. (*E*) Terminal branch length for all iso-1 insertions from *Ty1*/*Copia* and *Tc1*/*Mariner* superfamilies from maximum likelihood trees. (*F–H*) Maximum likelihood trees constructed from all defragmented insertions for *blood*, *Tc1-2*, and *G-element* families, and the *inset* shows terminal branch length quantification. Difference between groups is tested by Wilcoxon rank-sum test: (*) $P<0.05$, (**) $P<0.005$, (***) $P<0.0005$.

genome is also derived from the iso-1 strain. Presumably, the 75 nonreference insertions for iso-1 reflect the use of different isolates for the reference genome assembly and for the long-read data. Further clustering and parsing of all nonreference insertions across the eight strains resulted in a list of 3545 unique TE insertions of at least 200 bp in length (Supplemental Table S8). These insertions

belong to 165 of the 184 different families in our TE library. Ninety-four of these 165 TE families were classified as "active" because each of these included at least 10 nonreference insertions shared by no more than two strains, whereas the other 98 families were classified as "inactive" (Supplemental Table S9).

We used this compendium of insertions to test whether active TEs are significantly enriched within piCs using a binomial test to compare the observed overlaps with piCs to the average overlaps expected from 1000 random reshufflings of TE insertions (see Supplemental Methods; Kapusta et al. 2013). This analysis revealed that seven active TE families are significantly enriched in piCs, whereas one active and 10 inactive families are significantly depleted in piCs (Fig. 5A). All the inactive TE families that are significantly depleted in piCs belong to either DNA or LINE subclasses, with the exception that one active family depleted in piCs was *17.6*, a *Ty1/copia* superfamily member (Inouye et al. 1986) with 79 nonreference insertions. These results are consistent with the prediction of the "trap" model that piCs are enriched for active TE families but also composed of inactive families.

Next, we sought to distinguish family-level enrichment of TE insertions within rare, recently arisen piCs (smaller piCs [<10 kb] shared by no more than three strains) and within common larger "trap"-like piCs (>10 kb, shared by at least five strains). To increase the statistical power for this analysis, we used all TE sequences annotated by RepeatMasker in each genome, instead of only nonreference insertions. For iso-1, we find that only four TE families are significantly enriched within young piCs. All four are LTR retrotransposon families of the *Ty3/Mdg4* superfamily (*blood, Dm-412, flea, Stalker-2*), and all except *flea* belong to the *Mdg1* lineage (Fig. 5C,D; Costas et al. 2001; Bertocchi et al. 2020). All four families are also classified as active in this study as well as previous studies that examined TE insertion frequency among *D. melanogaster* populations (Kelleher and Barbash 2013; Kofler et al. 2015). In contrast, we find that numerous active and inactive families from all the TE subclasses are significantly enriched in large and common "trap-like" piCs, largely representative of the overall TE landscape of *D. melanogaster* (Fig. 5C,D). In the A1 genome, 20 TE families are significantly enriched in rare piCs. Again, these are predominantly LTR retrotransposons (14 families), but four DNA transposon families and two LINE families are also significantly enriched. (Fig. 5E,F). *1360*, a DNA transposon, is one of the four enriched DNA transposon families in A1 rare piCs but not in iso-1. It is possible these differences arise from changes in TE activity, the number of recent insertions, or strain background impacts TE's readily conversion into active piCs. *blood* insertions are neither enriched nor depleted in common piCs of both strains. Taken together, these analyses yield a contrasting portrait of TE composition in the two major types of piCs.

TE composition of piCs captures distinct steps in piC evolution

To further illuminate the evolution within piCs, we analyzed how the overall age and distribution of TEs of piCs change as they become more frequent in the population and presumably older. We plotted the mean percentage of divergence of individual TE insertions to their consensus sequences (a measure of TE age) across piCs and their flanking non-piC regions for each piC frequency class (strain-specific or shared by two to eight strains). We find that the divergence of TE insertions in rare piCs (shared by three or fewer strains) is markedly lower (3.5%–5%) than in their flanking regions (10%–15%) (Fig. 6A). In addition, the divergence of TE copies within piCs increases gradually with the frequency of the piCs to the extent that for the most common piCs (shared by seven and eight strains) the mean percentage of TE divergence is only slightly lower than in their flanking genomic regions (Fig. 6A). This apparent increase in average age of TE insertions as piCs become more common across strains provides weight to the inference that more common piCs represent evolutionarily older clusters relative to those that are strain specific or rare. This pattern also suggests that piCs are born from singleton TE insertions and grow by gradual accretion of TEs over time.

To further test this idea, we examined how TE coverage within piCs and surrounding regions changes as piCs become more common and presumably older (Fig. 6B). First, we observe that piCs generally show significantly higher TE coverage than their flanking genomic regions. Second, we find that strain-specific piCs, which presumably represent the youngest piCs, show high mean TE coverage in the middle at >60% (on average, 60 out every 100 bp is composed of TE sequence), which drops at the edges of piC coordinates to <20% (Fig. 6B). In contrast, more common piC groups show consistently higher TE coverage across their entire length. This pattern is consistent with a birth-and-growth process in which a piC emerges from individual TE insertion, but piRNA production spreads to flanking TE insertions as they insert near or within the piC.

