Received: 30 August 2021 | Revised: 24 November 2021 | Accepted: 1 December 2021

DOI: 10.1111/mec.16311

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

Evolutionary dynamics of piRNA clusters in *Drosophila*

Filip Wierzbicki[1,2] | Robert Kofler[1] | Sarah Signor[3]

[1]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria

[2]Vienna Graduate School of Population Genetics, Vienna, Austria

[3]Biological Sciences, North Dakota State University, Fargo, North Dakota, USA

Correspondence
Robert Kofler, Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria.
Email: rokofler@gmail.com

Sarah Signor, Vienna Graduate School of Population Genetics, Vienna, Austria.
Email: sarah.signor@ndsu.edu

Funding information
Austrian Science Fund, Grant/Award Number: P30036-B25; National Science Foundation, Grant/Award Number: NSF-EPSCoR-1826834 and NSF-EPSCoR-2032756; North Dakota EPSCoR STEM grants program

Abstract

Small RNAs produced from transposable element (TE)-rich sections of the genome, termed piRNA clusters, are a crucial component in the genomic defence against selfish DNA. In animals, it is thought the invasion of a TE is stopped when a copy of the TE inserts into a piRNA cluster, triggering the production of cognate small RNAs that silence the TE. Despite this importance for TE control, little is known about the evolutionary dynamics of piRNA clusters, mostly because these repeat-rich regions are difficult to assemble and compare. Here, we establish a framework for studying the evolution of piRNA clusters quantitatively. Previously introduced quality metrics and a newly developed software for multiple alignments of repeat annotations (Manna) allow us to estimate the level of polymorphism segregating in piRNA clusters and the divergence among homologous piRNA clusters. By studying 20 conserved piRNA clusters in multiple assemblies of four *Drosophila* species, we show that piRNA clusters are evolving rapidly. While 70%–80% of the clusters are conserved within species, the clusters share almost no similarity between species as closely related as *D. melanogaster* and *D. simulans*. Furthermore, abundant insertions and deletions are segregating within the *Drosophila* species. We show that the evolution of clusters is mainly driven by large insertions of recently active TEs and smaller deletions mostly in older TEs. The effect of these forces is so rapid that homologous clusters often do not contain insertions from the same TE families.

KEYWORDS
Drosophila, molecular evolution, piRNA clusters, transposable elements

1 | INTRODUCTION

Transposable elements (TEs) are short sequences of DNA that multiply within genomes (Mcclintock, 1956). TEs are widespread across the tree of life, often making up a significant portion of the genome (2.7%–25% in fruit flies, 45% in humans and 85% in maize (Piegu et al., 2006; Schnable et al., 2009; Lee & Langley, 2012). TEs also impose a severe mutational load on their hosts by producing insertions that disrupt functional sequences and mediate ectopic recombination (Lim, 1988, Levis et al., 1984, McGinnis et al., 1983). However, some TE insertions have also been associated with increases in fitness, for example due to changes in gene regulation, where they can act as enhancers, repressors or other regulators of complex gene expression patterns (Daborn et al., 2002, Gonzalez et al., 2008, Mateo et al., 2014, Casacuberta & Gonzalez, 2013). The distribution of fitness effects of TEs is not known, but the majority of insertions are thought to be deleterious (Yang & Nuzhdin, 2003, Dimitri et al., 2003, Lee & Langley, 2012, Adrion et al., 2017).

For a long time, TEs were thought to be solely counteracted at the population level (transposition/selection balance) (Charlesworth & Charlesworth, 1983, Barron et al., 2014). However, the discovery of a small RNA-based defence system revealed that they are also actively combated by the host (Brennecke et al., 2007, Lee & Langley, 2010, Blumenstiel, 2011). This host defence system relies upon PIWI-interacting RNAs (piRNAs) that bind to PIWI-clade proteins and suppress TE activity transcriptionally and post-transcriptionally (Brennecke et al., 2007, Gunawardane et al., 2007, Sienski et al., 2012, Le Thomas et al., 2013). For example, in *D. melanogaster*, post-transcriptional silencing of TEs is based on Aub and AGO3, which, guided by piRNAs, cleave TE transcripts in the cytoplasm (Kalmykova et al., 2005, Peters & Meister, 2007, Brennecke et al., 2007, Gunawardane et al., 2007). In the nucleus piRNAs guide PIWI to transcribed TEs which, aided by other proteins, transcriptionally silence TEs through the establishment of repressive chromatin marks (Sienski et al., 2012, Le Thomas et al., 2013, Darricarrere et al., 2013). These piRNAs are produced from discrete regions of the genome termed piRNA clusters, which largely consist of many TE fragments (Brennecke et al., 2007), although individual euchromatic TE insertions may also form piRNA-producing loci (Shpiz et al., 2014, Mohn et al., 2014). There is evidence that a single insertion of a TE into a piRNA cluster may be sufficient to initiate piRNA mediated silencing of the TE, although piRNA production may take several generations to peak (Marin et al., 2000, Josse et al., 2007, Zanni et al., 2013). Therefore, one hypothesis is that a newly invading TE proliferates in the host until a copy jumps into a piRNA cluster, which triggers the production of piRNAs that silence the TE (Bergman et al., 2006, Malone & Hannon, 2010, Goriaux et al., 2014, Ozata et al., 2019). Recent findings bring this hypothesis into question and suggest a possible alternative role for piRNA clusters. Gebert et al. (2021) found that individual piRNA clusters were dispensable for TE suppression and suggest that dispersed piRNA-producing TEs may largely maintain silencing of TEs (Gebert et al., 2021). These two hypotheses may represent two aspects of TE suppression, wherein piRNA clusters could be important to trigger the silencing of a TE but may be dispensable later on, where dispersed piRNA-producing loci maintain silencing of the TEs (Chen & Aravin, 2021).

Despite the central importance of piRNA clusters for the control of TEs, we know very little about how piRNA clusters evolve within and between species. For example, transposition into clusters would be advantageous to hosts if cluster insertions are indeed required for functional silencing of TEs. Then, a general expansion of piRNA clusters would be expected with the invasion of novel TEs. Such invasions may be quite frequent. For example, it is likely that four TE families invaded worldwide *D. melanogaster* populations within the last 100 years (Schwarz et al., 2021). Larger or more abundant piRNA clusters in turn will expand the functional target for transposition and may thus be favoured. In support of this hypothesis, it was suggested that piRNA clusters have largely been gained over the course of evolution (Chirn et al., 2015). However, these claims are difficult to evaluate as studying the evolution of piRNA clusters is challenging for several reasons. First, piRNA clusters are highly repetitive and very difficult to assemble; thus, high-quality ungapped assemblies of these repetitive regions are required (see, e.g., Wierzbicki et al., 2021). Second, it is challenging to unambiguously identify homologous clusters within and between species. Third, investigating the evolution of the composition of clusters requires reliable alignments of the highly repetitive piRNA clusters. Due to these challenges and the importance of these clusters for TE control, the evolutionary turnover of piRNA clusters is considered to be a central open question in TE biology (Czech et al., 2018).

Here, we investigate the evolution of piRNA clusters within and between four *Drosophila* species. By combining long-read-based assemblies with a recently developed approach for identifying homologous piRNA clusters (CUSCO (Wierzbicki et al., 2021)) and a newly developed software for generating multiple alignments of repetitive regions (Manna), we are able to shed light on the evolution of piRNA clusters. While piRNA clusters are 70%–80% conserved within species, they share almost no similarity between species as closely related as *D. melanogaster* and *D. simulans*. Many polymorphic (i.e. any variation within species irrespective of its population frequency) insertions and deletions within clusters are maintained in *Drosophila* populations. The evolutionary forces dictating the observed patterns appear to be large insertions of recently active TEs and smaller deletions of older TE insertion. Due to this rapid turnover, homologous piRNA clusters frequently do not contain insertions from the same TE families. Using our approach of combining CUSCO and Manna, we established a framework to study piRNA cluster evolution quantitatively within and between species.

2 | METHODS

2.1 | Overview of approaches for comparing the composition of homologous piRNA clusters

In principle, three different approaches for quantifying differences in the composition of homologous piRNA clusters are feasible (Table 1). With the 'Alignment First' approach, a multiple sequence alignment of the sequences of piRNA clusters is performed and repeats are annotated in the sequences of the piRNA clusters. To obtain estimates of the fraction of homologous TE insertions or the population frequency of TE insertions, it is however necessary to link the repeat annotation to the multiple sequence alignment (MSA). Unfortunately, there are currently no tools available for this task. Linking the MSA and the repeat annotation is challenging as it is not clear on how to best deal with complex alignments where for example parts of the TEs align to multiple insertions in the homologous clusters or with alignments between different TE families. MSAs at the nucleotide level are ignorant of the higher order annotations and may therefore align subsequences of related TE families (e.g. *Tirant* and *ZAM* (Marsano et al., 2000)), although this may not be desirable from an evolutionary perspective as mutations will rarely convert a TE insertion from one family to another one. The difficulty of linking the MSA to the repeat annotation motivated us to pursue

TABLE 1 Overview of approaches for quantifying differences in the composition of homologous piRNA clusters. Software NGMLR + Sniffles (Sedlazeck et al., 2018), ClustalW (Thompson et al., 1994) and RepeatMasker (Smit et al., 1996-2010)

Approach	Alignment first	Annotation first	SV caller
Steps of pipeline	1. Multiple sequence alignment (MSA) 2. Repeat annotation 3. Linking MSA and repeat annotation	1. Repeat annotation 2. Multiple alignment of repeat annotation	1. Align long reads to reference genome 2. Identify SVs 3. Test if SV sequence corresponds to TE
Tool, step 1	for example ClustalW	for example RepeatMasker	for example NGMLR
Tool, step 2	for example RepeatMasker	for example Manna	for example Sniffles
Tool, step 3	no tool available	-	for example RepeatMasker
Input	(i) Assemblies for samples of interest (ii) TE library	(i) Assemblies for samples of interest (ii) TE library	(i) A reference genome (ii) Long reads for a sample of interest (iii) TE library
Pros	(i) The entire sequence of piRNA clusters is aligned	(i) Simple pipeline (ii) Allows to avoid alignments between different TE families	(i) No assembly for strain of interest is required (the raw long reads are sufficient)
Cons	(i) No tool is available for linking the MSA and the repeat annotation (ii) Alignments between different TE families can not be avoided	(i) Solely annotated regions are aligned (ii) Less accurate inference of homologous insertions due to loss of sequence information	(i) Only pairwise alignments are feasible

an alternative approach and to develop Manna ('Annotation First'; Table 1). With this approach, the repeats are first annotated in the sequences of piRNA clusters and a multiple alignment is performed directly with the repeat annotations (Figure 2, Table 1). Information about the population frequency of TE insertions or the fraction of aligned TE sequences (i.e. the similarity) can be easily obtained from the resulting multiple alignments of the TE annotations. This approach is simple, requiring only few bioinformatics steps, identifies the most likely homologous TE insertions by a multiple alignment algorithm and allows to avoid alignments between different TE families. However, as a disadvantage this approach ignores information of sequences that are located between the annotated TEs.

If solely clusters among two samples need to be compared, it is also feasible to rely on a structural variant caller such as Sniffles (Table 1; Sedlazeck et al., 2018).

