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

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

Affiliation:
Department of Molecular Biology and Genetics, Cornell University, Ithaca, USA

*To whom correspondence should be addressed : sps257@cornell.edu, cf458@cornell.edu & ac347@cornell.edu

Abstract

Animal genomes are parasitized by a horde of transposable elements (TEs) whose mutagenic activity can have catastrophic consequences. The piRNA pathway is a conserved mechanism to repress TE activity in the germline via a specialized class of small RNAs associated with effector Piwi proteins called piwi-associated RNAs (piRNAs). piRNAs are produced from discrete genomic regions called piRNA clusters (piCs). While piCs are generally enriched for TE sequences and the molecular processes by which they are transcribed and regulated 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 evolution, we use a population genomics approach to compare piC activity and sequence composition across 8 geographically distant strains of *D. melanogaster* with high quality long-read genome assemblies. We perform extensive annotations of ovary piCs and TE content in each strain and test predictions of two proposed models of piC evolution. The *'de novo'* model posits that individual TE insertions can spontaneously attain the status of a small piC to generate piRNAs silencing the entire TE family. The 'trap' model envisions large and evolutionary stable genomic clusters where TEs tend to accumulate and serves as a long-term "memory" of ancient TE invasions and produce a great variety of piRNAs protecting against related TEs entering the genome. It remains unclear which model best describes the evolution of piCs. Our analysis uncovers extensive variation in piC activity across strains and signatures of rapid birth and death of piCs in natural populations. Most TE families inferred to be recently or currently active show an enrichment of strain-specific

insertions into large piCs, consistent with the trap model. By contrast, only a small subset of active LTR retrotransposon families is enriched for the formation of strain-specific piCs, suggesting that these families have an inherent proclivity to form *de novo* piCs. Thus, our findings support aspects of both *'de novo'* and 'trap' models of piC evolution. We propose that these two models represent two extreme stages along an evolutionary continuum, which begins with the emergence of piCs *de novo* from a few specific LTR retrotransposon insertions that subsequently expand by accretion of other TE insertions during evolution to form larger 'trap' clusters. Our study shows that piCs are evolutionarily labile and that TEs themselves are the major force driving the formation and evolution of piCs.

Introduction

Organisms have evolved many mechanisms to minimize the genomic instability caused by transposable element (TE) activity (Bingham et al. 1982; Hedges and Deininger 2007; Huang et al. 2012; Montgomery et al. 1991). In animals, the Piwi-associated RNA (piRNA) pathway is a conserved small RNA-based mechanism regulating TE activity in the germline (Brennecke et al. 2007; Houwing et al. 2007; Lau et al. 2006; Grimson et al. 2008). piRNAs are 23-35 nucleotide RNAs produced from discrete loci called piRNA clusters (piCs) that guide Piwi effector proteins to silence TEs (Saito and Siomi 2010; Ozata et al. 2019). The piRNA pathway presents features of an adaptive defense system against TE invasion (Brennecke et al. 2008; Kofler et al. 2018; Khurana et al. 2011; Yu et al. 2019) but little is known about the processes and principles driving its evolution. The genes encoding the effector proteins and processing factors involved in piRNA-mediated silencing display signatures of adaptive evolution (positive selection) in several species' lineages (Yi et al. 2014; Simkin et al. 2013; Vermaak et al. 2005; Palmer et al. 2018), which may indicate adaptation to rapidly changing TE sequences and new invasions (Cosby et al. 2019). While attempts have been made to explain the rapid evolution of piRNA pathway genes, (Wang et al. 2020; Parhad et al. 2020; Brand and Levine 2021) little is known about how piRNA-producing loci originate and evolve in flies, or any other species.

piRNAs are produced from single-stranded long non-coding RNA precursors that are transcribed from dispersed loci called piRNA clusters (Li et al. 2013; Mohn et al. 2014; Brennecke et al. 2007). piCs make up 0.1-3% of the genome of flies, mosquitoes, and mice and are enriched for TEs and other repeats such as DNA satellites but sometimes host gene sequences as well (Chirn et al. 2015; Brennecke et al. 2007; Ma et al. 2021; Roovers et al. 2015; Chen et al. 2021). The best characterized function of piRNAs is to repress TEs. Since TE activity and composition vary significantly between and within species, TEs themselves must be important drivers of piC evolution, but this has not been thoroughly tested. TEs exhibit high diversity in their mechanisms

of transposition and genomic distribution (Sultana et al. 2017; Charlesworth et al. 1994; Bartolomé et al. 2002; Wells and Feschotte 2020). In addition, differences in the spatial and temporal activity of TE families exist in animal germlines (Laski et al. 1986; Calvi and Gelbart 1994; Bogu et al. 2019; Yoth et al. 2022; Chang et al. 2022). Hence, it is likely that piCs evolve through diverse mechanisms to repress newly introduced TEs, which is predicted to create an arms race between TEs and piCs (Cosby et al. 2019; Luo et al. 2020; Parhad and Theurkauf 2019; Said et al. 2022). Indeed, some piCs in flies are specialized to repress specific subsets of TE families, such as *flamenco*, which is almost entirely composed of and dedicated to silencing *Ty3/mdg4* (formerly known as *gypsy*) retroviral-like elements (Zanni et al. 2013). Although detailed mechanistic features of piC regulation and function have been uncovered in the last decade (Ozata et al. 2019; Czech et al. 2018), there is very limited understanding of piC evolution. Thus far, studies of piC evolution have been restricted to either a few large piCs or to small conserved genic piCs (Gebert et al. 2021; Chirn et al. 2015; Zhang et al. 2020; Ellison and Cao 2020; Mohamed et al. 2020; Wierzbicki et al. 2023).

The organization of piCs is best characterized in *Drosophila melanogaster*. The genome-wide piC landscape in *D. melanogaster* ovary comprises of tens of large (>10 kb) loci and hundreds of smaller (<10 kb) loci (Brennecke et al. 2007; Chen et al. 2021). It is also known that most large clusters (>10 kb) reside in pericentromeric and sub-telomeric regions. Larger pericentromeric clusters are composed of tens to hundreds of diverse TE insertions, while the small clusters (<10 kb) often contain recent TE insertions (Shpiz et al. 2014; Miller et al. 2023; Robine et al. 2009). The architecture and composition of some large clusters suggests a ‘trap’ model for the evolution of piCs, wherein TE insertions of active families within clusters is selectively favored because of their presumed transposition repressive effects (Zanni et al. 2013; Bergman et al. 2006; Moon et al. 2018; Zhang et al. 2020). Over time, this process is predicted to lead to the accumulation of archival remnants of past TE invasions in piCs. Thus, these large piCs are thought to produce a

bank of diverse piRNAs related to previously encountered TEs as means to silence newly introduced TEs. On the other hand, the finding that small piCs can originate from recent TE insertions suggested a *'de novo'* model. Here, individual TE insertions are converted into piCs through an epigenetic licensing process guided by maternally deposited piRNAs (le Thomas et al. 2014; Brennecke et al. 2008; Olovnikov et al. 2013). The '*de novo*' model abrogates the need for active TEs to land into existing 'trap' clusters to come under the control of the piRNA pathway (Shpiz et al. 2014; Gebert et al. 2021).

The 'trap' and '*de novo*' models of piC evolution are not mutually exclusive, but each makes contrasting predictions on how the structure, composition and activity of piCs are expected to evolve in a population. In the 'trap' model, piCs are expected to be (1) fewer in number, larger (10-100 kb), and mainly peri-centromeric and sub-telomeric; (2) piCs are expected to be syntenically stable archive of sequences of active and inactive TE families, and (3) piCs should gradually undergo TE sequence turnover via internal sequence rearrangements such as insertions and deletions (INDELs). The latter process is likely to be a fundamental characteristic of 'traps', wherein insertions of young and active TE families would replace older insertions of inactive TE families. That is because there is a limit to the genomic space afforded to clusters due to the possibility of spreading of their silent chromatin over host genes (Blumenstiel et al. 2016; Huang et al. 2022; Lee and Karpen 2017). This would also ensure that piC sequences are representative of the ever-changing genomic TE content while maintaining a constant genomic space over time (Kofler 2020; Said et al. 2022). Initial discoveries of piC architecture and composition in *D. melanogaster* were consistent with the 'trap' model (Assis and Kondrashov 2009; Brennecke et al. 2007; Malone et al. 2009). Notably, the *flamenco* cluster on the X chromosome is proposed to act as a 'trap' specialized for the capture of *Ty3*/*mdg4* retrotransposons (Genzor et al. 2021; Zanni et al. 2013). Furthermore, evolution of piRNA-mediated repression of recently invading TEs, such as the *P*-element, documented cases where silencing of *P*-elements is established by individual

insertions within large known piCs – *42AB, #40, #3* and *X-TAS* both in laboratory and natural conditions (Khurana et al. 2011; Zhang et al. 2020; Moon et al. 2018; Ronsseray et al. 1991; Srivastav et al. 2019). While these studies suggest that horizontally acquired TEs come under the control of piRNA produced by insertions from the same family inserted into existing piCs, it remains unclear whether the 'trap' concept is broadly generalizable.