Discussion

To study piC evolution at a fine-scale resolution in *D. melanogaster*, we used population genomic methods to characterize piC variation across eight inbred strains. A crucial asset was the availability of high-quality genome assemblies for these strains (Chakraborty et al. 2019). This enabled us to produce de novo annotation of piCs for each strain from mapping inferred piRNAs from ovarian small RNA libraries we constructed and sequenced for two biological replicates sampled six months apart. Our piC annotations for the two replicates showed high reproducibility with >88% of piCs annotated in one replicate found in the second replicate (Fig. 1A; Supplemental Fig. S3). Also, to understand variation in sequence composition and age of piCs, it was necessary to produce libraries of TE consensus sequences representative of the eight strains analyzed here. By performing de novo discovery and reannotation of TE families for each genome, we identified 47 novel TE families (Fig. 4). Although further investigation is required to examine their evolutionary origins and relationship to known TE families, it appears that many of the novel TE families we annotated were highly diverged from known families and often "hidden" in highly repetitive regions that would likely be poorly assembled in short-read genome assemblies. These results stress the benefits of high-quality genome assemblies and the necessity to perform de novo TE discovery when new strains or geographical isolates are considered. This is true even for model species like *D. melanogaster*, for which TEs have been extensively cataloged, because previous TE identifications were mostly based on a single reference genome. Robust annotation of piCs and TEs allowed us to compare the activity and TE composition of piCs across strains and derive general principles about piC origin and evolution.

Our findings recapitulate aspects of the "trap" model of piC evolution. First, we find enrichment of insertions of active TE families and depletion of inactive families in piC (Fig. 5B). In addition, we also find that "trap"-like large piCs are enriched for diverse TE families, which are relatively younger in piC than in flanking non-piC regions (Figs. 5D, 6A). This is consistent with the idea that "trap" piCs represent an archive of past TE activity, which biases