2.2 | Validation of Manna

We thoroughly validated our approach for comparing the composition of piRNA clusters with simulated genomes (Figure 2, Figure S24). We simulated populations consisting of five haploid genomes (Figure S24a, grey). TE insertions with different population frequencies were introduced into the naive genomes using SimulaTE (Kofler et al., 2018). Next, we annotated TEs in these genomes using RepeatMasker (Smit et al., 1996-2010), aligned the annotations using Manna (Figure S24a) and tested whether the observed and expected population frequencies of the TE insertions agree. To avoid mismatches between different TE families, we set the gap penalty to a lower value than the mismatch penalty. As a consequence, the position of gaps and thus the ordering of the TE in the alignment will be arbitrary in some cases, for example when different TE families are inserted into orthologous regions of the aligned strains (Figure S24b). Ambiguous ordering may be a problem in the validation as it will make it difficult to infer the expected population frequency for the observed TE insertions in the alignment. We first simulated five genomes with 246 TE insertions (2 for each of 123 TE families found in *D. melanogaster*) with a random population frequency between 1 and 5 (Figure S24c). To avoid ambiguous ordering of the TE insertions, one genome contained each of the 246 TE insertions (Figure S24c, marked by red dot). Our approach accurately reproduced the ordering as well as the population frequency of each TE insertion (Figure S24c). This implies that the alignment is correct, that is solely homologous insertions were aligned. Finally, we validated our approach with a challenging scenario, where in addition to the position and the frequency also the haplotype was randomly selected (Figure S24d). As a consequence, no individual had all 246 TE insertions. Both insertions of a given TE family had the same population frequency, which allows us to unambiguously infer the expected population frequency for any TE insertion in the alignment (i.e. expectations are based on the identity of the family and not the order of the insertions; Figure S24d). We correctly estimated the population frequency for the vast majority of the TE insertions (230 out of 246; Figure S24d). For 16 TE insertions, the population frequency was slightly underestimated (Figure S24d). Several more validations of Manna and the code to reproduce the validations are available at https://sourceforge.net/p/manna/wiki/Home/#validation.

2.3 | Long-read assemblies and data

The two *D. simulans* lines *SZ232* and *SZ45* were collected in California from the Zuma Organic Orchard in Los Angeles, CA, on two consecutive weekends of February 2012 (Signor et al., 2017;

Signor New et al., 2017, Signor, 2020). We sequenced *SZ232* and *SZ45* on a MinION platform (Oxford Nanopore Technologies (ONT), Oxford, GB), with fast base-calling using guppy (v4.4.2) and assembled with Canu (v2.1) (Koren et al., 2017) and two rounds of polishing with Racon (v1.4.3) and Pilon (v1.23) (SRR3585779, SRR3585440) (Walker et al., 2014, Vaser et al., 2017, Signor, New, et al., 2017).

The *D. simulans* strain *m252* was collected 1998 in Madagascar, and the assembly was generated with PacBio reads (Nouhaud, 2018). The *D. simulans* strain w^{xD1} was originally collected by M. Green, likely in California, but its provenance has been lost (personal communication Jerry Coyne). The *D. melanogaster* strain *A4* was sampled 1963 in Koriba Dam (Zimbabwe) (King et al., 2012). The reference strain *Iso-1* of *D. melanogaster* was generated by crossing several laboratory strains, with largely unknown sampling data (Brizuela et al., 1994). *Canton-S* was sampled 1935 in Ohio (USA) (Anxolabehere et al., 1988). We could not obtain details on the sampling of the *D. sechellia* strain *sech25* (*Robertson 3C*) and the *D. mauritiana* strain *mau12* (*w12*) (Chakraborty et al., 2021). The assemblies of the *D. melanogaster* strain *A4* (GCA_003401745.1), the *D. simulans* strain w^{xD1} (GCA_004382185.1), the *D. sechellia* strain *sech25* (GCA_004382195.1) and the *D. mauritiana* strain *mau12* (GCA_004382145.1) are based on PacBio reads (Chakraborty et al., 2018, Chakraborty et al., 2021). The assembly of the *D. melanogaster* strain *Canton-S* (GCA_015832445.1) was generated using ONT reads (Wierzbicki et al., 2021). We obtained the assembly of the *D. melanogaster* reference strain *Iso-1* from FlyBase (release 6 (Hoskins et al., 2015)). For a subset of the analyses, we additionally used a different assembly of the *D. melanogaster* reference strain *Iso-1* (PBcR-BLASR (Khost et al., 2017)) and an assembly of the *D. simulans* strain *w501* (GCA_016746395.1).

2.4 | Identifying homologous piRNA clusters

Previously, we designed flanking sequences for 85 out of the 142 annotated piRNA clusters in *D. melanogaster* (Wierzbicki et al., 2021, Brennecke et al., 2007). We excluded piRNA clusters at the end of chromosomes where two flanking sequences cannot be found, as well as clusters on the fragmented U chromosome. We aligned the sequences flanking piRNA clusters in *D. melanogaster* to each assembly using bwa bwasw (0.7.17-r1188 (Li & Durbin, 2010)) and bwa mem -a (show alternative hits) to identify clusters that were not recovered by bwa bwasw. We identified homologous clusters as the regions between the aligned *D. melanogaster* flanking sequences (Wierzbicki et al., 2021) and excluded cluster sequences with internal gaps. We validated the homology of clusters with a reciprocal mapping approach. First, we designed independent sets of flanking sequences in the target strain (e.g. *D. simulans*) that did not overlap with the aligned *D. melanogaster* flanking sequences. Second, we aligned these reciprocal flanking sequences with bwa bwasw and bwa mem -a to release 5 of the *D. melanogaster* reference genome (piRNA clusters were annotated in release 5 (Brennecke et al., 2007)). Finally, we checked whether the coordinates of the annotated piRNA clusters were contained within the positions of the aligned reciprocal flanking sequences (Tables S2–S4).

2.5 | Assembly quality of piRNA clusters

Even when both flanking sequences align to the same contig, a piRNA cluster may be incorrectly assembled, for example if some internal sequences are missing in the assembly. We previously proposed that heterogeneity of the base coverage (e.g. due to repeat collapse) and an elevated soft-clip coverage (resulting from unaligned read termini) can be used to identify assembly errors in clusters (Wierzbicki et al., 2021). To examine these patterns in our assemblies, we aligned the long reads used for generating the assembly back to the respective assembly using minimap2 (v2.16-r922; v2.17-r954) (Li, 2018). The exception to this was *D. melanogaster Iso-1* where the long reads are not from the original assembly but from a slightly diverged substrain (Solares et al., 2018). As reference, we computed the 99% quantiles of the base and soft-clip coverage of complete BUSCO (Benchmarking Universal Single-Copy Orthologs (v3.0.2; v5.0.0) (Simao et al., 2015)) genes based on the *Diptera_odb10* data set. We solely considered reads with a minimum length of 5 kb and a mapping quality of 60. Regions where the base or the soft-clip coverage markedly deviates from the 99% quantile of the BUSCO genes could indicate an assembly error and serve as a guide to the quality of the overall cluster assembly.

2.6 | Aligning the annotations of piRNA clusters

To align the TE annotations of homologous piRNA clusters, we first extracted the sequences of the clusters from the assemblies with samtools (v1.9 (Li et al., 2009)) based on the positions of the aligned flanking sequences. Next, we annotated TEs in these sequences using RepeatMasker (open-4.0.7) with a *D. melanogaster* TE library and the parameters: -s (sensitive search), -nolow (disable masking of low complexity sequences) and -no_is (skip check for bacterial IS) (Smit et al., 2013-2015, Bao et al., 2015, Quesneville et al., 2005). Finally, we aligned the resulting repeat annotations with our novel tool Manna (see Results) using the parameters -gap 0.09 (gap penalty), -mm 0.1 (mismatch penalty) and -match 0.2 (match score).

2.7 | Visualizing piRNA clusters

For visualizing the composition and evolution of piRNA clusters, we annotated repeats in piRNA clusters using the *D. melanogaster* TE library and RepeatMasker (open-4.0.7 (Smit et al., 2013-2015, Bao et al., 2015, Quesneville et al., 2005)). We identified homologous sequences in piRNA clusters with blastn (BLAST 2.7.1+ (Altschul et al., 1990)) using default parameters. We visualized the annotation and the sequence similarity of piRNA clusters with Easyfig (v2.2.3 08.11.2016) (Sullivan et al., 2011) setting the similarity scale to a minimum of 70%. Finally, we merged the pairwise visualizations generated by Easyfig to allow comparing multiple clusters. A walkthrough for this pipeline is available at https://sourceforge.net/p/manna/wiki/piRNAclusterComparison-walkthrough/.

2.8 | piRNAs

We obtained previously published piRNA data from ovaries of *D. simulans* (ERR1821669) and *D. melanogaster* (ERR1821654) strains sampled from Chantemesle (France) (Asif-Laidin et al., 2017). We trimmed the adaptor sequence (TGGAATTCTCGGGTGCCAAG) with cutadapt (v3.4 (Martin, 2011)). We aligned the reads to the reference genomes (*D. melanogaster*: *Iso-1*, *D. simulans*: w^{xD1} with novoalign (V3.03.02; http://novocraft.com/) and obtained the coordinates of the piRNA clusters from the aligned flanking sequences (see above). We retained reads with a length between 23 and 29bp, normalized the abundance of these reads to a million mapped reads and visualized the abundance of ambiguously ($mq = 0$) and unambiguously ($mq > 0$) mapped reads along piRNA clusters with R (v3.6.1) and ggplot2 (v3.3.3) (R Core Team, 2012, Wickham, 2016).

3 | RESULTS

3.1 | Identification of homologous piRNA clusters

To shed light on the evolution of piRNA clusters, we compared the composition of clusters among related *Drosophila* species. *D. sechellia*, *D. mauritiana* and *D. simulans* are closely related, having an estimated divergence time of 0.7 million years, while *D. melanogaster* diverged from this group 1.4 million years ago (Figure 1a (Obbard et al., 2012)). We relied on long-read assemblies as they allow for end to end reconstruction of piRNA clusters and their TE content and thus promise to provide a complete picture of cluster evolution (Wierzbicki et al., 2021). Since we are interested in the evolution of clusters both within and between species, we obtained long-read assemblies of several strains for *D. melanogaster* and *D. simulans*. In total, we analysed nine long-read based assemblies, four of *D. simulans*, three of *D. melanogaster* and one each of *D. sechellia* and *D. mauritiana*. All analysed assemblies are of high quality based on classic quality metrics and metrics developed to assess the assembly quality of repetitive regions (Table S1 (Wierzbicki et al., 2021)).

The number of identified piRNA clusters varied considerably between the strains and species, ranging from 73 clusters in *D. melanogaster Iso-1* to 23 clusters in *D. simulans SZ45* (Figure 1b,c). To study the evolution of piRNA clusters between species, we focused on 20 piRNA clusters shared between *D. mauritiana*, *D. sechellia* and the three best assemblies of *D. melanogaster* and *D. simulans* (Figure 1c; yellow). Most notably, our analysis included clusters *42AB* (cluster 1), *20A* (cluster 2) and *38C* (cluster 5) but not *flamenco*. Except for cluster *20A*, which is an uni-strand cluster that is expressed in the germline and the soma, all analysed clusters are dual-strand clusters that are solely expressed in the germline (Mohn et al., 2014, Brennecke et al., 2007). By investigating the heterogeneity of the base coverage and the soft-clip coverage—two recently proposed metrics for identifying assembly errors in piRNA clusters (Wierzbicki et al., 2021)—we ascertained