The '*de novo*' model of piC evolution posits that TE silencing is driven by smaller clusters comprised of individual TE insertions of recent origin. If the '*de novo'* model is the primary mode of piC evolution, it should be expected that piCs are 1) abundant but smaller in size (1-10 kb), broadly distributed genome-wide, 2) highly polymorphic in population and typically strain-specific, and comprised mainly of insertions from active TE families, and 3) exhibit wholescale insertions and deletions, leading to loss or gain of entire clusters. These properties would ensure that piRNA production is biased toward active TE families. Several independent studies have provided observations that *de novo* formation of canonical piCs can emerge in the laboratory from artificial transgenes (Muerdter et al. 2012; de Vanssay et al. 2012) and in nature from recent TE insertions (Shpiz et al. 2014; Ryazansky et al. 2017). In summary, while both trap and *de novo* models of piC evolution have received empirical support, their relative contribution to piC evolution in *D. melanogaster* is unclear, and the mechanisms underlying the evolution of piCs remain broadly uncharacterized. In the present study, we found that there is extensive variation in sequence and activity of piCs across 8 *D. melanogaster* strains. These results lead us to develop a unified model of piC evolution that integrate components of *'de novo'* and 'trap' models supported by our findings.

100 ## Results

101 ### Extensive variation in the genomic landscape of piCs

102 To quantify piC variation in *D. melanogaster* we generate a comprehensive annotation of active
103 piCs in eight highly inbred strains. Seven of these strains are derived from natural populations of
104 distinct worldwide origins and have publicly available long-read genome assemblies (Chakraborty
105 et al. 2019). For each of these seven strains, we constructed and sequenced libraries of small
106 RNAs isolated from ovaries of two biological replicates sampled 6 months apart (**Supplementary**
107 **Fig. S1A, Table S1)** (see Methods)**.** In addition, we analyzed two ovarian small RNA libraries for
108 the reference iso-1 strain, generated as part of two independent studies (Shipz et al, 2014 and
109 Asif-Laidin et al, 2017). piCs are defined by high expression and density of 23-29 nt long and 1U-
110 biased small RNAs inferred to represent piRNAs (Brennecke et al. 2007; Mohn et al. 2014). Each
111 small RNA library is analyzed separately using the pipeline outlined briefly here and described in
112 more detail in Methods, which lists three methods to define piCs **(Fig. 1A, Supplementary Fig.**
113 **2B).** The *restrictive* and *proTRAC* methods serve the purpose of discovering moderately to highly
114 expressed piCs using uniquely and multi-mapping piRNAs respectively. The *permissive* method
115 is carried out mainly to validate low to moderately expressed piCs detected by *proTRAC* using
116 multi-mapping piRNAs.

117 The piC predictions of the *restrictive* and *proTRAC* methods were both tested for reproducibility
118 by comparing coordinates of piCs predicted independently from two replicates of each strain using
119 bedtools. To quantify inter-replicate reproducibility in piC annotations, bedtools *intersect* was used
120 with minimum required overlap of 20% of piC length for intra-strain replicates. Both *restrictive* and
121 *proTRAC* methods yielded highly reproducible piC coordinates across replicates of each strain
122 with >80% of piCs between two replicates overlapping over >75% of their respective length **(Fig.**
123 **1B,C, Supplementary Fig. 3A,B)**. However, inter-strain pairwise comparisons revealed that pairs
124 of strains share only an average of ~40% of their total piCs**.** The high confidence set of piCs from

125 *proTRAC* that either exhibited high expression or were supported by uniquely mapping piRNAs,
126 along with all piCs from the *restrictive* method for each replicate were combined to create a
127 replicate-specific 'master list' of piCs.

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

141 Genome-wide visualization of piC annotations across chromosomes reveals striking variability in
142 piC landscape across the 8 strains **(Fig. 1D)**. In aggregate, the total amount of genomic DNA
143 covered by active piCs in each strain ranged from 4.8 Mb to 6.3 Mb **(Fig. 1E)**. While their piC
144 landscape is broadly similar in terms of being denser within peri-centromeric and telomeric
145 heterochromatic regions (which are also characterized by low mappability scores) compared to
146 euchromatic regions, it is readily apparent that many individual clusters are present in only one or
147 a few strains, even within these heterochromatic regions. Smaller, euchromatic piCs are even
148 more strikingly variable across strains despite being characterized by higher mappability scores
149 **(Fig. 1F)**. Thus, from this first broad-scale view, it appears that the total amount of genomic space

150 occupied by piC within each strain is largely similar, but the positions of piCs is highly variable
151 across strains.

152 **Abundant strain-specific and strain-biased piCs**

153 To quantitatively assess the frequency of piCs across the 8 strains, we quantified the overlap of
154 piCs predicted independently for each of the 8 strains using the master list coordinates. To account
155 for changes in size of piCs among strains, we required a minimum positional overlap of only 1 bp
156 for piCs to be considered shared between strains. Even when using this non-conservative
157 criterion, we found that 568 (*restrictive*) to 906 (*proTRAC*) of piCs are active only in a single or a
158 few strains confirming that each strain has a very unique piC landscape **(Fig. 1F).** The results are
159 similar whether we used the piC predictions of the *restrictive* and *proTRAC* methods separately
160 or the combined master list **(Supplementary Table 2).** All strains exhibit 35-60 piCs that are
161 strictly unique to that strain and another ~30 piCs that could not be lifted-over and therefore are
162 likely to be strain-specific also **(Supplementary Fig. 4B)**. Thus, we can conservatively estimate
163 that each strain possesses between 50 and 100 piCs that are not shared by any of the other 7
164 strains examined. This is a conservative estimate because we only require 1 bp of overlap to
165 consider piCs to be shared, so we may be overestimating the number of shared piCs. In addition,
166 142 and 69 piCs ('*restrictive'* pipeline) are shared between two and three strains, respectively. All
167 such piCs, shared by 3 strains or less are together termed as 'rare' piCs. Rare piCs are not only
168 extremely abundant but also exhibit significant piRNA expression ranging from 20-100 RPM,
169 which is comparable to previously described canonical piCs like *38C*, *80EF* and *traffic jam* 3'UTR
170 **(Supplementary Table 3)**. Additionally, despite their small size, in aggregate, strain-specific piCs
171 contribute a substantial portion of the total genomic span of piCs (average of ~1 Mb) and 15-20%
172 of the total piC genomic length of each strain **(Supplementary Fig. 4C).**

173 Next, we examined the relationship between the size of piC and their level of sharing across
174 strains. First, we note that median piC length predicted from each library is highly similar, with