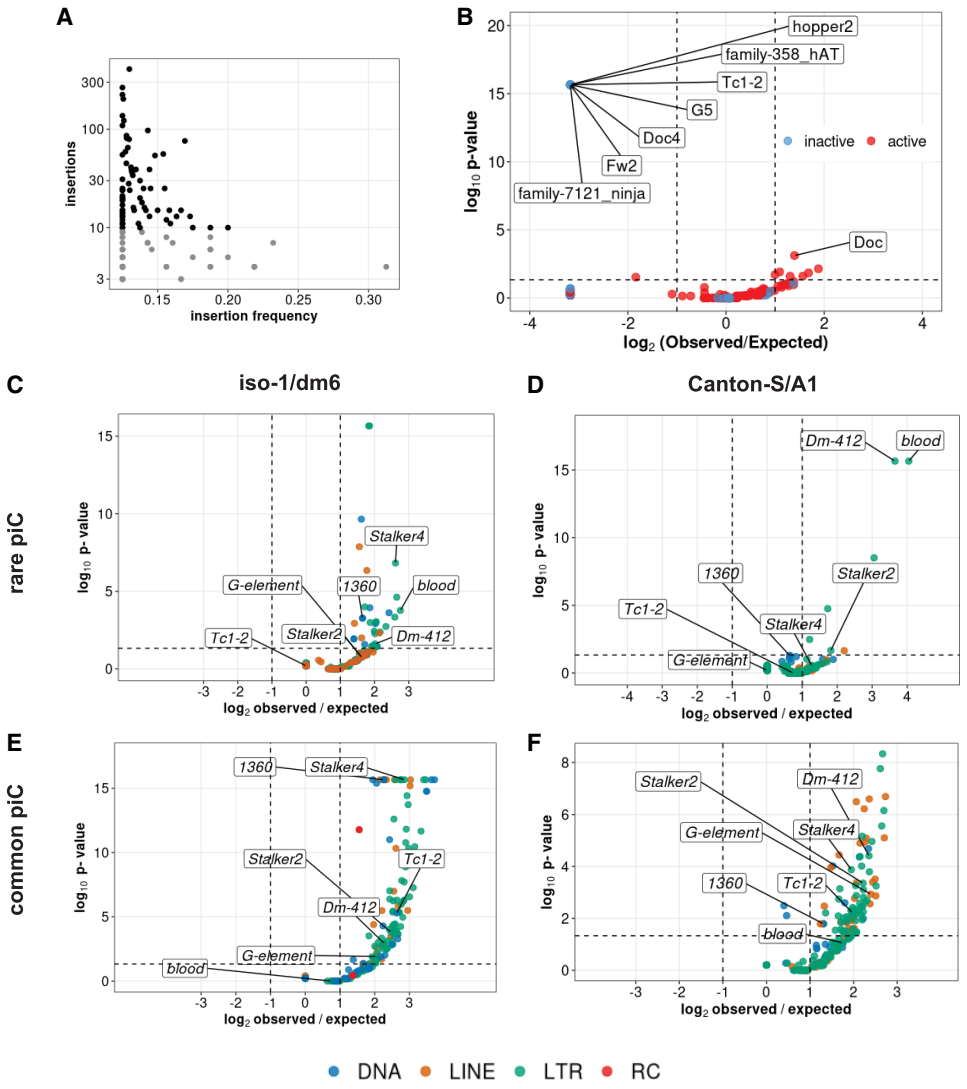

Figure 5. Insertions of only a few active LTR families associates with rare piCs. (*A*) Scatter plot of nonreference TE insertion counts and mean population frequencies of 176 TE families. Black dots highlight selected TEs classified as "active," whereas gray dots are deemed "inactive." (*B*) Enrichment analyses of nonreference insertions of TE families in master-list piCs using random shuffling of nonreference TE insertions. The *y*-axis *P*-values are from binomial tests conducted to compare observed counts to expected average overlaps of nonreference TE insertions to piCs for each family. (*C*–*F*) Enrichment analyses of TE families in master-list rare and common piCs of the A1/Can-S strain using random shuffling. All TE insertions (>200 bp) of strain iso-1 genome are randomly shuffled with constant piCs coordinates grouped into rare (shared by four strains or fewer) and common piCs (shared by five strains or more). *P*-values on *y*-axes are from binomial tests conducted to compare observed counts to expected average overlaps of TE insertions to piCs for each family. Statistically significant families are labeled.

piRNA production toward recently active TE families. Consistent with previously reported variation in piC activity (Rozhkov et al. 2010; Ellison and Cao 2020; Zhang et al. 2020), we find extensive intraspecific variation in piC activity, even for large "trap"-like piCs. For example, *42AB* and *38C* clusters show a significant loss in piRNA production in one or multiple strains (Fig. 2; Supplemental Fig. S6). What could cause the loss in piRNA production from large heterochromatic piCs? Our current lack of understanding of the *cis*-regulatory requirements for piC activity makes it difficult to determine whether changes in piRNA production are caused by genetic or epigenetic changes in the piCs. Consistent with previous studies (Ellison and Cao 2020; Wierzbicki et al. 2023), we observe considerable structural variation in large peri-centromeric piCs (Supplemental Fig. S7). It is possible that such structural changes result in changes in piRNA production, but further studies are needed to elucidate the mechanisms by which large and seemingly stable piCs lose their activity.

What can TE composition of piCs tell us about the coevolution of TEs and piCs? As previously reported (Kofler 2020; Wierzbicki et al. 2023), we find that diverse TE families (from all subclasses) are enriched in large, common piCs (Fig. 5). This enrichment may be explained by selection against new TE insertions in gene-rich euchromatic regions, which leads to accumulation of TEs in heterochromatic regions, where purifying selection is also weak (Charlesworth and Langley 1989; Bartolomé et al. 2002; Blumenstiel et al. 2002; Dolgin and Charlesworth 2006; Schrider et al. 2013). However, our genome-wide analysis revealed that SV enrichment in common piCs cannot be completely explained