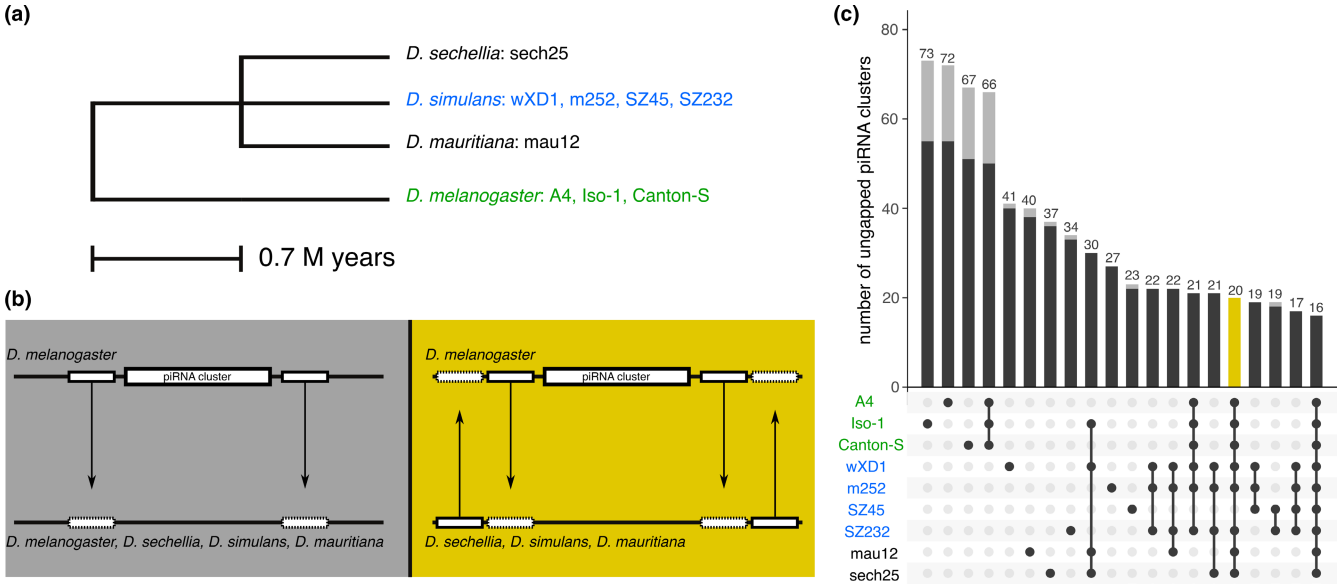

FIGURE 1 Overview of the species and piRNA clusters used in this work. (a) Phylogenetic tree showing the evolutionary distance between the four species investigated in this work (based on Obbard2012). The analysed strains are shown after the species name. (b) Our approach for finding homologous piRNA clusters in the different species and strains. Unique sequences flanking the piRNA clusters in *D. melanogaster* were aligned to a target strain. An homologous cluster was identified when both flanking sequences aligned to the same contig (grey). We confirmed homology of clusters by designing flanking sequences in the target strain and aligning them back to *D. melanogaster* reference genome (yellow, 'reciprocal flanks'). (c) Number of ungapped piRNA clusters found in different species, strains and sets of strains/species as indicated by the lines linking the samples of a set in the lower panel. Strains of *D. melanogaster* and *D. simulans* are shown in green and blue, respectively. Colours of bar (grey or yellow) correspond to the approach used for identifying homologous clusters (see b). Fraction of ungapped clusters unique to species or overlaps are shown in light grey

that the assemblies of the 20 clusters are of high quality (see Materials and Methods; Figures S1–S5). Based on publicly available small RNA data from ovaries of a *D. melanogaster* and *D. simulans* strain collected in Chantemesle (France (Asif-Laidin et al., 2017)), we estimate that 15 out of the 20 investigated clusters are expressed in both species (>10 reads per million; Figures S6–S8; this may be an underestimate as small RNAs were not derived from an assembled strain).

3.2 | Comparing the composition of homologous clusters

piRNA clusters are often referred to as 'TE graveyards' since they are thought to carry the remains of past TE invasions, which results in a high repeat content. This highly repetitive nature makes it difficult to compare the composition of homologous clusters, for example using multiple sequence alignments. We approached this problem inspired by the alignments of amino-acid sequences, which are performed at a higher level than the underlying nucleotide sequences. Here, we propose that multiple alignments may be performed with the TE annotations (e.g. generated by RepeatMasker) of piRNA clusters instead of the nucleotide sequences. For this reason, we developed Manna (multiple annotation alignment), a novel tool performing multiple alignments of annotations. Although primarily designed for annotations of repeats, it may work with the annotations of any feature. Manna performs a progressive alignment similar to that described by Feng & Doolittle (1987). Using a simple scoring scheme (Figure S9) and an adapted Needleman–Wunsch algorithm (Needleman & Wunsch, 1970), a guide tree is computed. Based on this tree, the most similar annotations are aligned first, followed by increasingly more distant annotations. For the scoring matrix, the score of each newly aligned annotation is computed as the average score of the previously aligned annotations (Feng & Doolittle, 1987).

This novel tool enables us to compare the composition of homologous clusters using the following approach: first, we align pairs of sequences flanking piRNA clusters to the assemblies, thereby identifying the positions of homologous clusters in each assembly (Figure 2a). Second, we extract the sequences delimited by these pairs of flanking sequences (Figure 2b). Third, we annotate repeats in the extracted sequences (Figure 2c) and solely retain the repeat annotation (Figure 2d). Finally, we align the repeat annotation with Manna (Figure 2e). Using simulated sequences with varying repeat contents, we carefully validated this approach for comparing the composition of homologous piRNA clusters (Materials and Methods).

Alignments with Manna allow us to quantify (i) the number of polymorphic and fixed TE insertions and (ii) the similarity *s* and the distance ($d = 1 - s$) among homologous clusters. The similarity (*s*) between clusters is computed as $s = 2 * a/(2 * a + u)$ where *a* and *u* are the total length of aligned and unaligned TE sequences, respectively (see, e.g., Figure S10). This similarity can be intuitively interpreted as the fraction of TE sequences that can be aligned between two (homologous) clusters.

Alignments with Manna do not incorporate unannotated sequence in between TEs (Figure 2c). Therefore, we additionally investigated the similarity among homologous clusters using a complementary approach: we identified similar sequences between clusters with BLAST (minimum identity 70% (Altschul et al., 1990)) and visualized these similarities and the repeat content of clusters with Easyfig (Figures S11–S15).

3.3 | Rapid evolution of piRNA clusters

To quantify the rate at which piRNA clusters evolve, we estimated the evolutionary turnover of the TE content of the 20 piRNA clusters using the similarity (*s*) as computed with Manna (see above). Based on the distance between the clusters ($d = 1 - s$), we additionally generated phylogenetic trees reflecting these distances (Figure 3a).

Strikingly, an average of solely 8.1% of the TE sequences can be aligned between the piRNA clusters of *D. melanogaster* and *D. simulans* (Figure 3a; Table S5). Among the 20 clusters, the similarity ranged from 0.0% for clusters 19 and 110 to 93.5% for cluster 114 (length weighted median: 3.7%; Table S5). Within the more closely related species of the *simulans*, complex 41.4% of the TE sequences can be aligned between *D. simulans* and *D. mauritiana* (range: 0.0%–100%; length weighted median: 32.7%) and 32.7% between *D. sechellia* and *D. simulans* (range: 0.0%–88.8%; length weighted median: 24.8%; Table S5). Our data thus suggest that the clusters of *D. simulans* are more closely related to *D. mauritiana* than to *D. sechellia*. This is in agreement with previous works, which suggested that *D. simulans* is more closely related to *D. mauritiana* than to *D. sechellia* (Lachaise & Silvain, 2004).

Given this rapid turnover of piRNA clusters between species, we hypothesized that clusters could also evolve rapidly within species. In agreement with this, we found that the average similarity of clusters within species is 73.12% for *D. melanogaster* (range: 33.3%–100%; length weighted median: 74.2%) and 74.7% for *D. simulans* (range: 0.0%–100%; length weighted median: 75%; Table S5). The similarity of clusters between species is significantly different from the similarity within species (Wilcoxon rank-sum tests: *Dmel* vs. *Dmel-Dsim* $W = 34.5$, $p = 0.000008$; *Dsim* vs. *Dmel-Dsim* $W = 53.5$, $p = 0.00008$; *Dsim* vs. *Dsim-Dsec* $W = 79$, $p = 0.001$; *Dsim* vs. *Dsim-Dmau* $W = 113$, $p = 0.02$). On average, 26% of the TE sequences in piRNA clusters cannot be aligned between two assemblies of the same species. The TE content of clusters is thus highly polymorphic within species.

Since the strains analysed in *D. simulans* and *D. melanogaster* were collected at very diverse time points and geographic locations, we speculated that the similarity among strains sampled from the same population may be higher. A comparison of 16 clusters shared between the Californian *D. simulans* strains *SZ232* and *SZ45*, which were collected at the same location and date, and an African strain (*m252*) and an old Californian strain (w^{xD1}, likely collected approximately 50 years prior) did not confirm this hypothesis (similarity between *SZ232* vs. *SZ45*: 72.5%; average similarity among all other

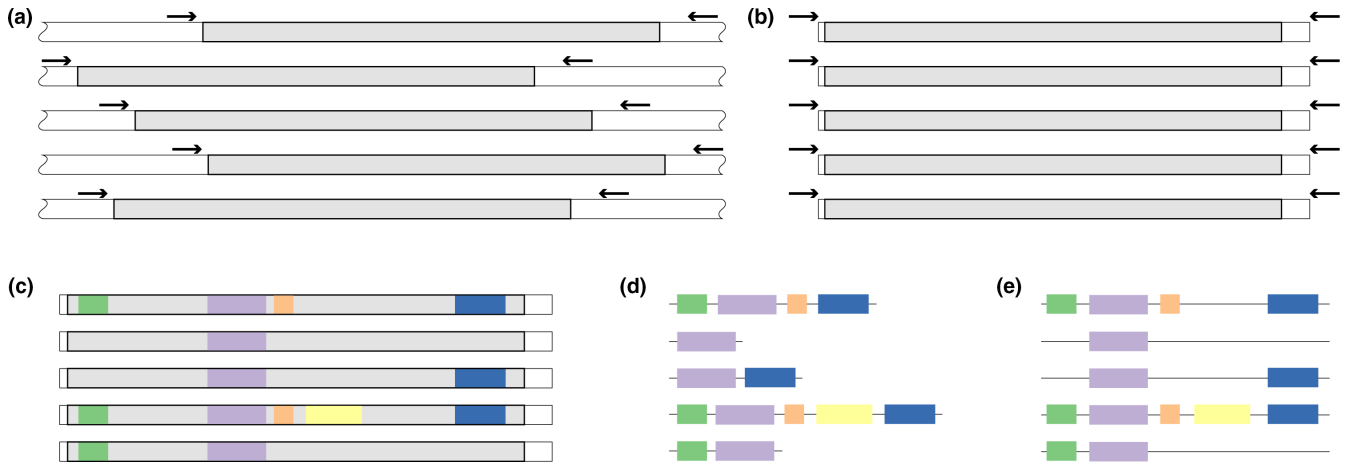

FIGURE 2 Overview of our approach for comparing the composition of piRNA clusters. (a) To identify homologous piRNA clusters (grey areas) in the strains, we mapped sequences flanking the piRNA clusters (black arrows) to the assemblies. (b) Regions delimited by the flanking sequences were extracted (i.e. the piRNA clusters plus the short sequences between the clusters and the flanking sequences). (c) Repeats were annotated in the extracted sequences. (d) Solely the repeat annotations were retained for further analysis. (e) The repeat annotations were aligned with Manna allowing us to compare the repeat content of piRNA clusters