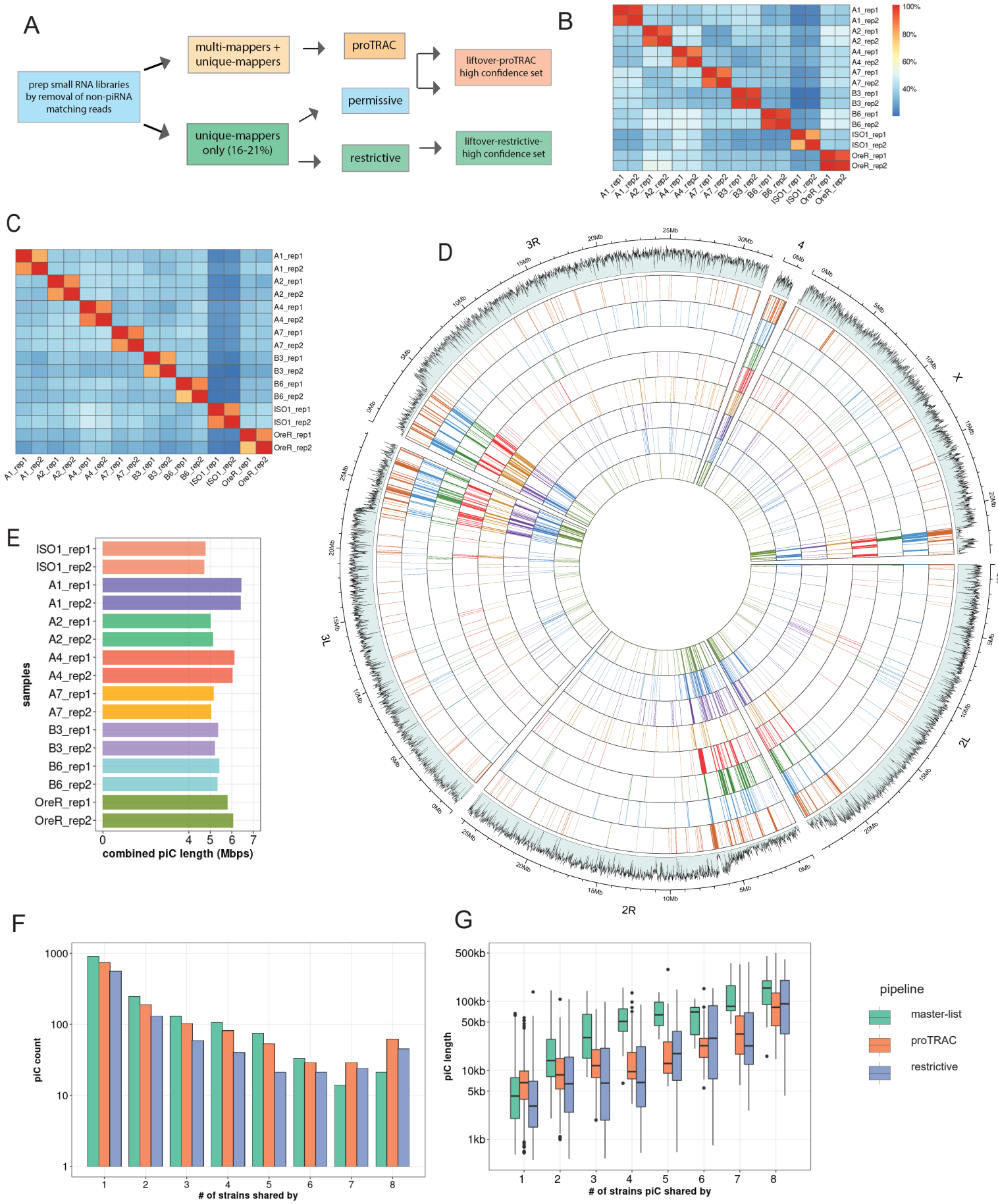
175 median length ranging from 5.7 kb to 7.5 kb **(Supplementary Fig. 4A)**. We find that piC size is
176 positively correlated with the level of sharing across strains, and this correlation holds true for all
177 prediction methods (Pearson r = 0.56 for proTRAC, 0.58 for restrictive, and 0.76 for master-list, *p-*
178 *value* < 2.2e-16) (**Fig. 1G)**. In other words, piCs detected in a single or a minority of the strains
179 (rare piCs) tend to be smaller (2-10 kb) than those shared by the majority of the strains (common
180 piCs). If we posit that rare piCs represent evolutionarily younger piCs than common piCs, this
181 relationship suggests that piCs are born relatively small and increase in size as they get older.
182 Alternatively, larger piCs may be more evolutionarily stable than smaller ones. We note, however,
183 that even large piCs can still be variable in activity across strains. For example, large well-known
184 piCs like *42AB* and *38C* are still only active in 6 to 7 of the 8 strains (see below). Taken together,
185 these results suggest that ovarian piCs are extremely labile and poorly conserved in activity across
186 *D. melanogaster* strains.

187 **Figure 1. Inter-strain variability of piCs in *D. melanogaster* strains.** (*A*) Summary of piC
188 prediction and annotation pipeline using restrictive, permissive and proTRAC pipelines. *(B)* Cross-
189 strain overlap (% of total piCs count) of independently predicted piCs for each replicate using the
190 restrictive method. *(C)* Combined genomic piC size predicted from each replicate small RNA
191 library independently in respective genome assemblies. *(D)* Genome-wide distribution of lifted-
192 over piCs in 7 DSPR strains and reference iso-1 strain. Bars along the circumference represent
193 presence of piCs in 10 kb bins for each chromosome. The outermost bar plot is piRNA mappability
194 scores, followed by iso-1 piCs, followed by piCs of 7 DSPR strains. *(E)* Combined piC length
195 predicted independently for each strain per small RNA library. *(F)* Population frequency of piCs in
196 7 DSPR strains and the reference iso-1 strain quantified after liftover to reference genome. *(G)*
197 piC length distribution by population frequency in kilobase-pairs (kb) quantified after liftover to
198 reference genome.

199

Figure 1

Srivastav et al, 2023



200 **Extensive variability in piRNA expression of piCs**

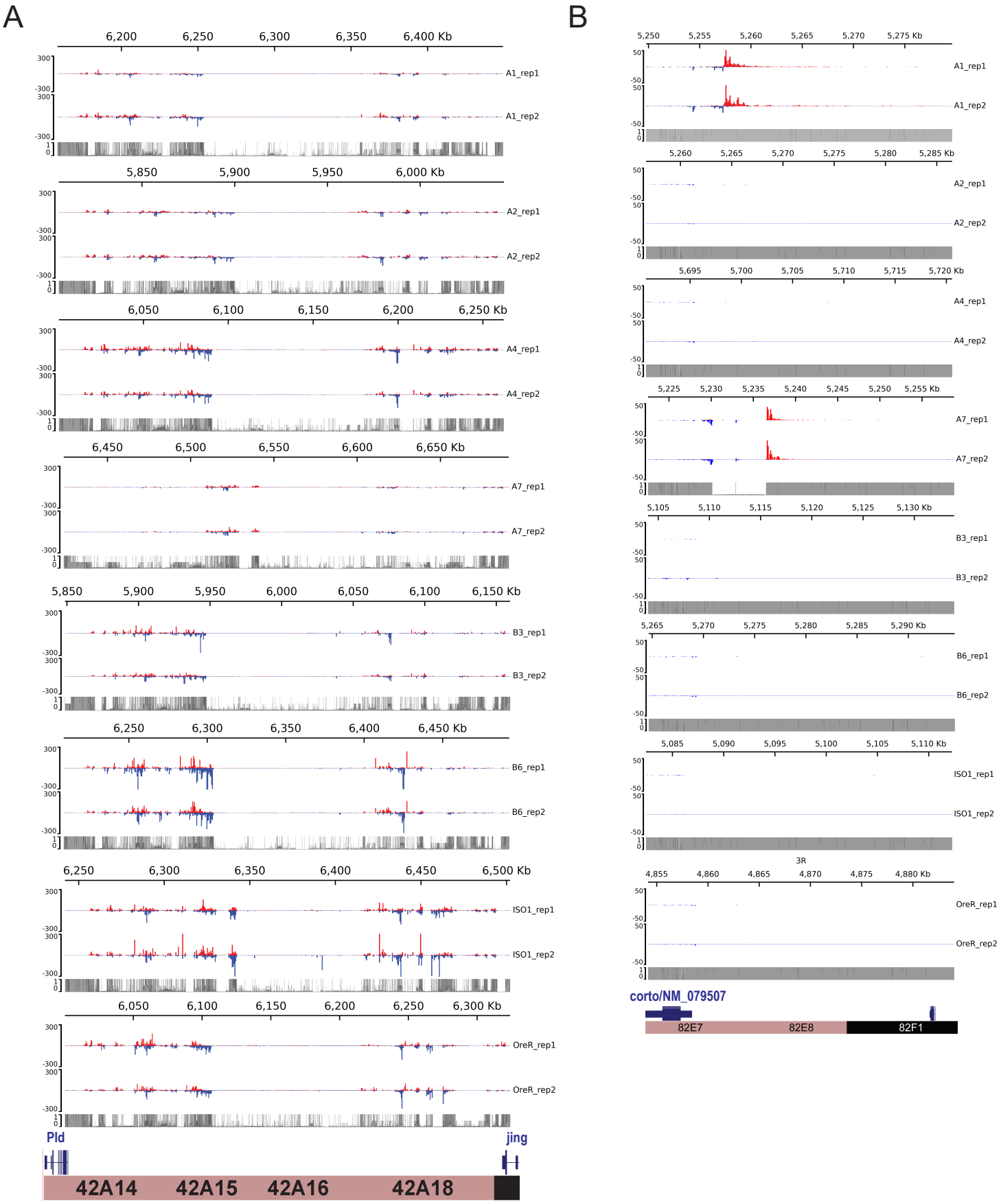
201 To illustrate the differences in activity of piCs among the 8 strains, we examine the piRNA
202 coverage profiles for *42AB* and *82E*, two dual-stranded piCs **(Fig 2A,B)**. The *42AB* cluster has
203 been extensively documented for its high piRNA expression (Brennecke et al. 2007; Klattenhoff
204 et al. 2009). We present normalized coverage of uniquely mapping piRNAs to the respective *42AB*
205 assemblies for the 8 strains using annotated flanking genes *Pld* and *jing* from both small RNA
206 library replicates. Additionally, to examine differences in read coverage due to mappability,
207 theoretical mappability scores are visualized along the length of the cluster in 100 bp bins **(Fig.**
208 **2A)**. Strains A1 and A7 have severely reduced (>20-fold) piRNA expression levels throughout
209 *42AB* compared to the other strains. Similarly, *38C* – a highly productive dual-stranded piC in iso-
210 1, exhibits significant variability in uniquely-mapping piRNAs across strains **(Supplementary Fig.**
211 **S6)**. Since *42AB* and 38C are active in 6 out of 8 strains, it is most parsimonious to conclude that
212 these piCs are relatively old but have lost activity in a subset of strains.