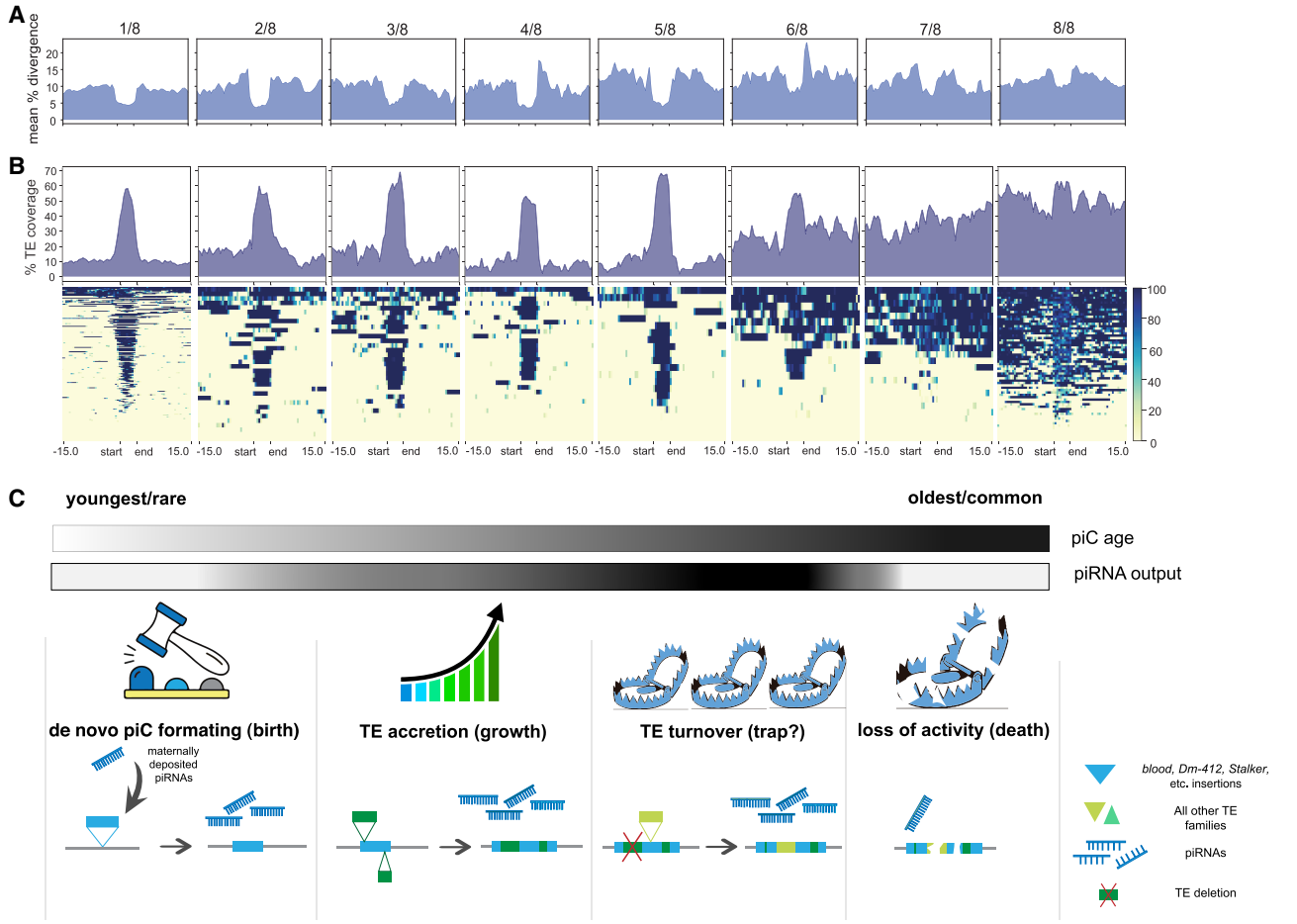

Figure 6. Age and distribution of TEs provide a portrait of intraspecific piC evolution. (*A*) TE age (by percentage of divergence) is plotted for piCs of increasing population frequency from *left* to *right*. Summary profile plot of mean percentage of TE insertion divergence in 500-bp windows in scaled piC regions and flanking non-piC regions of the reference genome. Each group represents piCs that are shared by 1/8 to 8/8 strains. (*B*) TE content of piCs of different stages are plotted. Summary profile plot and heatmap of mean TE coverage in 500-bp windows in scaled piC regions and flanking ±15-kb non-piC regions of the reference genome. Scaled piC regions are marked by start and end labels at the *bottom* of the heatmap. (*C*) New unified model of piC evolution: "Birth and death" is shown in four steps from *left* to *right*.

by their overlap with heterochromatin (Supplemental Fig. S8). Thus, the enrichment of diverse elements within large common piCs may be driven in part by the lesser deleterious effects of their accumulation in these regions and/or positive selection for their piRNA production.

Are particular TEs prone to give rise to piCs? To answer this question, we tested for enrichment of individual TE families within rare piCs, which likely represent young clusters. Only a small set of retrotransposon families are enriched within such clusters, and most belong to the *mdg1* subclade of LTR retrotransposons such as *blood* and *Dm-412* (Fig. 5). Why would these elements be prone to seed new piCs? There are two possibilities, which are not mutually exclusive. One possibility is that these TE families are readily licensed as piCs from maternally inherited piRNAs, thereby giving rise to the emergence of dispersed piC loci following recent transposition events (Le Thomas et al. 2014; Shpiz et al. 2014). Alternatively, it may be linked to their propensity to produce double-stranded RNA (dsRNA) and endogenous siRNAs, as endo-siRNAs production has been associated with the formation of piCs from transgenes (Olovnikov et al. 2013; Akulenko et al. 2018). Many LTR (including *blood*) and non-LTR retrotransposons are known to possess bidirectional promoters that can result in the formation of dsRNAs that stimulate the production of siRNAs (Watanabe et al. 2008; Russo et al. 2016; Hung and Slotkin 2021). It is possible that the inherent capacity of some retrotransposons to produce dsRNAs nucleate the formation of piCs. This idea has received support in a recent study showing that endogenous siRNA production precedes piC formation, and maternal inheritance of these siRNAs is required for licensing of piCs (Luo et al. 2023). However, we cannot rule out that these rare TE-specific piCs arose exclusively from maternally inherited piRNAs as piRNAs from such piCs could not be differentiated from the piRNAs produced from insertions in large common piCs (Supplemental Fig. S12).

In summary, a picture is emerging in which the rapid evolution of piCs among *D. melanogaster* strains is driven by the activity of TE families themselves. We propose a "birth-and-death" model of piC evolution, which combines components of the

conventional "trap" model and features of piCs found in this study (Bergman et al. 2006; Khurana et al. 2011; Shpiz et al. 2014; Gebert et al. 2021). In this model (Fig. 6C), we posit that piCs form frequently throughout the genome, often from recent TE insertions, with certain LTR retrotransposon families making a stronger contribution to seeding new piCs. Newly emerged piCs may increase in frequency and size owing to natural selection or drift, depending on factors such as their propensity to trigger genomic autoimmunity (Blumenstiel et al. 2016; Huang et al. 2022), ectopic recombination (Petrov et al. 2003; Sentmanat and Elgin 2012), and the establishment of a chromatin environment conducive for piRNA production (Le Thomas et al. 2014). Over time, these stabilized clusters may grow by "trapping" additional TE insertions, which will eventually result in large heterochromatic clusters such as *42AB*. Because of the host's limited capacity to maintain such piCs without incurring a fitness cost, those clusters that lack piRNAs targeting active TEs may gradually lose activity or become dispensable (Gebert et al. 2021). We can only speculate about the mechanisms that lead to the loss of activity of large piCs, but the rapid turnover of piC loci provide some possibilities. Because of re-establishment of female piC loci every generation from maternally deposited piRNAs, it is possible that a large piC may not get licensed if it only contains old insertions from inactive TE families, which are poorly represented in maternally deposited piRNAs. Because the maternally deposited piRNA pool is expected to diverge from generation to generation to incorporate new TEs, it may not contain enough homologous piRNAs for licensing of older piCs that lack recently active TEs. This seems consistent with high piRNA abundance of *42AB* in five of the eight strains, low in two of the eight strains, and extremely low in one strain. The most parsimonious conclusion from such continuous variation can be drawn that *42AB* activity is on the decline in *D. melanogaster* populations. Further studies are warranted to test this "birth-and-death" model. Our study provides a first in-depth view of piC evolution in *D. melanogaster* that is likely to stimulate other comparative studies of piRNA evolution.