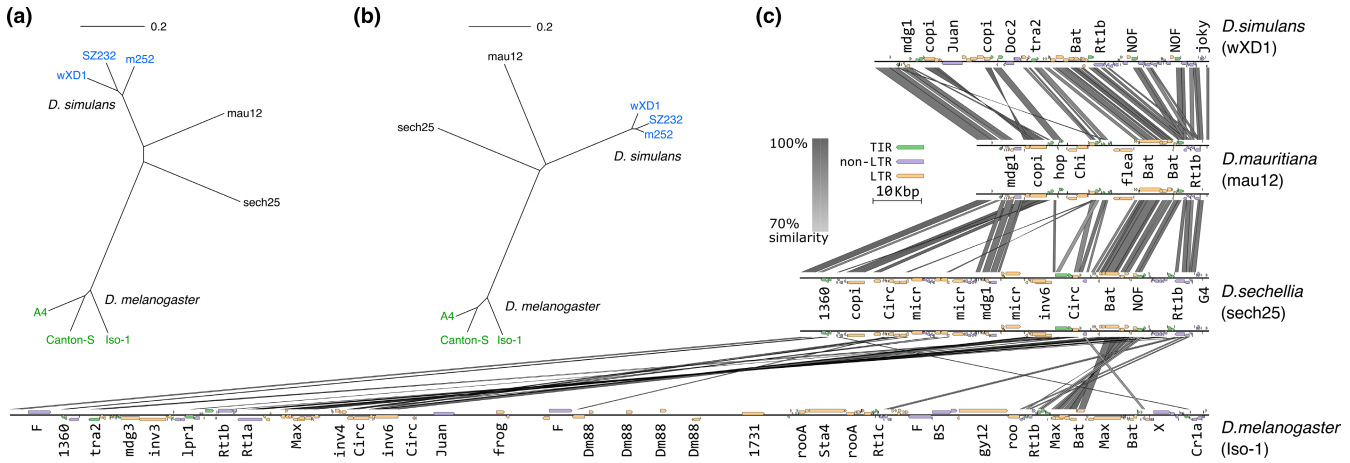

FIGURE 3 piRNA clusters are rapidly evolving in *Drosophila* species. (a) Phylogenetic tree summarizing the distance of the 20 piRNA clusters among the different strains and species weighted by the average cluster lengths. The distance is estimated by Manna as the fraction of unaligned TE sequences (scale bar shows a distance of 20%). Note that solely about 8.1% of the TE sequences can be aligned between the clusters of *D. melanogaster* and *D. simulans*. (b) Phylogenetic tree for the piRNA cluster *42AB* (cluster 1) based on alignments with Manna. (c) The evolution of piRNA cluster *42AB* in four *Drosophila* species visualized with Easyfig. Homology among the sequences (grey bars) was determined with BLAST. The grey scale indicates the degree of the sequence similarity. Homology blocks smaller than 400 bp are not shown. Insertions of TEs are shown as small rectangular arrows where the colour indicates the order (LTR, non-LTR and TIR). Family names are abbreviated

D. simulans strains: 75.8%; Table S6). The clusters of strains sampled from the same population are thus not necessarily more similar than the clusters of strains sampled from different regions and time points (although the results vary among the clusters).

Next, we aimed to investigate the evolution of cluster *42AB* (cluster 1) in more detail. In *D. melanogaster*, *42AB* is one of the largest clusters that may account for 20%–30% of all piRNAs (Brennecke et al., 2007). It is thus frequently highlighted as a canonical piRNA cluster (e.g. Czech et al., 2008, Mohn et al., 2014, Olovnikov et al., 2013, Andersen et al., 2017). A phylogenetic tree based on an alignment of annotated TEs shows that cluster *42AB* is rapidly evolving among the investigated *Drosophila* species (Figure 3b; for a tree for all other clusters, see Figure S16). The similarity of *42AB* between *D. simulans* and *D. melanogaster*, based on an alignment of TE annotations using Manna, is solely 4%. Within the *simulans* clade, the similarity of *42AB* between *D. simulans* and *D. mauritiana* is 29.6%, and between *D. simulans* and *D. sechellia*, it is 26.4% (Table S5). Within species, cluster *42AB* is more variable in *D. melanogaster* (similarity: 77.5%) than in *D. simulans* (similarity: 90.3%; Table S5). As alignments with Manna only capture similarities of annotated TEs, we also visualized the

evolution of cluster *42AB* using BLAST and Easyfig (Figure 3c). This approach confirms our findings. Cluster *42AB* has few sequence similarities between *D. melanogaster* and *D. simulans* and a higher level of sequence similarity among the species of the *simulans* complex (Figure 3c). We conclude that cluster *42AB* is rapidly evolving in the investigated species (Figure 3c). For a visualization of the sequence similarity of all 20 clusters in the four species, see Figures S11–S15.

Thus far, we showed that the sequence of piRNAs clusters is evolving very quickly between and within species. However, it is possible that this rapid evolution is due to rearrangements within piRNA clusters (Gebert et al., 2021), while the TE content of clusters actually remains stable. We addressed this question by quantifying the number of insertions from each TE family in each cluster, and determining if at least one insertion of a given family is present in a given cluster in *D. simulans*, *D. melanogaster* or both species (an insertion in any of the three strains of each species was considered as a presence). For example, we considered *blood* to be present in cluster *42AB* in both species when a single *blood* insertion was found in *42AB* of *A4* (*D. melanogaster*) and *m252* (*D. simulans*) but not in any other strain of the two species. The rapid evolution of piRNA clusters does not appear to be due to rearrangements, as the presence of TE families was also not conserved across species (Figure 4). Out of 321 TE families in piRNA clusters, only 76 were present in both species (families present in more than one cluster were counted multiple times). 164 were private to *D. melanogaster* and 81 to *D. simulans* (Figure 4). A similar observation can be made when we compare the TE composition of piRNA clusters among *D. simulans*, *D. mauritiana* and *D. sechellia* (Figure S17).

We thus conclude that piRNA clusters are rapidly evolving in *Drosophila* species, such that the average, only about 8% of TEs sequences were aligned between the closely related *D. melanogaster* and *D. simulans*. Furthermore, homologous clusters frequently contain different TE families.

3.4 | piRNA clusters in *D. melanogaster* and *D. simulans* genotypes

The amount of variation in the composition of piRNA clusters within species is an important open question as this variation may lead to piRNA clusters regulating different TE families in different strains (Chen & Aravin, 2021). We thus investigated variation in the piRNA clusters of *D. melanogaster* and *D. simulans* in more detail, incorporating several genotypes from each species. An alignment of the 20 clusters with Manna in the three strains of *D. melanogaster* (*Dmel*) and *D. simulans* (*Dsim*) shows that clusters in *D. melanogaster* contain more TEs than in *D. simulans* (*Dmel* = 1,002, *Dsim* = 547). The majority of these insertions are fixed (*Dmel* = 647, *Dsim* = 362; Figure 5a), but a substantial number of TE insertions is segregating in one (*Dmel* = 229, *Dsim* = 118) or two genotypes (*Dmel* = 126, *Dsim* = 67). Despite these differences in the TE abundance among the two species, the site frequency spectrum of the cluster insertions is very similar between *D. melanogaster* and *D. simulans* (Chi-squared test $p = 0.20$; Figure 5a). The large number of polymorphic cluster insertions is not contingent upon a single outlier-genotype since all genotypes from both species carried abundant polymorphic cluster insertions (*D. melanogaster*: *CS* = 191, *A4* = 153, *Iso1* = 137; *D. simulansSZ232* = 106, w^{xD1} = 97, *m252* = 49, Figure S18a). The polymorphic cluster insertions were distributed over 17 clusters in *D. melanogaster* and 12 clusters in *D. simulans* (Figure S18a). In agreement with the higher TE content of *D. melanogaster* clusters, piRNA clusters in *D. melanogaster* were substantially longer than in *D. simulans* (Wilcoxon rank-sum test $W = 2192$, $p = 0.040$; Figure S18b). The total size of the piRNA clusters in *D. melanogaster* was about double that of the clusters in *D. simulans* (average over all three strains *Dmel* = 817,770, *Dsim* = 452,591). In both species, segregating cluster insertions were on the average longer than fixed ones (*D. melanogaster*: *seg* = 1115, *fix* = 591, Wilcoxon rank-sum test $W = 122302$, $p = 0.089$; *D. simulans*: *seg* = 798, *fix* = 470, Wilcoxon rank-sum test $W = 38248$, $p = 0.0065$).

In addition, the amount of polymorphisms segregating in strains sampled from the same population (*SZ232*, *SZ45*) is similar to the amount of polymorphism sampled in strains from different locations (*m252*, Africa) and time points (Chi-square test, with different minimum sizes of indels; $p_{100bp} = 0.12$, $p_{500bp} = 0.87$, $p_{1000bp} = 0.98$; Figures S19, S20). While overall polymorphism was similar amongst strains, the amount of fixed and segregating TE insertions varies across the clusters of both species, albeit not significantly (Chi-squared test $p_{dmel} = 0.08$, $p_{sim} = 0.14$; Figure 5b). Some clusters in *D. melanogaster* mostly have fixed TEs such as cluster 96 (*fix* = 83, *seg* = 14) and cluster 142 (*fix* = 31, *seg* = 4), but other clusters, like cluster 1 (*fix* = 153, *seg* = 114) and cluster 45 (*fix* = 36, *seg* = 41), have large proportions of segregating TEs (Figure 5b). Similarly in *D. simulans*, some clusters such as cluster 1 (*fix* = 89, *seg* = 12) and cluster 29 (*fix* = 29, *seg* = 2) have largely fixed TEs, whereas cluster 5 (*fix* = 26, *seg* = 75) and cluster 86 (*fix* = 20, *seg* = 22) contain many segregating TE insertions. This raises the possibility that clusters evolve at different rates, with some clusters evolving faster than others. Additionally, the evolutionary turnover of the clusters might differ among species; for example, cluster *42AB* (cluster 1) evolves faster in *D. melanogaster*, whereas cluster *5* evolves faster in *D. simulans* (Figure 5b).

Our analysis is based on the consensus sequences of *D. melanogaster* TEs (Quesneville et al., 2005). We asked if this could lead to a bias where TE insertions in *D. simulans* clusters are less readily identified than in *D. melanogaster*. Such a bias could lead to a lower density of TEs in piRNA clusters of *D. simulans* as compared to *D. melanogaster*. We found that the density of TE insertions in piRNA clusters is very similar in the two species (TE insertions per kb *Dmel* = 0.994, *Dsim* = 0.985). However, cluster insertions in *D. simulans* were, on the average, slightly shorter than in *D. melanogaster* (average length *Dmel* = 777, *Dsim* = 581; Wilcoxon rank-sum test $W = 300760$, $p = 0.0015$). Although this is in agreement with previous works suggesting that TEs in *D. simulans* are shorter than in *D. melanogaster* (Lerat et al., 2011, Vieira et al., 2012), it could also be a technical artefact where, for example, terminal regions of TEs are