213 The *82E* cluster is a smaller piC (~25 kb) we selected because it is also highly expressed in some
214 strains but is euchromatic, overlapping the 5' UTR of the *corto* gene. It is only identified as a piC
215 in strains A1 and A7 in this study using the *restrictive* pipeline and piC expression profiles clearly
216 show that piRNA expression across the *82E* region is only detectable in these two strains **(Fig.**
217 **2B).** Other the other hand, *82E* has no detectable expression in 6 of 8 strains. Additionally, we
218 show the average of normalized piRNA coverage from two replicates for the 8 strains in respective
219 genome assemblies using the DrosOmics browser (Coronado-Zamora et al. 2023). Examples of
220 variability in normalized piRNA expression across strains for piCs is shown. *80EF* (left) and *80F9*
221 (right) are two peri-centromeric piCs on chr3L (**Supplementary Fig. 5A)**. *80EF*, a previously
222 described Rhino-dependent piC, is a common piC, detected across all strains with significant
223 piRNA production. *80F9*, however, is a less common piC with extremely variable piRNA coverage
224 (50-fold) between strains and is annotated as a piC in only 6 strains. piRNA coverage of *Trypsin*

225 genes-associated piC (left) and *eEF1alpha1* associated piC (right) in **Supplementary Fig. 5B**
226 also are consistent with their detection by annotation pipelines in the respective strains.
227 Comparison of such syntenic piCs between strains in their native genome assemblies provide
228 validation of the variable activity of piCs across strains presented earlier from annotation pipelines
229 **(Fig. 1G).**

230 **Figure 2. Natural variation in expression of uniquely mapping piRNAs from *42AB* and**
231 ***82EF*.** *(A)* Uniquely mapping piRNA expression profiles of *42AB* piCs for the 8 strains with two
232 small RNA library replicates. Expression values are in reads per million (RPM) for 100 bp bins.
233 Mappability scores (0-1) is shown for 100 bp bins of each respective *42AB* genomic assembly.
234 *(B)* Uniquely mapping piRNA expression profile of *82EF* piCs for the 8 strains with two small
235 RNA library replicates. Expression values are in reads per million (RPM) for 100 bp bins.
236 Mappability scores (0-1) is shown for 100 bp bins of each respective *42AB* genomic assembly.
237
238
239
240
241
242
243
244
245
246
247
248

Figure 2


A
6,200 6,250 6,300 6,350 6,400 Kb
A1_rep1
A1_rep2
5,850 5,900 5,950 6,000 Kb
A2_rep1
A2_rep2
6,050 6,100 6,150 6,200 6,250 Kb
A4_rep1
A4_rep2
6,450 6,500 6,550 6,600 6,650 Kb
A7_rep1
A7_rep2
5,850 5,900 5,950 6,000 6,050 6,100 6,150 Kb
B3_rep1
B3_rep2
6,250 6,300 6,350 6,400 6,450 Kb
B6_rep1
B6_rep2
6,250 6,300 6,350 6,400 6,450 6,500 Kb
ISO1_rep1
ISO1_rep2
6,050 6,100 6,150 6,200 6,250 6,300 Kb
OreR_rep1
OreR_rep2
Pld
jing
42A14
42A15
42A16
42A18
B
5,250 5,255 5,260 5,265 5,270 5,275 Kb
A1_rep1
A1_rep2
5,260 5,265 5,270 5,275 5,280 5,285 Kb
A2_rep1
A2_rep2
5,695 5,700 5,705 5,710 5,715 5,720 Kb
A4_rep1
A4_rep2
5,225 5,230 5,235 5,240 5,245 5,250 5,255 Kb
A7_rep1
A7_rep2
5,105 5,110 5,115 5,120 5,125 5,130 Kb
B3_rep1
B3_rep2
5,265 5,270 5,275 5,280 5,285 5,290 Kb
B6_rep1
B6_rep2
5,085 5,090 5,095 5,100 5,105 5,110 Kb
ISO1_rep1
ISO1_rep2
3R
4,855 4,860 4,865 4,870 4,875 4,880 Kb
OreR_rep1
OreR_rep2
corto/NM_079507
82E7
82E8
82F1

249 **Structural variation in piCs supports both 'trap*'* and '*de novo'* models.**

250 To better understand the mutational processes underlying the divergence of piCs among strains,
251 we examined the contribution of inter-strain structural variants (SVs), namely insertions and
252 deletions (INDELs), within the piC genomic regions. INDELs are predicted to affect piCs differently
253 under the 'trap*'* and *'de novo'* models of piC evolution. Under the 'trap' model, we expect that the
254 rate of insertions, likely representing the 'trapping' of TE insertions within common piCs would be
255 balanced by that of deletions as to stabilize the size of the piCs. Thus, we expect a similar
256 frequency of insertions and deletions within large piCs acting as traps. By contrast, we predict that
257 clusters born 'de novo' would be dominated by insertions, likely corresponding to recent
258 transposition events. To systematically detect SVs genome-wide for each of the strains relative to
259 the iso-1 reference strain (Chakraborty et al. 2019; Solares et al. 2018), we mapped raw long
260 sequencing reads for each strain to the iso-1 genome and called SVs using three independent SV
261 callers (See Methods). SVs were then genotyped, filtered, and retained for downstream analyses
262 if supported by at least two callers. SVs from all strains were then collapsed to construct a list of
263 unique SVs that consisted of 2274 insertions and 4409 deletions relative to the reference genome.
264 INDELs were then evolutionarily polarized into insertions vs. deletions by comparison of each
265 variant to *D. simulans* and *D. sechellia* reference strains, which enable inference of the ancestral
266 state (see methods). Polarization led to loss of ~55% of INDEL calls as the ancestral or derived
267 state of the loci could not be determined due to conflicts in calls between the two outgroup species.
268 After this filtering, 1183 insertions and 1873 deletions were retained for analysis, of which 30% of
269 insertions and 20% of deletions overlapped with master-list piCs **(Fig. 3A)**.

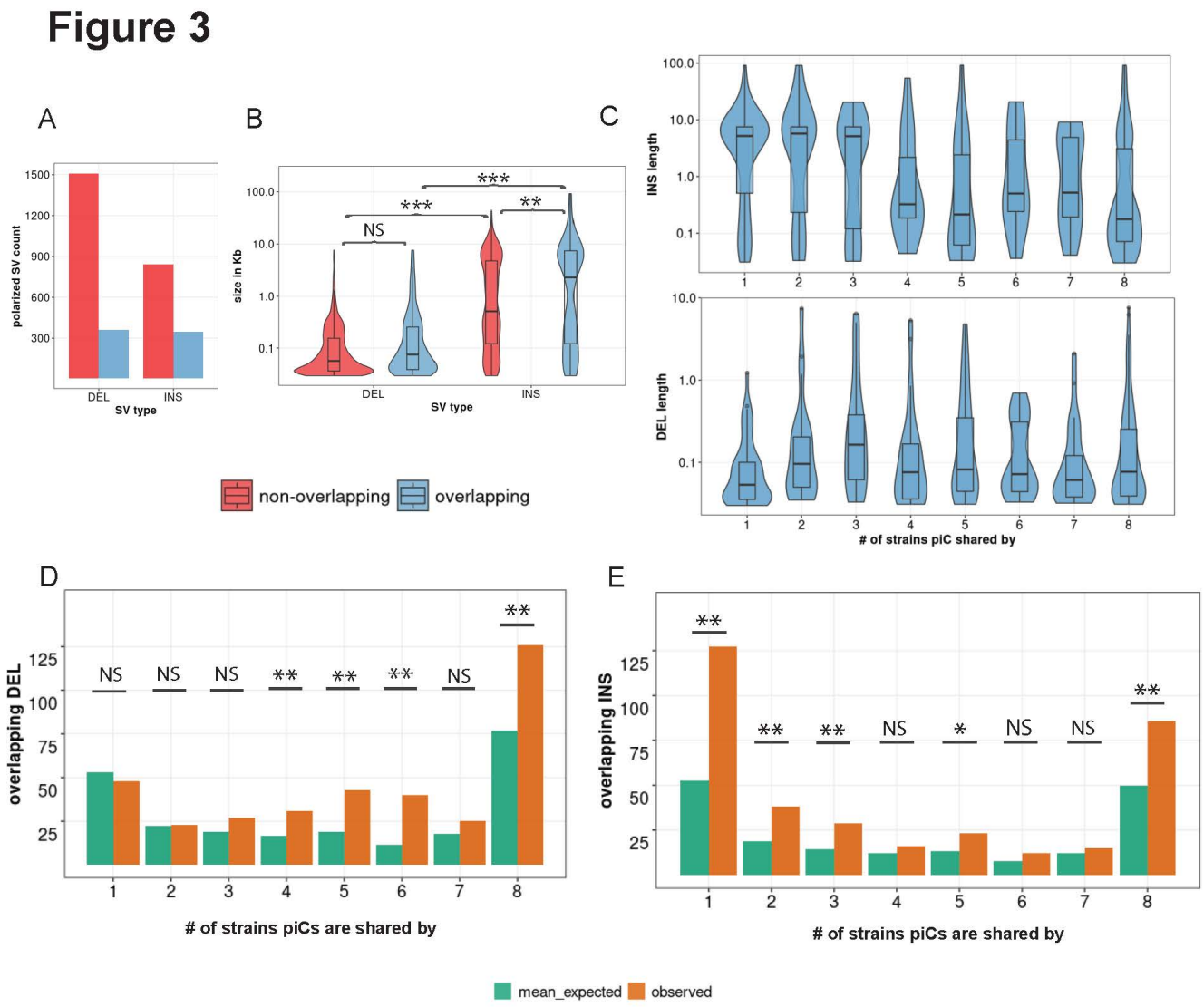
270 We examined the size distribution of INDELs overlapping piC and non-piC regions of the genome
271 to then test the predictions for INDELs associated with piC variation. First, genome-wide,
272 insertions range from 30 bp to 91 kb with a median of 612 bp, while deletions range from 30 bp to
273 7.6 kb with a lower median of 208 bp compared to insertions, which is consistent with previous