Methods

Fly stocks

DSPR founder stocks of A1 (b1_paired), A2 (b3841_paired), A4 (b1_3852), A7 (t7_paired), B3 (b3864_paired), B6 (t1_paired), and Oregon-R were a gift from Anthony Long (UCI). All stocks were maintained on standard cornmeal medium at 22°C under a 12-h day–night cycle.

Small RNA library construction and sequencing

Small RNA libraries were constructed by size fractionation on urea-polyacrylamide gel electrophoresis as described by Ma et al. (2021), and additional details are provided in the Supplemental Methods. All libraries were quantified using Qubit 3.0, pooled into replicate-1 and replicate-2 groups, and analyzed on a Agilent Bioanalyzer. Single-end 75-bp Illumina sequencing was performed for all libraries on NextSeq 500 at the Cornell Biotechnology Resource Center.

piC annotation

Active piC annotation was conducted independently for each replicate of each strain using a custom pipeline adapted from previously described methods (Rosenkranz and Zischler 2012; Mohn et al. 2014). Detailed annotation steps of each pipeline are provided in the Supplemental Methods and outlined in Supplemental Figure S2.

Structural variation detection and filtering

Raw long reads for the seven DSPR strains, the iso-1 reference strain, *D. simulans* (wxd1), and *D. sechellia* (sech25) were mapped to the *D. melanogaster* iso-1 release 6 (GCA_000001215.4) without the Y Chromosome with minimap2.1 map-pb --N3, and resulting SAM file was converted to BAM and sorted (Li et al. 2009; Li 2018). Three SV callers—Sniffles-2.0 (Smolka et al. 2024), cuteSV-1.0.13 (Jiang et al. 2020), and svim-2.0 (Heller and Vingron 2019)—were used for SV detection. Filtering, genotyping, and collapsing of SV calls are detailed in the Supplemental Methods. Filtered and polarized SV calls are reported in Supplemental Table S5.

De novo TE annotation

To create a comprehensive and accurate representative TE library representing the TE insertions contained in the eight strains, de novo TE annotation was conducted using several computational tools. A summary of all major steps is presented in a flow chart in Figure 4A. Briefly, canonical FlyBase TE consensus sequences were filtered to include only TEs that best represent the TE insertion landscape of each strain using RepeatMasker-4.1.0 results (Smit 1999; Larkin et al. 2021). FlyBase TE families with at least three copies of >200 bp and <1% divergence were retained. This resulted in a reference library for the eight strains with 129 TE families. Next, RepeatModeler2 was run on the seven DSPR genomes and reference iso-1 strain followed by removal of non-TE repeats like tRNA, satellites, rRNA, etc., as well as TE subfamilies using bash scripts (Flynn et al. 2020). Putative novel TE family consensus was identified (see Supplemental Methods), and each genome was remasked with combined TE library (novel and reference TEs).

Nonreference TE insertion detection from long reads

Raw long reads used in SV detection were also used for unique TE insertion analyses. Reads (>1 kb) were mapped to iso-1 reference genome (without Y-linked contigs and all contigs <20 kb) using minimap2.1 default parameters, and the resulting SAM file was converted to BAM file, sorted, and indexed using SAMtools (Li et al. 2009; Li 2018). TLDR, a de novo TE detection program (Ewing et al. 2020), was run for each strain using the comprehensive *D. melanogaster* TE library curated in this study. High-confidence TLDR insertion calls for all eight strains are reported in Supplemental Table S7.

Data access

All raw and processed sequencing data generated in this study have been submitted to the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA1013381. Relevant source data and methods for figures are available as Supplemental Material. Additional raw data files, TE consensus sequences, and scripts are available as Supplemental Code and at GitHub (https://github.com/kerogens101/Dmel_piC).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

This work was supported by National Institutes of Health grants R01-GM119125 to A.G.C. and R35-GM122550 to C.F. S.P.S. is supported by the Distinguished Scholar Award from the Cornell Center for Vertebrate Genomics. We thank Jullien Flynn for her assistance in RepeatModeler2 runs and consensus filtering and Michelle Stitzer for recommendations on structural variant analysis. We also thank Dan Barbash, Justin Blumenstiel, and Andrew Grimson for helpful discussions on the analyses and members of the Clark and Feschotte laboratories for valuable feedback on the manuscript.

Author contributions: S.P.S. performed all experiments. S.P.S. performed all computational analysis and created all figures with input from A.G.C. and C.F. S.P.S., A.G.C., and C.F. wrote the manuscript. Initial project conception was by S.P.S. and A.G.C.