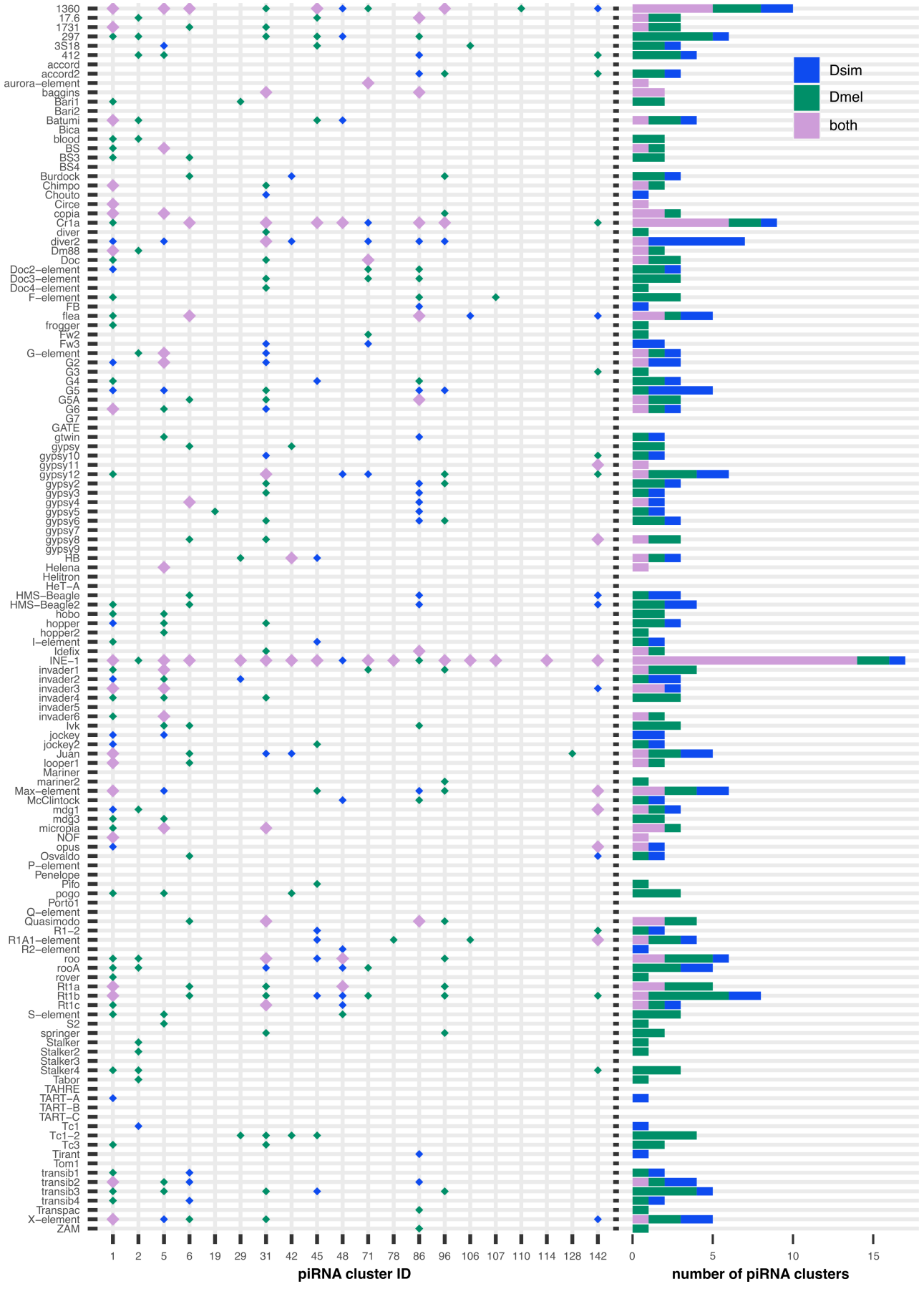

FIGURE 4 Overview of the TE content of piRNA clusters in *D. simulans* and *D. melanogaster*. For each piRNA cluster (x-axis), we indicate whether a given TE family (y-axis) has at least one insertion in *D. melanogaster* (green), *D. simulans* (blue) or in both species (purple). We considered insertions in any of the three assemblies of *D. melanogaster* and *D. simulans*. The right panel summarizes the abundance of the families in piRNA clusters. Note that the TE content of the clusters varies dramatically between the species

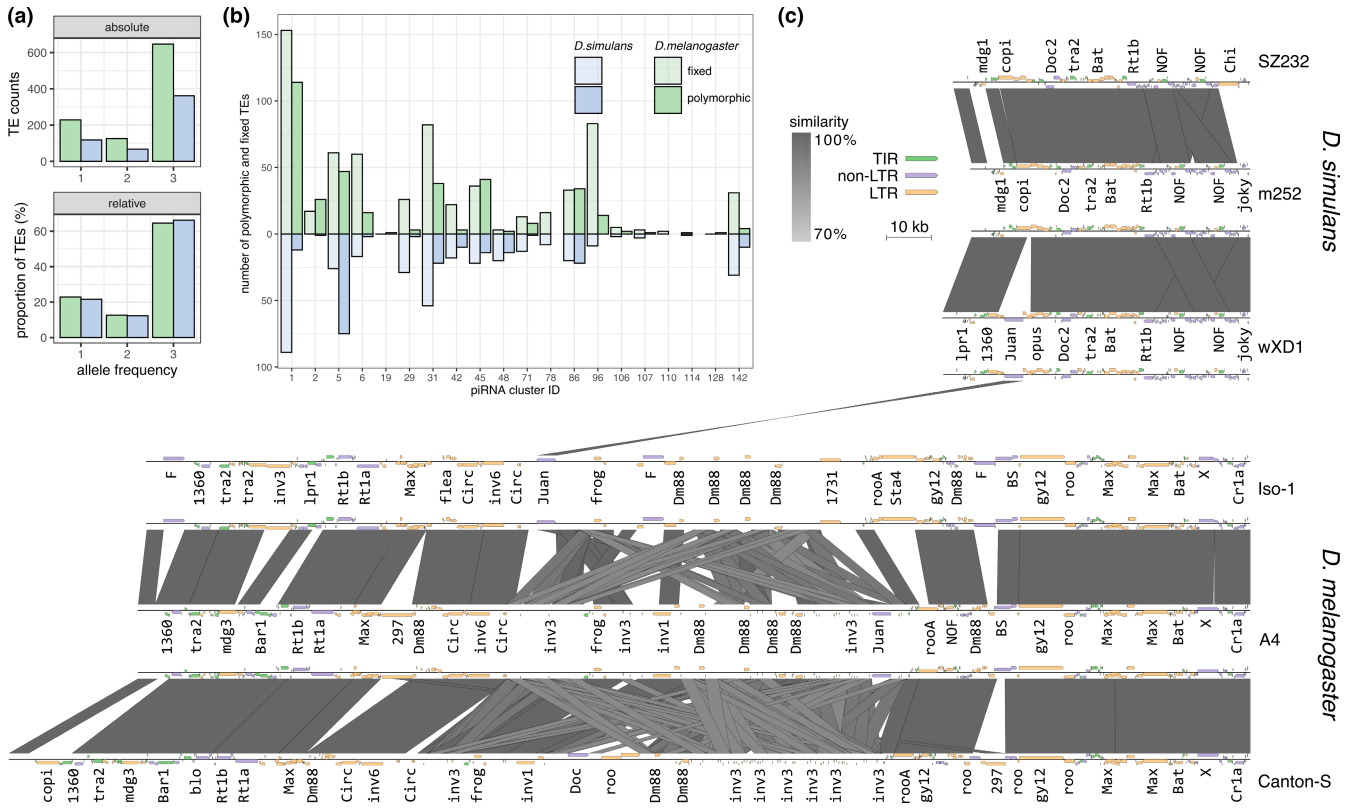

FIGURE 5 Rapid evolution of piRNA clusters within *D. melanogaster* and *D. simulans*. (a) Population frequencies of TE insertions in all 20 piRNA clusters of *D. melanogaster* (green) and *D. simulans* (blue). The absolute (top) and relative (bottom) TE abundance are shown. Insertions occurring in three individuals are fixed. (b) Numbers of fixed (transparent) and polymorphic (opaque) sites for each piRNA cluster in *D. melanogaster* (green) and *D. simulans* (blue). (c) Composition of cluster *42AB* in 3 strains of *D. melanogaster* and *D. simulans*. Grey bars indicate regions of similarity among two assemblies of *42AB* (minimum length 3 kb). TE families are coloured by order (LTR, non-LTR and TIR)

due to a high divergence from the sequence in the repeat library not annotated as TEs in *D. simulans*. To estimate the extend of missed TEs in the two species, we annotated TEs with a more extensive but less well curated repeat library containing diverse repeats from different *Drosophila* species (RepBase (Bao et al., 2015)). This analysis suggests that we may have missed some TEs in both species, albeit slightly more in *D. simulans* (fraction (%) of TEs in genome; *D. melanogaster Iso-1*: *consensus* = 16.3, *repbase* = 19.1; *D. simulans*w^{xD1}: *consensus* = 14.7, *repbase* = 19.3).

Finally, we investigated the composition of cluster *42AB* in more detail (Figure 5c). Cluster *42AB* is, consistently among the strains, shorter in *D. simulans* than in *D. melanogaster* (Figures 5c; S18b). The density of TEs in cluster *42AB* is higher in *D. simulans* (TEs per kb *Dmel* = 0.79, *Dsim* = 1.41) possibly due to the shorter TE insertions (average length of TEs in *42AB Dmel* = 920 *bp*, *Dsim* = 658 *bp*). While there is considerable sequence conservation in both species, the *D. melanogaster 42AB* cluster bears no resemblance to *42AB* in *D. simulans*, other than containing a *Juan* element, which is likely not a homologous insertion (Figure 5c). The number of segregating insertions is larger in *D. melanogaster* than in *D. simulans*, suggesting that *42AB* is evolving faster in *D. melanogaster* (Figure 5b,c). For a visualization of the sequence similarity of all clusters in the different assemblies of *D. melanogaster* and *D. simulans*, see Figures S11–S15.

We conclude that piRNA clusters are highly polymorphic in both species, that clusters have a similar TE density in both species, and that most clusters are shorter in *D. simulans* than in *D. melanogaster*. Furthermore, clusters may evolve at different rates among and within species.

3.5 | Evolutionary forces shaping the composition of piRNA clusters

Many diverse evolutionary forces may act on the TE content of piRNA clusters, such as mutations, insertion bias, negative or positive

selection and drift (Kofler, 2019, Kelleher et al., 2018, Lu & Clark, 2010, Brennecke et al., 2007, Zhang et al., 2020). While we cannot distinguish among these forces, we can shed light on their joint effect by investigating the abundance of insertions and deletions segregating in piRNA clusters. We determined the number of insertions and deletions segregating in piRNA clusters of the *D. simulans* strains by polarizing segregating indels using *D. mauritiana* as outgroup. Among the analysed species, the clusters of *D. simulans* and *D. mauritiana* are most closely related and thus best suited for this analysis (Figure 3a). We used TE insertions with a minimum length of 100 bp and considered indels resulting from presence/absence polymorphisms in the alignment and indels resulting from length differences between aligned TEs sequences. We found that 69 deletions and 199 insertions are segregating in piRNA clusters of *D. simulans* (Figure 6a). These indels were distributed over 12 of the investigated 20 piRNA clusters (Figure S21). Insertions were, on the average, longer than deletions (average length $\bar{l}_{ins} = 703\,bp$, $\bar{l}_{del} = 229\,bp$; Wilcox rank-sum test $W = 4778.5$, $p = 0.0002$). Most indels were found in three of the 20 clusters: cluster 5 (104 indels), cluster 45 (30 indels) and cluster 86 (28 indels; Figure S21). Because de novo TE insertions will likely be large, we separately analysed long indels (≥ 2000). We found 15 long insertions and a single long deletion. The most abundant long insertions were due to the TE families *roo*, *1360*, *G5*, *invader3* and *Max-element* (two for each family). These families are likely active in *D. simulans* as many insertions of these families segregate at low population frequencies (Kofler et al., 2015, Signor, 2020). Finally, we asked if insertions are occurring with younger TE families than deletions. While we do not have direct estimates for the age of TE families in *D. simulans*, we may use the average population frequency of all insertions of a family as proxy for age. Insertions of recently active families will mostly have a low frequency, whereas old families will mostly have fixed insertions. Using the frequency estimates of Kofler et al. (2015), we found that families with insertions in piRNA clusters have a significantly lower average population frequency than families with deletions ($\bar{f}_{ins} = 0.27$, $\bar{f}_{del} = 0.50$; Wilcoxon rank-sum test $W = 8896$, $p = 7.3e-07$ Figure 6b).

In summary, the evolutionary dynamics of piRNA clusters are governed by many insertions and few deletions, where insertions are on the average larger than deletions. Furthermore, insertions usually involve recently active families, whereas deletions mostly happen in older families.

4 | DISCUSSION

Here, we established a framework for studying the evolution of piRNA clusters quantitatively, used that framework to analyse the composition of 20 piRNA clusters in four *Drosophila* species, and showed that piRNA clusters are evolving rapidly.