274 structural variant profiling of *D. melanogaster* strains (Dopman and Hartl 2007; Zichner et al. 2013;
275 Huang et al. 2014). However, insertions overlapping piCs have a median length of 2.2 kb, whereas
276 insertions non-overlapping piCs have a smaller median length of 512 bp (Kruskal-Wallis test,
277 χ^2=10.812, df=1, *p*-value = 0.001). Meanwhile, deletions overlapping piCs have a similar length
278 distribution than those non-overlapping piCs (Kruskal-Wallis test, χ^2=0.72404, df= 1, *p*-value =
279 0.39) **(Fig. 3B)**. We also compared the length distribution of INDELs in piCs grouped by the
280 number of strains with which they are shared. We found that strain-specific or rare piCs (shared
281 by less than 4 strains) are associated with relatively large insertions (median length of 5.2 kb),
282 whereas common piCs (shared by more than half of the strains) have a median insertion length
283 of less than 1 kb **(Fig. 3C)**. However, the length distribution of deletions was similar between rare
284 and common piCs. In sum, rare piCs are uniquely associated with relatively large insertions, which
285 is consistent with the idea that these piCs emerged *de novo* from recent TE insertions.

286 Next, we tested whether piCs are enriched for INDELs relative to the rest of the genome. To do
287 this, we compared the INDEL counts overlapping piCs for each of the cluster frequency categories
288 with those expected based on 1000 sets of randomly shuffled INDELs. We found that deletions
289 were significantly enriched in common piCs, but not in rare piCs **(Fig. 3D)**. Insertions were strongly
290 enriched both in rare and common piCs **(Fig. 3E)**. These results may be confounded by the
291 location of piCs within constitutive heterochromatin, where the rate of SVs is generally high
292 (Chakraborty et al. 2021; Montgomery et al. 1991). However, we found that only ~28% of all piCs
293 lie within constitutive heterochromatin boundaries of the reference genome assembly and INDELs
294 were significantly enriched in piCs even when we compared them to heterochromatic regions
295 **(Supplementary Fig. S7)** (Riddle et al. 2011).Thus, the SV enrichment we observe within
296 common piCs is unlikely to be solely driven by their location within constitutive heterochromatin.
297 Overall, we conclude that piCs are subject to a high rate of structural genomic change relative to
298 the rest of the genome, which likely contributes to their rapid evolutionary turnover. Additionally,

we found that common piCs are enriched for both insertions and deletions, which is consistent with these clusters evolving as 'traps'. By contrast, rare piCs are only enriched for insertions, which supports the notion that these are generally young clusters born '*de novo*' from recent TE insertions.

Figure 3. Common piCs exhibit 'trap' like sequence turnover. *(A)* Observed counts of INDELs overlapping and non-overlapping with piCs. *(B)* Length distribution of INDELs overlapping and non-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 carried out using poverlap. Variants overlapping with piCs of differing population frequencies is compared to expected mean overlap counts genome wide. (p -value <0.05=*, <0.005=**)



328 **TE re-annotation of each strain uncovers ~3 Mb of unannotated TE DNA.**

329 While our analysis of SVs within piCs supports that these events are important drivers of piC
330 evolution, it does not directly address the role of TE activity. To assess the contribution of TEs to
331 the composition and changes in the activity of piCs across strains, we carried out *de novo*
332 annotation of TEs in each of the strain genome assemblies. This was necessary because many
333 TE consensus sequences present in the reference TE library for *D. melanogaster* (FlyBase
334 release 2019_05) were discovered and curated more two decades ago using primarily the iso-1
335 and Oregon-R strains (Bartolomé et al. 2002; Kaminker et al. 2002; Bowen and McDonald 2001).
336 However, recent advances in long-read sequencing technology have provided a means to obtain
337 a more unbiased view of the repetitive landscape of *Drosophila* genomes, revealing novel TE
338 families (Ellison and Cao 2020; Rech et al. 2022; Han et al. 2022). We developed a TE annotation
339 pipeline based on RepeatModeler2 (for discovery) (Flynn et al. 2020; Smit 1999), RepeatMasker
340 (for annotation) and additional tools to distinguish novel TEs from known TEs and curate a
341 comprehensive TE library for the 8 strains used in this study **(Fig. 4A,** See methods**)**.

342 TE family sequence assemblies by RepeatModeler2 for each strain were aligned to the reference
343 TEs using the 80-80-80 rule (Wicker et al. 2007). RepeatModeler2 sequences already present in
344 the reference TE library were removed. Next, the remaining RepeatModeler2 sequences were
345 used to re-mask the genomes to examine them for novel TE families. Over 5000 insertions (>500
346 bp in size, median of 676 bp), very similar (<5% divergence) to their respective RepeatModeler2
347 family consensus are discovered for each strain (example of strain B6 in **Fig. 4B**). These novel
348 insertions resulted in masking of an additional 2.5 Mb to 4 Mb in each genome assembly that
349 would have been missed or mis-annotated as highly diverged insertions by masking with only the
350 reference TE library **(Fig. 4C)**. In summary, a refined and comprehensive TE library was created
351 with a combination of 129 reference TE consensus sequences and 45 uncharacterized consensus
352 sequences that capture all TE insertions genome-wide and reflect their relative age.

353 **piC TE composition is represented by younger LTR insertions than any other TE subclass.**

354 Using the new TE library described above, we sought to compare the age and composition of TEs
355 within piCs to that of the rest of the genome. To infer the age of each family, we use the median
356 sequence divergence of each insertion to their family consensus. To examine TE composition, we
357 grouped TEs into the major subclasses and superfamilies represented in *Drosophila*: non-LTR
358 retrotransposons (LINE), LTR retrotransposons (*Ty1*/*copia*; *Ty3*/*mdg4*; *BEL*/*Pao* superfamilies),
359 Rolling Circle (RC) transposons, and cut-and-paste DNA transposons. We found that TE copies
360 from all three LTR superfamilies were significantly younger in piCs than non-piC regions **(Fig. 4D)**.
361 Conversely, TE copies from LINE, RC and DNA subclasses were not significantly different in age
362 in piCs than in non-piC regions. To corroborate these results using an independent method to
363 date insertions, we built phylogenetic trees from all copies for one LTR superfamily (*Ty1*/*copia*)
364 and one DNA transposon superfamily (*Tc1*/*mariner*) and used terminal branch lengths to estimate
365 their relative age (Carr et al. 2012). We chose these superfamilies because they were of moderate
366 abundance and therefore manageable for multiple sequence alignments and phylogenetic
367 analyses. The results of these analyses yielded the same trend observed genome-wide using
368 sequence divergence from consensus sequences whereby the *Ty1*/*copia* LTR retrotransposons
369 (n=135) overlapping piCs are significantly younger than non-overlapping ones, while *Tc1*/*mariner*
370 (n=89) DNA transposons show no such bias **(Fig. 4E)**.