References

Akkouche A, Mugat B, Barckmann B, Varela-Chavez C, Li B, Raffel R, Pélisson A, Chambeyron S. 2017. Piwi is required during *Drosophila* embryogenesis to license dual-strand piRNA clusters for transposon repression in adult ovaries. *Mol Cell* **66:** 411–419.e4. doi:10.1016/j.molcel.2017.03.017

Akulenko N, Ryazansky S, Morgunova V, Komarov PA, Olovnikov I, Vaury C, Jensen S, Kalmykova A. 2018. Transcriptional and chromatin changes accompanying de novo formation of transgenic piRNA clusters. *RNA* **24:** 574–584. doi:10.1261/rna.062851.117

Aravin AA, Hannon GJ, Brennecke J. 2007. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318:** 761–764. doi:10.1126/science.1146484

Asif-Laidin A, Delmarre V, Laurentie J, Miller WJ, Ronsseray S, Teysset L. 2017. Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in *Drosophila*. *DNA Res* **24:** 459–472. doi:10.1093/dnares/dsx017

Bartolomé C, Maside X, Charlesworth B. 2002. On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*. *Mol Biol Evol* **19:** 926–937. doi:10.1093/oxfordjournals.molbev.a004150

Baumgartner L, Handler D, Platzer SW, Yu C, Duchek P, Brennecke J. 2022. The *Drosophila* ZAD zinc finger protein Kipferl guides Rhino to piRNA clusters. *eLife* **11:** e80067. doi:10.7554/eLife.80067

Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and duplication generate networks of transposable element sequences in the *Drosophila melanogaster* genome. *Genome Biol* **7:** R112. doi:10.1186/gb-2006-7-11-r112

Bertocchi NA, Torres FP, Deprá M, da Silva Valente VL. 2020. Evolutionary study of the *412/mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons in Diptera. bioRxiv doi:10.1101/2020.09.24.311225

Bingham PM, Chapman CH. 1986. Evidence that white-blood is a novel type of temperature-sensitive mutation resulting from temperature-dependent effects of a transposon insertion on formation of white transcripts. *EMBO J* **5:** 3343–3351. doi:10.1002/j.1460-2075.1986.tb04649.x

Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the P element, a P-strain-specific transposon family. *Cell* **29:** 995–1004. doi:10.1016/0092-8674(82)90463-9

Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting chromatin domains. *Mol Biol Evol* **19:** 2211–2225. doi:10.1093/oxfordjournals.molbev.a004045

Blumenstiel JP, Erwin AA, Hemmer LW. 2016. What drives positive selection in the *Drosophila* piRNA machinery? The genomic autoimmunity hypothesis. *Yale J Biol Med* **89:** 499–512.

Bowen NJ, McDonald JF. 2001. *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res* **11:** 1527–1540. doi:10.1101/gr.164201

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128:** 1089–1103. doi:10.1016/j.cell.2007.01.043

Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322:** 1387–1392. doi:10.1126/science.1165171

Bucheton A, Paro R, Sang HM, Pelisson A, Finnegan DJ. 1984. The molecular basis of I-R hybrid dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. *Cell* **38:** 153–163. doi:10.1016/0092-8674(84)90536-1

Carr M, Bensasson D, Bergman CM. 2012. Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *PLoS One* **7:** e50978. doi:10.1371/journal.pone.0050978

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10:** 4872. doi:10.1038/s41467-019-12884-1

Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, Montooth KL, Meiklejohn CD, Larracuente AM, Emerson JJ. 2021. Evolution of genome structure in the *Drosophila simulans* species complex. *Genome Res* **31:** 380–396. doi:10.1101/gr.263442.120

Chang N-C, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. 2022. Zebrafish transposable elements show extensive diversification in age, genomic distribution, and developmental expression. *Genome Res* **32:** 1408–1423. doi:10.1101/gr.275655.121

Charlesworth B, Langley CH. 1989. The population genetics of *Drosophila* transposable elements. *Annu Rev Genet* **23:** 251–287. doi:10.1146/annurev.ge.23.120189.001343

Chen P, Kotov AA, Godneeva BK, Bazylev SS, Olenina LV, Aravin AA. 2021. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in *D. melanogaster* male germline. *Genes Dev* **35:** 914–635. doi:10.1101/gad.345041.120

Chirn G-W, Rahman R, Sytnikova YA, Matts JA, Zeng M, Gerlach D, Yu M, Berger B, Naramura M, Kile BT, et al. 2015. Conserved piRNA expression from a distinct set of piRNA cluster loci in eutherian mammals. *PLoS Genet* **11:** e1005652. doi:10.1371/journal.pgen.1005652

Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation, and cooption. *Genes Dev* **33:** 1098–1116. doi:10.1101/gad.327312.119

Costas J, Valadé E, Naveira H. 2001. Structural features of the *mdg1* lineage of the *Ty3/gypsy* group of LTR retrotransposons inferred from the phylogenetic analyses of its open reading frames. *J Mol Evol* **53:** 165–171. doi:10.1007/s002390010206

Di Nocera PP, Graziani F, Lavorgna G. 1986. Genomic and structural organization of *Drosophila melanogaster* G elements. *Nucleic Acids Res* **14:** 675–691. doi:10.1093/nar/14.2.675

Dolgin ES, Charlesworth B. 2006. The fate of transposable elements in asexual populations. *Genetics* **174:** 817–827. doi:10.1534/genetics.106.060434

Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in *Drosophila melanogaster*. *Proc Natl Acad Sci USA* **104:** 19920–19925. doi:10.1073/pnas.0709888104