We relied on highly contiguous long-read-based assemblies to investigate the evolution of piRNA clusters. Since few long-read-based assemblies for the investigated species are available, the small number of analysed assemblies may limit the statistical power of some conclusions in this study. For future work, it will thus be important to obtain more high-quality genomes for different *Drosophila* species. An important question is whether assembly problems could be responsible for some of our conclusions, such as the rapid evolution of piRNA clusters. All assemblies used in this work are of high quality based on classical quality metrics and metrics specifically developed to assess the assembly quality of piRNA clusters (Table S1; Figures S1–S5, S22 (Wierzbicki et al., 2021)). Furthermore, our conclusions are robust when analyses are based on a subset of high-quality assemblies and clusters. The similarity among the clusters of the two species is also low when solely analysing the best assembly for *D. simulans* and *D. melanogaster* (7.2%; Figure S22; Table S7). When we compare an

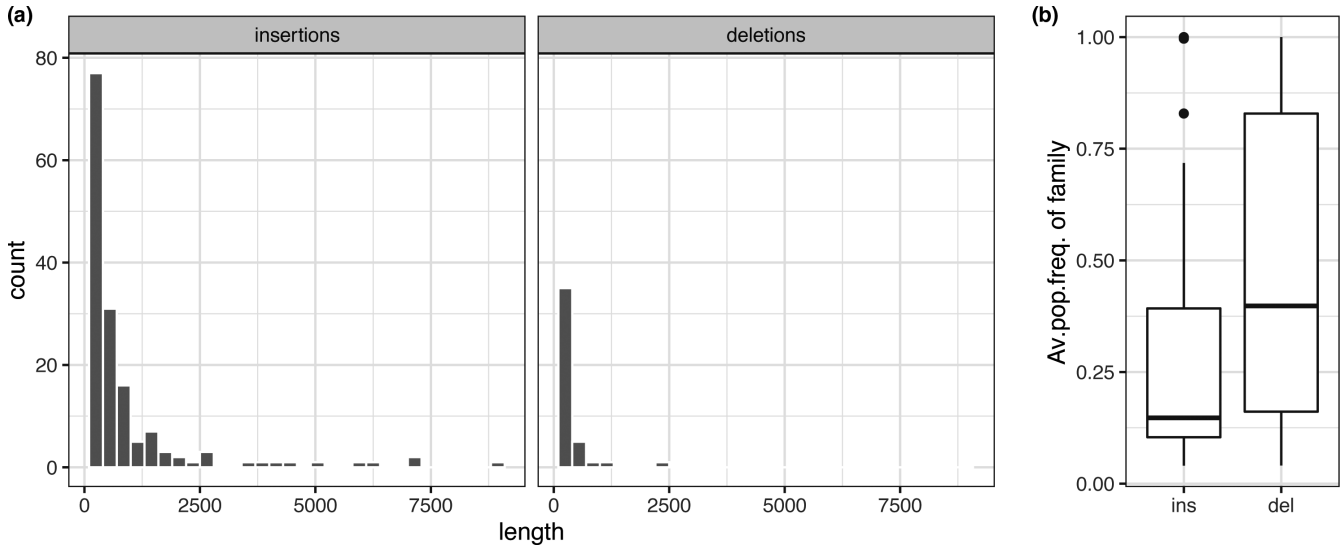

FIGURE 6 Overview of insertions and deletions in piRNA clusters of *D. simulans*. The clusters of *D. mauritiana* were used to polarize the indels. (a) Histograms showing the abundance and length of insertions and deletions. (b) Age of the families of insertions (ins) and deletions (del) in piRNA clusters, where the average population frequency (av.pop.freq.) of the family is used as a proxy for the age

additional assembly for each of *D. melanogaster* (PBcR-BLASR) and *D. simulans* (*w501*), there is again a very low similarity between the two species (9.8%; Table S7). Finally, our key findings hold when solely a set of the 10 clusters with the highest quality is analysed (Figures S22, S23). We thus argue that our results are largely robust with respect to assembly quality.

Another important question is whether the 20 piRNA clusters included in the analysis are a representative set of the 141 piRNA clusters of *D. melanogaster*. The 20 analysed piRNA clusters account for 55% of uniquely mapping piRNAs (Brennecke et al., 2007), suggesting that the clusters are at the least representative for the majority of the generated piRNAs. However, piRNA clusters were excluded from our analysis for three reasons: (i) clusters were at the end of a chromosome or on the unassembled U chromosome, which did not allow us to identify suitable flanking sequences; (ii) a cluster could not be assembled in all species without gaps, possibly due to complex repeat content; and (iii) we could not identify conserved flanking sequences in all species such that the homology of a cluster could be established. While the first point likely does not introduce a bias, the last two points could potentially result in a bias towards shorter or less complicated clusters. An analysis of five additional clusters shared among the 'best' assembly of *D. melanogaster* and *D. simulans* (*Canton-S* and w^{xD1}; Figure S22; Table S8) shows that the similarity of these 5 additional clusters between *D. melanogaster* and *D. simulans* (8.5%) is very close to our estimate based on the 20 clusters and 3 strains for each species (8.1%; Tables S5, S9). We thus think that our 20 clusters are largely a representative set of the piRNA clusters. To gain a more comprehensive picture, it will be important to extend the analysis performed in the present work to a larger number piRNA clusters. It is possible that investigating alternate flanking sequences could lead to an increase in the number of clusters, and rapid advances in sequencing technology will increase the number of contiguously assembled clusters. However, a comparison between species will always be less than entirely comprehensive, as clusters may not be shared between species of interest or the flanking sequences may have degraded beyond recognition. In agreement with this, previous research has noted that many piRNA clusters are species-specific (Gebert et al., 2021, Chirn et al., 2015).

This and other works established synteny of piRNA clusters based on sequences flanking the cluster up- and downstream (Gebert et al., 2021, Chirn et al., 2015). It is unclear if this is the best approach for finding homologous clusters. In principle, it is possible to use the sequence (or annotation) of piRNA clusters directly to search for the homologous clusters in an assembly of interest (e.g. with BLAST). However, given how rapidly piRNA clusters evolve, where solely 8% of TE sequences can be aligned between *D. melanogaster* and *D. simulans*, it is doubtful whether this approach will be able to correctly establish homology of the piRNA clusters. We quantified the similarity of clusters and the amount of polymorphism in clusters with our novel multiple alignment tool Manna. As a major innovation, this tool performs a multiple alignment with repeat annotations rather than the raw sequences. While this approach provides invaluable insight into the evolution of piRNA clusters, it does ignore some information such as divergence of the TEs. Alignments of clusters at the nucleotide level may be more sensitive, but this approach has its own problems (see Materials and Methods).

We found that *D. simulans* has fewer TE insertions in piRNA clusters than *D. melanogaster*. That this is a real pattern is supported by the similar density of TEs in the two species within the piRNA clusters (indicating no obvious presence of unannotated TEs in *D. simulans*) and the shorter length of piRNA clusters. Given that *D. melanogaster* and *D. simulans* share the vast majority of the TE families, likely because of shared ancestral TEs and a high rate of horizontal transfer between the species (Sanchez-Gracia et al., 2005, Schwarz et al., 2021, Kofler et al., 2015, Lerat et al., 2011), it seems unlikely that many TE families specific to *D. simulans* have been missed. In agreement with this, the TE abundance in both species was only slightly increased when a more comprehensive but less well curated TE library was used.

Based on the sequences flanking piRNA clusters, previous work has shown that the synteny of clusters is evolving quickly (i.e. flanking sequences can frequently not be found in other species) (Gebert et al., 2021, Chirn et al., 2015). Our study complements this work by confirming that homologous clusters are frequently not found among closely related species, and by showing, for the first time, that the content of piRNA clusters is evolving rapidly.

This raises the important question which evolutionary forces drive the evolution of piRNA clusters. In principle, the following forces could act on piRNA clusters. First, different types of mutations, such as insertions due to recent TE activity, the deletion bias observed in *Drosophila* or major rearrangements, for example due to ectopic recombination mediated by TE insertions, may contribute to the rapid turnover of piRNA clusters (Petrov et al., 1996, Langley et al., 1988). Many TE families are active in *Drosophila* species, so recent insertions may be an important driver of cluster evolution (Kofler et al., 2015). Also, genomic rearrangements have been implicated in the evolution of clusters (Assis & Kondrashov, 2009, Gebert et al., 2021). Second, selection (positive or negative) may contribute to the rapid evolution of piRNA clusters. Theory suggests that an invading TE is silenced by multiple segregating TE insertions distributed over many piRNA clusters (Kofler, 2019, Kelleher et al., 2018). This hypothesis has been confirmed experimentally by recent works investigating the distribution of cluster insertions in natural and experimental populations that were recently invaded by a TE (Zhang et al., 2020, Kofler et al., 2018). Theory further suggests that these segregating cluster insertions could be positively selected as haplotypes with a cluster insertion will accumulate few TEs overall and will thus be less deleterious than haplotypes without a cluster insertion (Kofler, 2019, Kelleher et al., 2018, Lu & Clark, 2010). However, the expected shift in the site frequency spectrum of positively selected cluster insertions is rather subtle and thus difficult to detect experimentally (Kofler, 2019). In agreement with this, a recent work did not detect evidence that cluster insertions are positively selected (Zhang et al., 2020). One drawback of this particular study is the lack of reconstruction of the entire piRNA cluster in each strain (P-element insertion sites were identified based

on alignments of short reads to a reference genome) (Zhang et al., 2020). As a consequence, P-element insertions will not be found if adjacent sequences are not conserved and the population frequency of the insertions may be estimated unreliably if the P-element inserted into repetitive regions. However, positive selection of cluster insertions could lead to an accumulation of TE insertions in piRNA clusters. Third, an insertion bias could also lead to an accumulation of TE insertions in piRNA clusters. It is likely that at least some TEs, such as the P-element, have a pronounced insertion bias into piRNA clusters (Ajioka & Eanes, 1989, Zhang et al., 2020, Kofler et al., 2018, Karpen & Spradling, 1992). It is an important open question whether other TE families also have such an insertion bias into piRNA clusters. Alternatively, piRNA clusters may attract TE insertions, for example due to protein–protein interactions (Brennecke et al., 2007, Vermaak & Malik, 2009). Finally, genetic drift could have a strong influence on the evolution of piRNA clusters. Apart from drift of cluster insertions or whole cluster haplotypes, drift may also act on the epigenetically transmitted information that determines the position of piRNA clusters. The information about the position of piRNA clusters is likely not hard coded into the DNA sequence (e.g. by motifs) but rather transmitted epigenetically by the population of maternally deposited piRNAs (LeThomas et al., 2014, Le Thomas, Stuwe, et al., 2014). Stochastic variation in the composition and the amount of maternal transmitted piRNAs could thus lead to a rapid turnover of the location of piRNA clusters. Such a rapid turnover would likely relax selection on individual cluster insertions and make detection of positive selection on cluster insertions even more challenging.

This raises the question as to which of these processes are active in the piRNA clusters investigated in the present work. The TE content of piRNA clusters is rapidly evolving, and we found that more insertions than deletions were segregating in piRNA clusters of *D. simulans*. The insertions were usually longer and occurring in younger TE families than the deletions. Most insertions are therefore likely due to recent activity of TE families in piRNA clusters. Nevertheless, some insertions (and deletions) could also be due to repeat expansion (and repeat collapse) or genomic rearrangements. A crucial question is whether the observed larger number of insertions in piRNA clusters is due to neutral processes or other forces such as positive selection on cluster insertions and an insertion bias into piRNA clusters. To distinguish between these possibilities, one would need adequate control regions, that is a regions that do not produce piRNAs, but otherwise have very similar properties to piRNA clusters (pericentromeric regions with a similar size, number, recombination rate and TE content). It is unfortunately challenging to find suitable control regions. Additionally, larger numbers of high-quality assemblies for the two *Drosophila* species may be necessary to reliably detect subtle shifts in the site frequency spectrum of the cluster insertions as expected under positive selection. However, the properties of the deletions in piRNA clusters (short and mostly in older TEs) can likely be explained by the deletion bias observed in *Drosophila*. The gradual erosion of TEs by a deletion bias could also explain why segregating insertions (likely young) are on average longer than fixed insertions (likely old). Another important open question is whether stochastic forces or other processes such as selection and insertion biases could cause differences in the rate of evolution among the piRNA clusters. It is for example possible that positive selection is stronger in clusters producing many piRNAs than in clusters producing few.