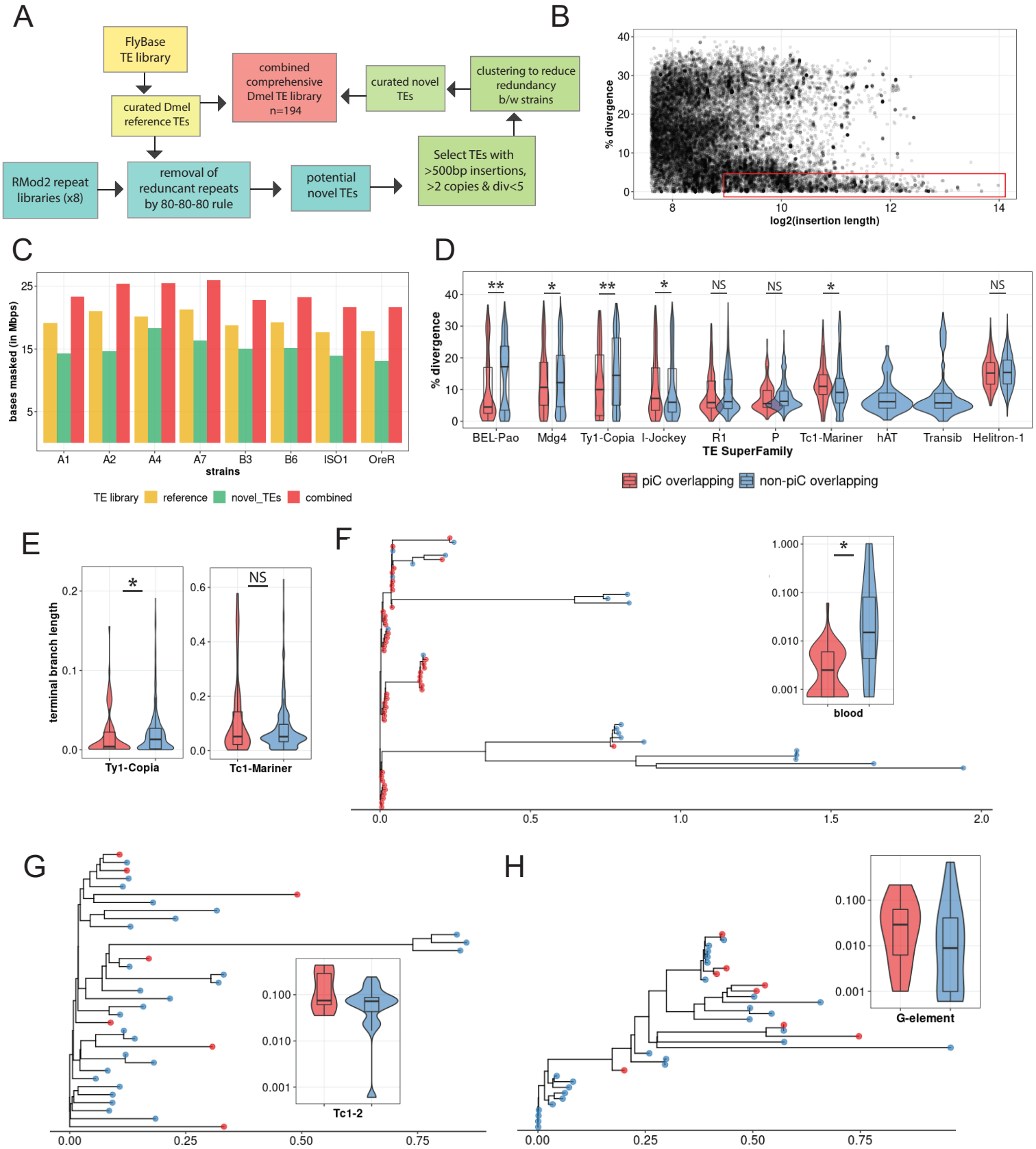
371 To examine whether these trends hold at the level of individual TE families, we selected one family
372 with moderate copy number from the LTR, LINE and DNA subclass and compared the age of piC-
373 overlapping and non-overlapping copies within each family. As a representative *Ty3*/*mdg4* LTR
374 superfamily, we analyzed *blood*, a family with 63 copies in the iso-1 strain that is known to be
375 transpositionally active (Bingham and Chapman 1986; Kofler et al. 2015). Consistent with the
376 trend observed at the level of the LTR superfamily, we found that 43 out of 63 *blood* insertions are
377 associated with piCs. Most of these are very recent insertions with median terminal branch length

378 of <0.002, which is significantly shorter than of insertions not overlapping piCs (Wilcoxon rank sum
379 test, *p*-value = 0.014) **(Fig. 4F)**. In other words, piC overlapping *blood* insertions are significantly
380 younger than the non-overlapping ones. As a representative of the *Tc1/mariner* superfamily of
381 DNA transposons, we analyzed *Tc1*-2, a family with 35 copies in the iso-1 genome. Consistent
382 with the trend observed at the level of the entire superfamily, the age of *Tc1-2* copies overlapping
383 piC is not significantly different than that of non-piC overlapping copies (Wilcoxon rank sum test,
384 *p*-value = 0.903) **(Fig. 4G)**. Analyzing the *G*-element LINE family, which counts 35 copies in iso-
385 1 and is still active (di Nocera et al. 1986), we found that the age of piC-overlapping copies is not
386 significantly different from non-overlapping copies (Wilcoxon rank sum test, *p*-value 0.855) and
387 the youngest *G*-element insertions according to terminal branch length do not overlap piCs **(Fig.
388 4H)**. Taken together, these results suggest that young LTR retrotransposon insertions tend to be
389 enriched in piCs, but this trend is not observed for other TE subclasses and superfamilies.

390 **Figure 4. *de novo* TE annotation uncovers ~3Mb of hidden TEs and reveals strong
391 associations of young LTR TEs with piCs than any other TE subclasses.** *(A)* TE annotation
392 pipeline using RepeatModeler2 and RepeatMasker to create the comprehensive TE library**.** *(B)*
393 Abundance of extremely similar and long TE insertions from RepeatMasker output of strain B6
394 using novel TE consensus library. *(C)* Differences in million base-pairs (Mbps) masked in
395 RepeatMasker results using novel-only, reference-only, and combined TE library. *(D)* Divergence
396 estimates for all defragmented iso-1 insertions (>250 bp) from RepeatMasker output. Insertions
397 with >1 bp overlap with master-list iso-1 piCs were considered piC overlapping. Difference
398 between groups is tested by Wilcox ranked-sum test. *(E)* Terminal branch length for all iso-1
399 insertions from *Ty1*/*Copia* and *Tc1*/*Mariner* superfamilies from maximum likelihood trees. *(F-H)*
400 Maximum likelihood trees constructed from all defragmented insertions for *blood*, *Tc1-2*, and *G-*
401 *element* families and the inset shows terminal branch length quantification. Difference between
402 groups is tested by Wilcox ranked-sum test; p-value <0.05=*, <0.005= **, <0.0005=***.

Figure 4

Srivastav et al, 2023



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

404 To examine the role of recent transposition events in driving piC sequence composition, we
405 established a set of non-reference TE insertions (absent from the reference genome) in each of
406 the 8 strains using the raw long read data available for each. Briefly, we applied TLDR (a long-
407 read TE insertion detection tool) (Ewing et al. 2020) with a cut-off of at least 2 supporting reads
408 per 10X genome coverage to remove false positives and enrich for germline insertions (see
409 Methods). Using these parameters, we identified 285 to 857 non-reference TE insertions for each
410 of the 7 DSPR lines but only 75 insertions for iso-1, which is expected since the reference genome
411 is also derived from the iso-1 strain. Presumably, the 75 non-reference insertions for iso-1 reflect
412 the use of different isolates for the reference genome assembly and for the long read data. Further
413 clustering and parsing of all non-reference insertions across the 8 strains resulted in a list of 3545
414 unique TE insertions of at least 200 bp in length. These insertions belong to 165 of the 184 different
415 families in our TE library. Ninety-four of these 165 TE families were classified as "active" when
416 they included at least 5 non-reference insertions shared by no more than 2 strains, while the other
417 98 families were classified as "inactive" **(Fig. 5A).**