Ellison CE, Cao W. 2020. Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of *Drosophila melanogaster*. *Nucleic Acids Res* **48:** 290–303. doi:10.1093/nar/gkz1080

Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW, Faulkner GJ. 2020. Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol Cell* **80:** 915–928.e5. doi:10.1016/j.molcel.2020.10.024

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117:** 9451–9457. doi:10.1073/pnas.1921046117

Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. 2021. Large *Drosophila* germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol Cell* **81:** 3965–3978.e5. doi:10.1016/j.molcel.2021.07.011

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. doi:10.1101/gr.275478.121

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455:** 1193–1197. doi:10.1038/nature07415

Han S, Dias GB, Basting PJ, Viswanatha R, Perrimon N, Bergman CM. 2022. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. *Nucleic Acids Res* **50:** e124. doi:10.1093/nar/gkac794

Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads. *Bioinformatics* **35:** 2907–2915. doi:10.1093/bioinformatics/btz041

Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov D V, Blaser H, Raz E, Moens CB, et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* **129:** 69–82. doi:10.1016/j.cell.2007.03.026

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al. 2014. Natural variation in genome

architecture among 205 *Drosophila melanogaster* genetic reference panel lines. *Genome Res* **24:** 1193–1208. doi:10.1101/gr.171546.113

Huang Y, Shukla H, Lee YCG. 2022. Species-specific chromatin landscape determines how transposable elements shape genome evolution. *eLife* **11:** e81567. doi:10.7554/eLife.81567

Hung YH, Slotkin RK. 2021. The initiation of RNA interference (RNAi) in plants. *Curr Opin Plant Biol* **61:** 102014. doi:10.1016/j.pbi.2021.102014

Inouye S, Hattori K, Yuki S, Saigo K. 1986. Structural variations in the *Drosophila* retrotransposon. *Nucleic Acids Res* **14:** 4765–4778. doi:10.1093/nar/14.12.4765

Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol* **21:** 189. doi:10.1186/s13059-020-02107-y

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, Frise E, Wheeler DA, Lewis SE, Rubin GM, et al. 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol* **3:** research0084.1. doi:10.1186/gb-2002-3-12-research0084

Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay LA, Bourque G, Yandell M, Feschotte C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* **9:** e1003470. doi:10.1371/journal.pgen.1003470

Kelleher ES, Barbash DA. 2013. Analysis of piRNA-mediated silencing of active TEs in *Drosophila melanogaster* suggests limits on the evolution of host genome defense. *Mol Biol Evol* **30:** 1816–1829. doi:10.1093/molbev/mst081

Kelleher ES, Azevedo RBR, Zheng Y. 2018. The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biol Evol* **10:** 3038–3057. doi:10.1093/gbe/evy218

Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE. 2011. Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147:** 1551–1563. doi:10.1016/j.cell.2011.11.042

Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS, Nowosielska A, et al. 2009. The *Drosophila* HP1 homolog rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* **138:** 1137–1149. doi:10.1016/j.cell.2009.07.014

Kofler R. 2019. Dynamics of transposable element invasions with piRNA clusters. *Mol Biol Evol* **36:** 1457–1472. doi:10.1093/molbev/msz079

Kofler R. 2020. piRNA clusters need a minimum size to control transposable element invasions. *Genome Biol Evol* **12:** 736–749. doi:10.1093/gbe/evaa064

Kofler R, Nolte V, Schlötterer C. 2015. Tempo and mode of transposable element activity in *Drosophila*. *PLoS Genet* **11:** e1005406. doi:10.1371/journal.pgen.1005406

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. doi:10.1093/nar/gkaa1026

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* **313:** 363–367. doi:10.1126/science.1130164

Lawlor MA, Cao W, Ellison CE. 2021. A transposon expression burst accompanies the activation of Y-chromosome fertility genes during *Drosophila* spermatogenesis. *Nat Commun* **12:** 6854. doi:10.1038/s41467-021-27136-4

Le Thomas A, Stuwe E, Li S, Du J, Marinov G, Rozhkov N, Chen Y-CA, Luo Y, Sachidanandam R, Toth KF, et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes Dev* **28:** 1667–1680. doi:10.1101/gad.245514.114

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34:** 3094–3100. doi:10.1093/bioinformatics/bty191

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079. doi:10.1093/bioinformatics/btp352

Luo Y, He P, Kanrar N, Fejes Toth K, Aravin AA. 2023. Maternally inherited siRNAs initiate piRNA cluster formation. *Mol Cell* **83:** 3835–3851.e7. doi:10.1016/j.molcel.2023.09.033

Ma Q, Srivastav SP, Gamez S, Dayama G, Feitosa-Suntheimer F, Patterson EI, Johnson RM, Matson EM, Gold AS, Brackney DE, et al. 2021. A mosquito small RNA genomics resource reveals dynamic evolution and host responses to viruses and transposons. *Genome Res* **31:** 512–528. doi:10.1101/gr.265157.120

Miller DE, Dorador AP, Van Vaerenberghe K, Li A, Grantham EK, Cerbin S, Cummings C, Barragan M, Egidy RR, Scott AR, et al. 2023. Off-target piRNA gene silencing in *Drosophila melanogaster* rescued by a transposable element insertion. *PLoS Genet* **19:** e1010598. doi:10.1371/journal.pgen.1010598

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. doi:10.1016/j.cell.2014.04.031

Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. *Genetics* **129:** 1085–1098. doi:10.1093/genetics/129.4.1085

Olovnikov I, Ryazansky S, Shpiz S, Lavrov S, Abramov Y, Vaury C, Jensen S, Kalmykova A. 2013. De novo piRNA cluster formation in the *Drosophila* germ line triggered by transgenes containing a transcribed transposon fragment. *Nucleic Acids Res* **41:** 5757–5768. doi:10.1093/nar/gkt310

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* **20:** 89–108. doi:10.1038/s41576-018-0073-3

Palmer WH, Hadfield JD, Obbard DJ. 2018. RNA-interference pathways display high rates of adaptive protein evolution in multiple invertebrates. *Genetics* **208:** 1585–1599. doi:10.1534/genetics.117.300567

Pasquesi GIM, Perry BW, Vandewege MW, Ruggiero RP, Schield DR, Castoe TA. 2020. Vertebrate lineages exhibit diverse patterns of transposable element regulation and expression across tissues. *Genome Biol Evol* **12:** 506–521. doi:10.1093/gbe/evaa068

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. *Mol Biol Evol* **20:** 880–892. doi:10.1093/molbev/msg102

Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat Commun* **13:** 1948. doi:10.1038/s41467-022-29518-8

Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, Tolstorukov MY, Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, et al. 2011. Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res* **21:** 147–163. doi:10.1101/gr.110098.110

Rosenkranz D, Zischler H. 2012. proTRAC: a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13:** 5. doi:10.1186/1471-2105-13-5

Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR, Hannon GJ, Evgen'ev MB. 2010. Small RNA-based silencing strategies for transposons in the process of invading *Drosophila* species. *RNA* **16:** 1634–1645. doi:10.1261/rna.2217810

Russo J, Harrington AW, Steiniger M. 2016. Antisense transcription of retrotransposons in *Drosophila*: an origin of endogenous small interfering RNA precursors. *Genetics* **202:** 107–121. doi:10.1534/genetics.115.177196

Said I, McGurk MP, Clark AG, Barbash DA. 2022. Patterns of piRNA regulation in *Drosophila* revealed through transposable element clade inference. *Mol Biol Evol* **39:** msab336. doi:10.1093/molbev/msab336

Saint-Leandre B, Capy P, Hua-Van A, Filée J. 2020. piRNA and transposon dynamics in *Drosophila*: a female story. *Genome Biol Evol* **12:** 931–947. doi:10.1093/gbe/evaa094

Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of spontaneous mutational events in *Drosophila melanogaster*. *Genetics* **194:** 937–954. doi:10.1534/genetics.113.151670

Sentmanat MF, Elgin SCR. 2012. Ectopic assembly of heterochromatin in *Drosophila melanogaster* triggered by transposable elements. *Proc Natl Acad Sci USA* **109:** 14104–14109. doi:10.1073/pnas.1207036109

Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. 2014. Euchromatic transposon insertions trigger production of novel Pi- and endo-siRNAs at the target sites in the *Drosophila* germline. *PLoS Genet* **10:** e1004138. doi:10.1371/journal.pgen.1004138

Simkin A, Wong A, Poh Y-P, Theurkauf WE, Jensen JD. 2013. Recurrent and recent selective sweeps in the piRNA pathway. *Evolution* **67:** 1081–1090. doi:10.1111/evo.12011

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657–663. doi:10.1016/S0959-437X(99)00031-3

Smolka M, Paulin LF, Grochowski CM, Horner DW, Mahmoud M, Behera S, Kalef-Ezra E, Gandhi M, Hong K, Pehlivan D, et al. 2024. Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* doi:10.1038/s41587-023-02024-y

Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS. 2018. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage. *Adv Genet* **8:** 3143–3154. doi:10.1534/g3.118.200162

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* **18:** 292–308. doi:10.1038/nrg.2017.7

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453:** 539–543. doi:10.1038/nature06908

Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. *Annu Rev Genet* **54:** 539–561. doi:10.1146/annurev-genet-040620-022145

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* **8:** 973–982. doi:10.1038/nrg2165

Wierzbicki F, Kofler R, Signor S. 2023. Evolutionary dynamics of piRNA clusters in *Drosophila*. *Mol Ecol* **32:** 1306–1322. doi:10.1111/mec.16311

Yi M, Chen F, Luo M, Cheng Y, Zhao H, Cheng H, Zhou R. 2014. Rapid evolution of piRNA pathway in the teleost fish: implication for an adaptation to transposon diversity. *Genome Biol Evol* **6:** 1393–1407. doi:10.1093/gbe/evu105

Yu T, Koppetsch BS, Pagliarani S, Johnston S, Silverstein NJ, Luban J, Chappell K, Weng Z, Theurkauf WE. 2019. The piRNA response to retroviral invasion of the koala genome. *Cell* **179:** 632–643.e12. doi:10.1016/j.cell.2019.09.002

Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S. 2013. Distribution, evolution, and diversity of retrotransposons at the *flamenco* locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci USA* **110:** 19842–19847. doi:10.1073/pnas.1313677110

Zhang S, Pointer B, Kelleher ES. 2020. Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Res* **30:** 566–575. doi:10.1101/gr.251546.119

Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO. 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res* **23:** 568–579. doi:10.1101/gr.142646.112

Received May 8, 2023; accepted in revised form May 7, 2024.