The available evidence suggests that piRNA clusters are larger in *D. melanogaster* than in *D. simulans*. This could be due to two, not mutually exclusive, reasons: first, the clusters are growing in the *D. melanogaster* lineage, or second the clusters are shrinking in the *D. simulans* lineage. Our analysis of insertions and deletions suggests that even in *D. simulans*, the clusters are evolving largely by insertions. If piRNA clusters were shrinking in the *D. simulans* lineage, we would not expect to see mostly insertions segregating in *D. simulans* populations. Therefore, it seems more likely that the piRNA clusters are expanding in both lineages but in *D. melanogaster* more than in *D. simulans*. This raises the question if the size of piRNA clusters could be subject to a runaway process, where larger clusters will accumulate more insertions of active TEs, which, when positively selected, will lead to even larger clusters. This further raises the question whether some forces counteract the expansion of piRNA clusters. Rare and large genomic rearrangements may be an option.

We showed that the sequence and the TE content of piRNA clusters are rapidly evolving. This raises another important question—Are the positions of piRNA clusters also rapidly changing? Since the information about the position of piRNA clusters is epigenetically transmitted (see above), fluctuations in the population of maternally transmitted piRNAs could lead to changes in the size and position of piRNA clusters. In agreement with this, a recent work suggests that many clusters in *Drosophila* are solely found in a single species (Gebert et al., 2021). The turnover of the location of piRNA clusters within and among species is an important open question for future research.

Another important question is whether the observed rapid turnover of piRNA clusters is in conflict with the prevailing view on how TE invasions are stopped: the trap model holds that an invading TE is stopped when a copy of the TE jumps into a piRNA cluster (Bergman et al., 2006, Malone & Hannon, 2009, Zanni et al., 2013, Ozata et al., 2019). For the trap model to work, it is crucial that the trap (i.e. the piRNA clusters) has a minimum size of about 0.2%–3% of the genome (Kofler, 2020). The number and genomic location of the piRNA clusters has little impact (Kofler, 2019) (except if an organism has a single piRNA cluster in non-recombining regions). As long as piRNA clusters account for at least 0.2%–3% of a genome, as is likely that case in *D. melanogaster* and *D. simulans*, we do not think that the rapid turnover of piRNA clusters is in conflict with the trap model.

Finally, our work raises the question as to the consequences of rapid evolution of the composition and possibly also location of the loci responsible for silencing TEs. One consequence of such a high turnover is that silencing of TEs may be evolutionary unstable since some individuals in a population may end up without a cluster insertion for a given TE family. A high turnover of piRNA-producing loci could thus explain the low level of activity observed for many TE families in *Drosophila* Nuzhdin, 1999, since the TE will be active in the individuals that do not produce cognate piRNAs. It is however

also possible that silencing of TEs is maintained by a large number of dispersed TE insertions that are not part of piRNA cluster but nevertheless generate piRNAs (Gebert et al., 2021, Mohn et al., 2014, Shpiz et al., 2014). In agreement with this, deletion of large piRNA clusters in *D. melanogaster* did not lead to an upregulation of TEs, likely due to a large number of dispersed piRNA-producing TE insertion (Gebert et al., 2021). If silencing against a TE is effectively based on a large and redundant number of loci, then the rapid turnover of the clusters may not lead to destabilization of the silencing of a TE, which implies that piRNA clusters may largely evolve neutrally.

ACKNOWLEDGEMENTS

We thank all members of the Institute of Population Genetics for feedback and support. This work was supported by the Austrian Science Fund (FWF) grant P30036-B25 to RK and by the National Science Foundation Established Program to Stimulate Competitive Research (NSF-EPSCoR-1826834), the North Dakota EPSCoR STEM grants programme and NSF-EPSCoR-2032756 to SS.

AUTHOR CONTRIBUTIONS

FW, RK and SS conceived this work. SS assembled the two *D. simulans* strains. RK developed Manna. FW, RK and SS analysed the data. FW, RK and SS wrote the manuscript.

OPEN RESEARCH BADGES

This article has earned an Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available at https://sourceforge.net/projects/manna/files/publicationdata/.

DATA AVAILABILITY STATEMENT

The long reads of the two *D. simulans* strains are available at NCBI (PRJNA736739; PRJNA736415). The genome assemblies are available at https://sourceforge.net/projects/manna/files/assemblies/. The novel software for a multiple alignments of annotations, Manna, is available at https://sourceforge.net/projects/manna/. A manual and the validations are available at https://sourceforge.net/p/manna/wiki/Home/. The TE library and list of TE names used in this work are available at https://sourceforge.net/projects/manna/files/pirnaclustercomparison/resources/. All scripts used in this work are available at https://sourceforge.net/projects/manna/files/publicationdata/.

ORCID

Filip Wierzbicki https://orcid.org/0000-0002-6171-2461
Robert Kofler https://orcid.org/0000-0001-9960-7248
Sarah Signor https://orcid.org/0000-0003-2401-0644

REFERENCES

Adrion, J. R., Song, M. J., Schrider, D. R., Hahn, M. W., & Schaack, S. (2017). Genome-wide estimates of transposable element insertion and deletion rates in Drosophila melanogaster. *Genome Biology and Evolution*, *9*(5), 1329–1340. https://doi.org/10.1093/gbe/evx050

Ajioka, J. W., & Eanes, W. F. (1989). The accumulation of p-elements on the tip of the x chromosome in populations of Drosophila melanogaster. *Genetics Research*, *53*(1), 1–6.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andersen, P. R., Tirian, L., Vunjak, M., & Brennecke, J. (2017). A heterochromatin-dependent transcription machinery drives piRNA expression. *Nature*, *549*(7670), 54–59. https://doi.org/10.1038/nature23482

Anxolabéhère, D., Kidwell, M. G., & Periquet, G. (1988). Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. *Molecular Biology and Evolution*, *5*(3), 252–269.

Asif-Laidin, A., Delmarre, V., Laurentie, J., Miller, W. J., Ronsseray, S., & Teysset, L. (2017). Short and long-term evolutionary dynamics of subtelomeric piRNA clusters in Drosophila. *DNA Research*, *24*(5), 459–472. https://doi.org/10.1093/dnares/dsx017

Assis, R., & Kondrashov, A. S. (2009). Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proceedings of the National Academy of Sciences*, *106*(17), 7079–7082. https://doi.org/10.1073/pnas.0900523106

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Barrón, M. G., Fiston-Lavier, A.-S., Petrov, D. A., & González, J. (2014). Population genomics of transposable elements in Drosophila. *Annual Review of Genetics*, *48*(1), 561–581.

Bergman, C. M., Quesneville, H., Anxolabèhéere, D., & Ashburner, M. (2006). Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biology*, *7*(11), R112.

Blumenstiel, J. P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. *Trends in Genetics*, *27*(1), 23–31. https://doi.org/10.1016/j.tig.2010.10.003

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, *128*(6), 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043

Brizuela, B. J., Elfring, L., Ballard, J., Tamkun, J. W., & Kennison, J. A. (1994). Genetic analysis of the brahma gene of Drosophila melanogaster and polytene chromosome subdivisions 72AB. *Genetics*, *137*(3), 803–813. https://doi.org/10.1093/genetics/137.3.803

Casacuberta, E., & González, J. (2013). The impact of transposable elements in environmental adaptation. *Molecular Ecology*, *22*(6), 1503–1517. https://doi.org/10.1111/mec.12170

Chakraborty, M., Chang, C. H., Khost, D. E., Vedanayagam, J., Adrion, J. R., Liao, Y., Montooth, K., Meiklejohn, C. D., Larracuente, A. M., & Emerson, J. J. (2021). Evolution of genome structure in the Drosophila simulans species complex. *Genome Research*, *31*, 380–396.

Chakraborty, M., Vankuren, N. W., Zhao, R., Zhang, X., Kalsow, S., & Emerson, J. J. (2018). Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nature Genetics*, *50*(1), 20–25. https://doi.org/10.1038/s41588-017-0010-y

Charlesworth, B., & Charlesworth, D. (1983). The population dynamics of transposable elements. *Genetics Research*, *42*(1), 1–27. https://doi.org/10.1017/S0016672300021455

Chen, P., & Aravin, A. A. (2021). Transposon-taming pirnas in the germline: Where do they come from? *Molecular Cell*, *81*(19), 3884–3885. https://doi.org/10.1016/j.molcel.2021.09.017

Chirn, G. W., Rahman, R., Sytnikova, Y. A., Matts, J. A., Zeng, M., Gerlach, D., Yu, M., Berger, B., Naramura, M., Kile, B. T., & Lau, N. C. (2015).

Conserved piRNA expression from a distinct set of piRNA cluster loci in Eutherian mammals. *PLoS Genetics*, *11*(11), e1005652. https://doi.org/10.1371/journal.pgen.1005652

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., Perrimon, N., Kellis, M., Wohlschlegel, J. A., Sachidanandam, R., Hannon, G. J., & Brennecke, J. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature*, *453*(7196), 798–802. https://doi.org/10.1038/nature07007

Czech, B., Munafò, M., Ciabrelli, F., Eastwood, E. L., Fabry, M. H., Kneuss, E., & Hannon, G. J. (2018). piRNA guided genome defense: From biogenesis to silencing. *Annual Review of Genetics*, *52*(1), 131–157.

Daborn, P. J., Yen, J. L., Bogwitz, M. R., Le Goff, G., Feil, E., Jeffers, S., Tijet, N., Perry, T., Heckel, D., Batterham, P., Feyereisen, R., Wilson, T. G., & ffrench Constant, R. H. (2002). A single P450 allele associated with insecticide resistance in Drosophila. *Science*, *297*(5590), 2253–2256.

Darricarrere, N., Liu, N., Watanabe, T., & Lin, H. (2013). Function of Piwi, a nuclear Piwi/Argonaute protein, is independent of its slicer activity. *Proceedings of the National Academy of Sciences*, *110*(4), 1297–1302. https://doi.org/10.1073/pnas.1213283110

Dimitri, P., Junakovic, N., & Arcá, B. (2003). Colonization of heterochromatic genes by transposable elements in Drosophila. *Molecular Biology and Evolution*, *20*(4), 503–512. https://doi.org/10.1093/molbev/msg048

Feng, D.-F., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, *25*(4), 351–360.

Gebert, D., Neubert, L. K., Lloyd, C., Gui, J., Lehmann, R., Teixeira, F. K., Gebert, D., Neubert, L. K., Lloyd, C., Gui, J., Lehmann, R., & Teixeira, F. K. (2021). Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Molecular Cell*, *81*, 1–14. https://doi.org/10.1016/j.molcel.2021.07.011

González, J., Lenkov, K., Lipatov, M., Macpherson, J. M., & Petrov, D. A. (2008). High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *PLOS Biology*, *6*(10), e251. https://doi.org/10.1371/journal.pbio.0060251

Goriaux, C., Théron, E., Brasset, E., & Vaury, C. (2014). History of the discovery of a master locus producing piRNAs: The amenco/COM locus in Drosophila melanogaster. *Frontiers in Genetics*, *5*, 257.

Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., & Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in Drosophila. *Science*, *315*(5818), 1587–1590.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., & Svirskas, R. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Research*, *25*(3), 445–458.

Josse, T., Teysset, L., Todeschini, A.-L., Sidor, C. M., Anxolabèhéere, D., & Ronsseray, S. (2007). Telomeric transsilencing: an epigenetic repression combining RNA silencing and heterochromatin formation. *PLoS Genetics*, *3*(9), 1633–1643.