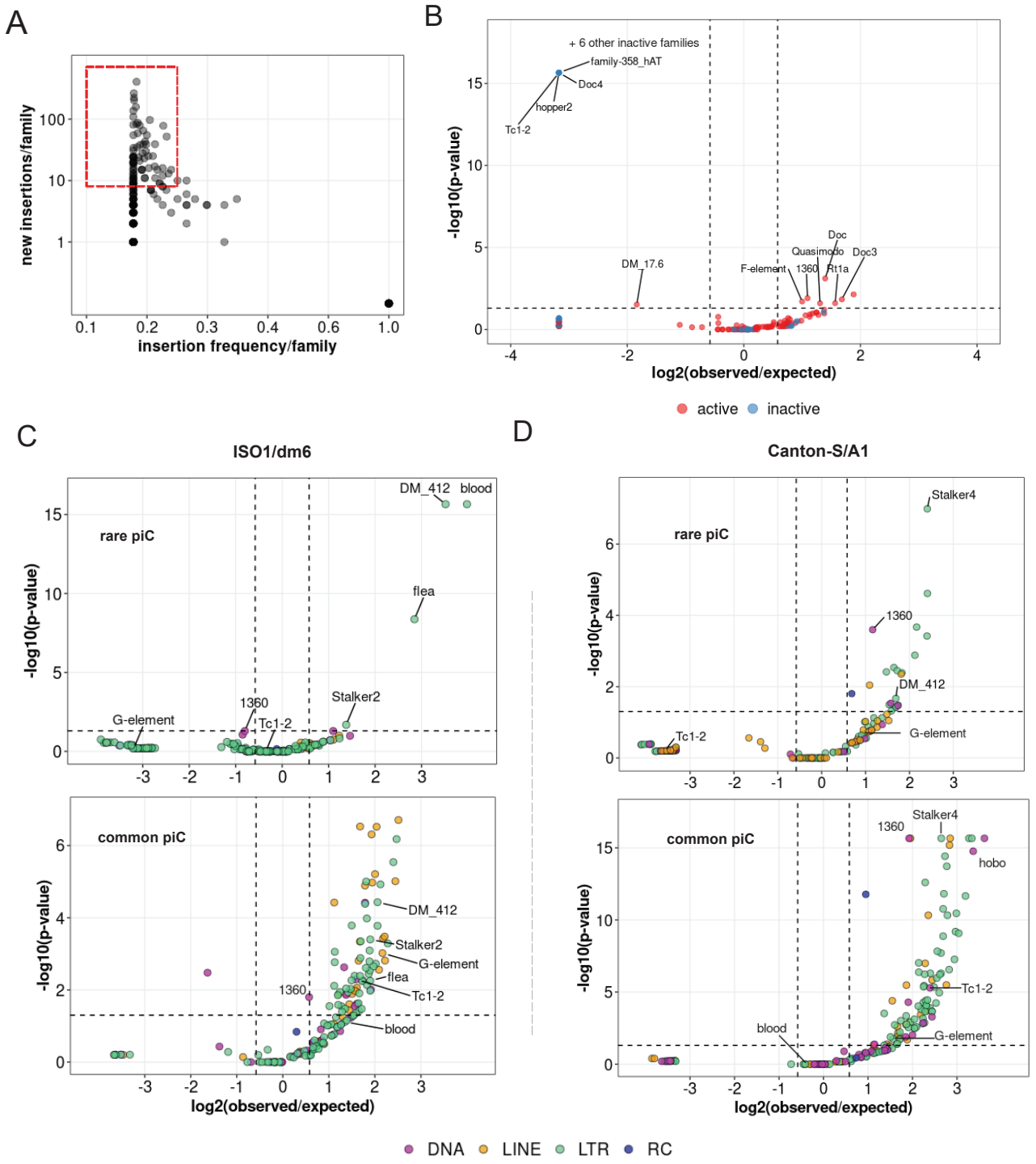
418 We used this compendium of insertions to test whether TE families are significantly enriched or
419 depleted within piCs based on their activity, using a binomial test to compare the observed
420 overlaps with piCs to the average overlaps expected from 1000 random reshufflings of piC
421 coordinates (see Methods) (Kapusta et al. 2013). This analysis revealed that 7 active TE families
422 are significantly enriched in piCs, while 1 active and 10 inactive families are significantly depleted
423 in piCs **(Fig. 5B).** All the inactive TE families that are significantly depleted in piCs belong to either
424 DNA or LINE subclasses, while the only active family depleted in piCs was 17.6, a *Ty1/copia*
425 superfamily member (Inouye et al. 1986) with 79 non-reference insertions. These results are
426 consistent with the prediction of the 'trap' of model that piCs are enriched for active TE families
427 but are also composed of inactive families.

428 Next, we sought to distinguish family-level enrichment of TE insertions within rare '*de novo*' piCs
429 (smaller piCs (<10 kb) shared by no more than 3 strains) and within common larger 'trap'-like piCs
430 (>10 kb, shared by at least 5 strains). To increase statistical power for this analysis, we used all
431 TE sequences annotated by RepeatMasker in each genome, instead of only non-reference
432 insertions. For iso-1, we found that only 4 TE families are significantly enriched within young piCs.
433 Interestingly, all 4 are LTR retrotransposon families of the *Ty3/Mdg4* superfamily (*blood, 412, flea,*
434 *Stalker-2*) and all except *flea* belong to the *Mdg1* lineage (Bertocchi et al. 2020; Costas et al.
435 2001), suggesting that this lineage of elements may be prone to seed new piCs. All four families
436 are also classified as active in this study as well as previous studies that examined TE insertion
437 frequency among *D. melanogaster* populations (Kofler et al. 2015; Kelleher and Barbash 2013).
438 By contrast, we found that numerous active and inactive families from all TE subclasses are
439 significantly enriched in large and common "trap-like" piCs, largely representative of the overall
440 TE landscape of *D. melanogaster* **(Fig. 5C)**. In the A1 genome, 20 TE families are significantly
441 enriched in rare piCs. Again, these are predominantly LTR retrotransposons (14 families), but 4
442 DNA transposon families and 2 LINE families are also significantly enriched. **(Fig. 5D)**.
443 Interestingly, *blood* insertions are neither enriched nor depleted in common piCs of both strains.
444 Taken together, these analyses yield a contrasting portrait of TE composition in the two major
445 types of piCs.

446 **Figure 5. Insertions of only few active LTR families associates with rare piCs.** *(A)* Scatter plot of
447 non-reference TE insertion counts and mean population frequencies of 184 TE families. Red box
448 highlights selected TEs classified as 'active'. *(B)* Enrichment analyses of TE families in master-
449 list piCs using random shuffling. Y-axis P-values are from binomial tests conducted to compare
450 observed counts to expected average overlaps of de novo TE insertions to piCs for each family.
451 *(C-D)* Enrichment analyses of TE families in master-list rare and common piCs of A1/Can-S
452 strain using random shuffling. *P* -values on y-axes are from binomial test conducted to compare
453 observed counts to expected average overlaps of de novo TE insertions to piCs for each family.
454 Names of some of the statistically significant families are shown.

(E-F) Enrichment analyses of TE families in the master-list of rare and common piCs of iso-1 using random shuffling. *P*-values on y-axes are from binomial tests conducted to compare observed counts to expected average overlaps of de novo TE insertions to piCs for each family. Names of some of the statistically significant families are shown.

Figure 5



455 **TE composition of piCs captures distinct steps in piC evolution**

456 To further illuminate the evolution of piCs, we analyzed how the overall age and distribution of
457 TEs of piCs change as they become more frequent, and presumably older. We plotted mean
458 percent divergence of individual TE insertions to their consensus sequences (a measure of TE
459 age) across piCs and their flanking non-piC regions for each piC frequency class (strain-specific
460 or shared by 2-8 strains). We found that the divergence of TE insertions in rare piCs (shared by 3
461 or less strains) is markedly lower (3.5-5%) than in their flanking regions (10-15%) **(Fig. 6A)**. In
462 addition, the divergence of TE copies within piCs increases gradually with the frequency of the
463 piCs to the extent that for the most common piCs (shared by 7 and 8 strains) the mean percent
464 TE divergence is only slightly lower than in their flanking genomic regions. This apparent increase
465 in age of TE insertions as piCs become more frequent provides weight to the inference that more
466 common piCs represent evolutionary older clusters relative to those that are strain-specific or rare.
467 It also suggests that piCs are born from *de novo* TE insertions and grow by gradual accretion of
468 TEs over time.