Kalmykova, A. I., Klenov, M. S., & Gvozdev, V. A. (2005). Argonaute protein PIWI controls mobilization of retrotransposons in the Drosophila male germline. *Nucleic Acids Research*, *33*(6), 2052–2059. https://doi.org/10.1093/nar/gki323

Karpen, G. H., & Spradling, A. C. (1992). Analysis of subtelomeric heterochromatin in the Drosophila minichromosome Dp1187 by single P element insertional mutagenesis. *Genetics*, *132*(3), 737–753. https://doi.org/10.1093/genetics/132.3.737

Kelleher, E. S., Azevedo, R. B. R., & Zheng, Y. (2018). The evolution of small-RNA-mediated silencing of an invading transposable element. *Genome Biology and Evolution*, *10*(11), 3038–3057. https://doi.org/10.1093/gbe/evy218

Khost, D. E., Eickbush, D. G., & Larracuente, A. M. (2017). Single-molecule sequencing resolves the detailed structure of complex satellite dna loci in Drosophila melanogaster. *Genome Research*, *27*(5), 709–721.

King, E. G., Merkes, C. M., McNeil, C. L., Hoofer, S. R., Sen, S., Broman, K. W., Long, A. D., & Macdonald, S. J. (2012). Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Research*, *22*(8), 1558–1566.

Kofler, R. (2019). Dynamics of transposable element invasions with piRNA clusters. *Molecular Biology and Evolution*, *36*(7), 1457–1472. https://doi.org/10.1093/molbev/msz079

Kofler, R. (2020). piRNA clusters need a minimum size to control transposable element invasions. *Genome Biology and Evolution*, *12*(5), 736–749. https://doi.org/10.1093/gbe/evaa064

Kofler, R., Nolte, V., & Schlötterer, C. (2015). Tempo and mode of transposable element activity in Drosophila. *PLOS Genetics*, *11*(7), e1005406. https://doi.org/10.1371/journal.pgen.1005406

Kofler, R., Senti, K.-A., Nolte, V., Tobler, R., & Schlötterer, C. (2018). Molecular dissection of a natural transposable element invasion. *Genome Research*, *28*(6), 824–835. https://doi.org/10.1101/gr.228627.117

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Lachaise, D., & Silvain, J.-F. (2004). How two afrotropical endemics made two cosmopolitan human commensals: the Drosophila Melanogaster-D. Simulans palaeogeographic riddle. *Genetica*, *120*(1), 17–39.

Langley, C. H., Montgomery, E., Hudson, R., Kaplan, N., & Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. *Genetics Research*, *52*(03), 223–235. https://doi.org/10.1017/S0016672300027695

Le Thomas, A., Marinov, G. K., & Aravin, A. A. (2014). A transgenerational process defines piRNA biogenesis in Drosophila virilis. *Cell Reports*, *8*(6), 1617–1623. https://doi.org/10.1016/j.celrep.2014.08.013

Le Thomas, A., Rogers, A. K., Webster, A., Marinov, G. K., Liao, S. E., Perkins, E. M., Hur, J. K., Aravin, A. A., & Tóth, K. F. (2013). Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes and Development*, *27*(4), 390–399. https://doi.org/10.1101/gad.209841.112

Le Thomas, A., Stuwe, E., Li, S., Du, J., Marinov, G., Rozhkov, N., Chen, Y. C. A., Luo, Y., Sachidanandam, R., Toth, K. F., Patel, D., & Aravin, A. A. (2014). Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes and Development*, *28*(15), 1667–1680. https://doi.org/10.1101/gad.245514.114

Lee, Y. C. G., & Langley, C. H. (2010). 710 Transposable elements in natural populations of Drosophila melanogaster. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1544), 1219–1228.

Lee, Y. C. G., & Langley, C. H. (2012). Long-term and short-term evolutionary impacts of transposable elements on Drosophila. *Genetics*, *192*(4), 1411–1432.

Lerat, E., Burlet, N., Biémont, C., & Vieira, C. (2011). Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene*, *473*(2), 100–109. https://doi.org/10.1016/j.gene.2010.11.009

Levis, R., O'Hare, K., & Rubin, G. M. (1984). Effects of transposable element insertions on RNA encoded by the white gene of Drosophila. *Cell*, *38*(2), 471–481. https://doi.org/10.1016/0092-8674(84)90502-6

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, *26*(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

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*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lim, J. K. (1988). Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. *Proceedings of the National Academy of Sciences*, *85*(23), 9153–9157. https://doi.org/10.1073/pnas.85.23.9153

Lu, J., & Clark, A. G. (2010). Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in Drosophila. *Genome Research*, *20*(2), 212–227.

Malone, C. D., & Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell*, *136*(4), 656–668. https://doi.org/10.1016/j.cell.2009.01.045

Malone, C. D., & Hannon, G. J. (2010). Molecular evolution of piRNA and transposon control pathways in Drosophila. *Cold Spring Harbor Symposia on Quantitative Biology*, *74*, 225–234. https://doi.org/10.1101/sqb.2009.74.052

Marin, L., Lehmann, M., Nouaud, D., Izaabel, H., Anxolabèhéere, D., & Ronsseray, S. (2000). P-element repression in Drosophila melanogaster by a naturally occurring defective telomeric P copy. *Genetics*, *155*(4), 1841–1854.

Marsano, R. M., Moschetti, R., Caggese, C., Lanave, C., Barsanti, P., & Caizzi, R. (2000). The complete Tirant transposable element in Drosophila melanogaster shows a structural relationship with retrovirus-like retrotransposons. *Gene*, *247*(1–2), 87–95. https://doi.org/10.1016/S0378-1119(00)00115-3

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

Mateo, L., Ullastres, A., & González, J. (2014). A transposable element insertion confers xenobiotic resistance in Drosophila. *PLoS Genetics*, *10*(8), e1004560. https://doi.org/10.1371/journal.pgen.1004560

McClintock, B. (1956). Controlling elements and the gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*, 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

McGinnis, W., Shermoen, A. W., & Beckendorf, S. K. (1983). A transposable element inserted just 5' to a Drosophila glue protein gene alters gene expression and chromatin structure. *Cell*, *34*(1), 75–84. https://doi.org/10.1016/0092-8674(83)90137-X

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*(6), 1364–1379. https://doi.org/10.1016/j.cell.2014.04.031

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nouhaud, P. (2018). *Long-read based assembly and annotation of a Drosophila simulans genome*. bioRxiv. https://doi.org/10.1101/425710

Nuzhdin, S. V. (1999). Sure facts, speculations, and open questions about the evolution of transposable element copy number. *Genetica*, *107*(1–3), 129–137.

Obbard, D. J., Maclennan, J., Kim, K.-W., Rambaut, A., O'Grady, P. M., & Jiggins, F. M. (2012). Estimating divergence dates and substitution rates in the Drosophila phylogeny. *Molecular Biology and Evolution*, *29*(11), 3459–3473. https://doi.org/10.1093/molbev/mss150

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 Research*, *41*(11), 5757–5768. https://doi.org/10.1093/nar/gkt310

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). PIWI-interacting RNAs: small RNAs with big functions. *Nature Reviews Genetics*, *20*(2), 89–108.

Peters, L., & Meister, G. (2007). Argonaute proteins: Mediators of RNA silencing. *Molecular Cell*, *26*(5), 611–623.

Petrov, D. A., Lozovskaya, E. R., & Hartl, D. L. (1996). High intrinsic rate of DNA loss in Drosophila. *Nature*, *384*(6607), 346–349. https://doi.org/10.1038/384346a0

Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition driven genomic expansions in Oryza australiensis, a wild relative of rice. *Genome Research*, *16*(10), 1262–1269.

Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., & Anxolabèhéere, D. (2005). Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biology*, *1*(2), 166–175. https://doi.org/10.1371/journal.pcbi.0010022

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. http://www.R-project.org

Sánchez-Gracia, A., Maside, X., & Charlesworth, B. (2005). High rate of horizontal transfer of transposable elements in Drosophila. *Trends in Genetics: TIG*, *21*(4), 200–203. https://doi.org/10.1016/j.tig.2005.02.001

Schnable, P. S., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T. A., Minx, P., Reily, A. D., Courtney, L., Kruchowski, S. S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B., Rock, S. M., Belter, E., Du, F., Kim, K., … Kumari, S. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*(5956), 1112–1115.

Schwarz, F., Wierzbicki, F., Senti, K.-A., & Kofler, R. (2021). Tirant stealthily invaded natural drosophila melanogaster populations during the last century. *Molecular Biology and Evolution*, *38*(4), 1482–1497.

Sedlazeck, F. J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., & Schatz, M. C. (2018). Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods*, *15*(6), 461–468. ISSN 15487105. https://doi.org/10.1038/s41592-018-0001-7

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 Genetics*, *10*(2), e1004138. https://doi.org/10.1371/journal.pgen.1004138

Sienski, G., Dönertas, D., & Brennecke, J. (2012). Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell*, *151*(5), 964–980. https://doi.org/10.1016/j.cell.2012.10.040

Signor, S. (2020). Transposable elements in individual genotypes of Drosophila simulans. *Ecology and Evolution*, *10*(7), 3402–3412.

Signor, S. A., Abbasi, M., Marjoram, P., & Nuzhdin, S. V. (2017). Conservation of social effects (ψ) between two 803 species of Drosophila despite reversal of sexual dimorphism. *Ecology and Evolution*, *7*(23), 10031–10041.

Signor, S. A., New, F. N., & Nuzhdin, S. (2017). A large panel of Drosophila simulans reveals an abundance of common variants. *Genome Biology and Evolution*, *10*(1), 189–206. https://doi.org/10.1093/gbe/evx262

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015) BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19): 3210–3212.

Smit, A. F. A., Hubley, R., & Green, P. (1996–2010). *RepeatMasker Open-3.0*. URL http://www.repeatmasker.org

Smit, A. F. A., Hubley, R., & Green, P. (2013–2015). *RepeatMasker Open-4.0*. URL http://www.repeatmasker.org

Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., Emerson, J. J., & Hawley, R. S. (2018). Rapid low-cost assembly of the Drosophila melanogaster reference genome using low-coverage, long-read sequencing. G3: Genes, Genomes. *Genetics*, *8*(10), 3143–3154.

Sullivan, M. J., Petty, N. K., & Beatson, S. A. (2011). Easyfig: a genome comparison visualizer. *Bioinformatics*, *27*(7), 1009–1010.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment

through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, *22*(22), 4673–4680.

Vaser, R., Sovic, I., Nagarajan, N., & Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. https://doi.org/10.1101/gr.214270.116

Vermaak, D., & Malik, H. S. (2009). Multiple roles for heterochromatin protein 1 genes in Drosophila. *Annual Review of Genetics*, *43*, 467–492.

Vieira, C., Fablet, M., Lerat, E., Boulesteix, M., Rebollo, R., Burlet, N., Akkouche, A., Hubert, B., Mortada, H., & Biémont, C. (2012). A comparative analysis of the amounts and dynamics of transposable elements in natural populations of Drosophila melanogaster and Drosophila simulans. *Journal of Environmental Radioactivity*, *113*, 83–86. https://doi.org/10.1016/j.jenvrad.2012.04.001

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer Nature. ISBN 978-3-319-24277-4.

Wierzbicki, F., Schwarz, F., Cannalonga, O., & Kofler, R. (2021). Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. *Molecular Ecology Resources*, 1–20. https://doi.org/10.1111/1755-0998.13455

Yang, H.-P., & Nuzhdin, S. V. (2003). Fitness costs of Doc expression are insufficient to stabilize its copy number in Drosophila melanogaster. *Molecular Biology and Evolution*, *20*(5), 800–804. https://doi.org/10.1093/molbev/msg087

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 amenco locus reflect the regulatory properties of piRNA clusters. *Proceedings of the National Academy of Sciences*, *110*(49), 19842–19847.

Zhang, S., Pointer, B., & Kelleher, E. S. (2020). Rapid evolution of piRNA-mediated silencing of an invading transposable element was driven by abundant de novo mutations. *Genome Research*, *30*(4), 566–575. https://doi.org/10.1101/gr.251546.119

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Wierzbicki, F., Kofler, R., & Signor, S. (2023). Evolutionary dynamics of piRNA clusters in *Drosophila*. *Molecular Ecology*, 32, 1306–1322. https://doi.org/10.1111/mec.16311