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

479

Figure 6

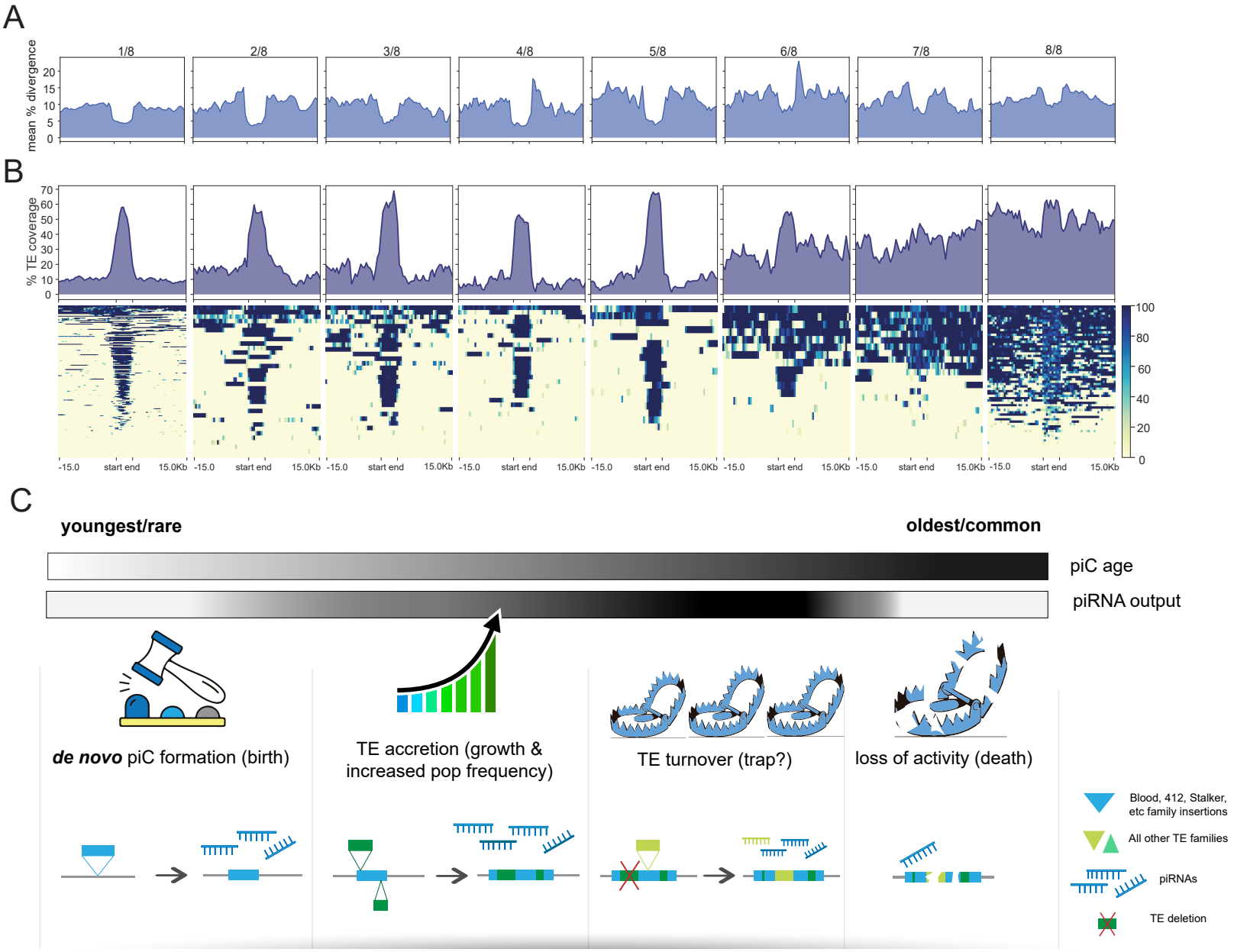


Figure 6. Age and distribution of TEs provide a portrait of intra-specific piC evolution. *(A)* Summary profile plot of mean percent divergence in 500 bp windows in scaled piC regions and flanking non-piC regions of the reference genome. Each group represents piCs that are shared by 1/8 to 8/8 strains. *(B)* 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. *(C)* New unified model of piC evolution in four steps from left to right.

480 ## Discussion

481 To study piC evolution at fine-scale resolution in *D. melanogaster*, we used population genomics
482 methods to characterize piC variation across eight inbred strains. A crucial asset was the
483 availability of high-quality genome assemblies for these strains (Chakraborty et al. 2019). This
484 enabled us to produce *de novo* annotation of piCs for each strain after mapping inferred piRNAs
485 from ovarian small RNA libraries we constructed and sequenced for two biological replicates
486 sampled six months apart. Our piC annotations for the two replicates exhibited high reproducibility
487 with >95% of piCs annotated in one replicate found in the second replicate **(Fig. 1A,B,**
488 **Supplementary Fig. 3)**. Also, to understand variation in sequence composition and age of piCs,
489 it was necessary to produce libraries of TE consensus sequences representative of the eight
490 strains analyzed here. By performing *de novo* discovery and re-annotation of TE families for each
491 genome, we identified 49 novel TE consensus **(Fig. 4)**. While further investigation is required to
492 examine their evolutionary origins and relationship to known TE families, it appears that many of
493 the novel TE families we annotated were highly diverged from known families and often "hidden"
494 in highly repetitive regions that would likely be poorly assembled in short-read genome
495 assemblies. These results stress the benefits of high-quality genome assemblies and the
496 necessity to perform *de novo* TE discovery when new individuals, strains or geographical isolates
497 are considered. This is true even for model species like *D. melanogaster,* where TEs have been
498 extensively cataloged, because previous TE identifications were mostly based on a single
499 reference genome. Robust annotation of piCs and TEs allowed us to compare in detail the activity
500 and TE composition of piCs across strains and test the generality of two contrasting models of piC
501 evolution.

502 We sought to distinguish which of the two models – '*de novo*' or 'trap' -- best captures piC evolution
503 in *D. melanogaster*. First, the chromosomal location and distribution of piCs sampled in this study
504 are largely consistent with the '*de novo*' model. We do, however, identify a small subset of piCs

505 (20-30) that are shared across most or all strains, a characteristic that is congruent with the 'trap'
506 model **(Fig. 1C,E)**. Second, we observe extensive variation in piC activity and abundant strain-
507 specific piCs, features supporting the '*de novo*' model **(Fig. 1F)**. We also find a positive correlation
508 between piC length and frequency suggesting that piCs are born small but grow in size as they
509 become more common in the population **(Fig. 1G)**. Third, INDELs associated with piCs exhibit
510 predicted signatures of both '*de novo*' and '*trap*' models but only in specific groups of piCs – rare
511 and common respectively **(Fig. 3)**. Fourth, the age of TE insertions within piCs are consistent with
512 the 'trap' model, whereby active or recently active TE families are enriched, while inactive ones
513 are depleted **(Fig. 4,5)**. Thus, our findings recapitulate predictions of both '*de novo*' and trap
514 models of piC evolution.

515 Overall, our results support the idea that piCs are primarily born '*de novo*', but a small subset of
516 large heterochromatic clusters are more evolutionarily stable and appear to behave as 'traps'. For
517 example, piCs like *flamenco*, *20A*, and *h52-3* show robust piRNA expression in all strains
518 analyzed. However, several large piCs like *42AB, 38C, Myo81F* show significant loss in piRNA
519 production in one or multiple strains **(Fig. 2, Supplementary Fig 6)**. We were able to rule out
520 mappability differences as a confounding factor **(Fig. 2)**. What then causes the loss in piRNA
521 production from large heterochromatic piCs? Our current lack of understanding of the cis-
522 regulatory requirements for piC activity makes it difficult to determine whether changes in piRNA
523 production are caused by genetic or epigenetic changes in the piCs. We and others (Wierzbicki
524 et al. 2021b; Ellison and Cao 2020) observe considerable structural variation among strains in
525 large peri-centromeric piCs**,** including *38C* and *42AB.* It is possible that such structural changes
526 result in changes in piRNA production, but further studies are needed to elucidate the mechanisms
527 by which large and seemingly stable piCs lose their activity.

528 What can TE composition of piCs tell us about the coevolution of TEs and piCs? As predicted and
529 previously reported (Chen et al. 2021; Brennecke et al. 2007; Ellison and Cao 2020; Wierzbicki et

530 al. 2023; Kofler et al. 2015), we found that diverse TE families (from all subclasses) are enriched
531 in large, common piCs **(Fig. 5)**. However, we found that this trend is driven mostly by younger
532 LTR insertions **(Fig. 4)**. This enrichment may be explained by selection against *de novo* TE
533 insertions in gene-rich euchromatic regions, which leads to unrestricted accumulation of TEs in
534 heterochromatic regions and where purifying selection is also weak (Blumenstiel et al. 2002;
535 Schrider et al. 2013; Charlesworth and Langley 1989; Dolgin and Charlesworth 2006). However,
536 our genome-wide analysis revealed that SV enrichment in common piCs cannot be completely
537 explained by their overlap with heterochromatin **(Supplementary Fig 7)**. Thus, the enrichment of
538 LTR elements within large common piCs may be driven in part by their insertion preference and/or
539 by selection for their repression.

540 Are particular TEs prone to give rise to piCs *de novo*? To answer this question, we tested for
541 enrichment of individual TE families within rare piCs. Interestingly, only a small set of
542 retrotransposon families are enriched within such clusters and most belong to the *mdg1* subclade
543 of LTR retrotransposons such as *blood*, *412,* and *Stalker* families **(Fig. 5C,D)** (Kapitonov and
544 Jurka 2003; Nefedova and Kim 2009). Why would these elements be prone to seed new piCs?
545 We hypothesize that this may be linked to their propensity to produce double-stranded RNA
546 (dsRNA) and endogenous siRNAs. It is well known that many LTR and non-LTR retrotransposons
547 possess bi-directional promoters that can result in the formation of dsRNAs that stimulate the
548 production of siRNAs (Hung and Slotkin 2021; Watanabe et al. 2008), including *blood* and *412*
549 (Russo et al. 2016). Because endo-siRNAs production has been associated with the formation of
550 transgenic piCs in flies (Olovnikov et al. 2013; Le Thomas et al. 2014), it is possible that the
551 propensity of these retrotransposon families to produce dsRNAs nucleate the formation of piCs.
552 This idea has received support in recent study showing that endogenous siRNA production
553 precedes piRNA cluster formation and maternal inheritance of these siRNAs is required for
554 licensing of piRNA clusters (Luo et al. 2022). Thus, we hypothesize that a subset of

555 retrotransposon insertions prone to produce dsRNA enter the endo-siRNA pathway which in turn
556 promote the birth of new piCs at these loci. In other words, the rapid evolution of piCs among *D.*
557 *melanogaster* strains may be driven by the activity of a few TE families.

558 Based on all these observations, we propose a 'birth-and-death' model of piC evolution, which
559 combines components of both 'trap' and '*de novo'* models (Moon et al. 2018; Shpiz et al. 2014;
560 Zhang et al. 2020; Bergman et al. 2006). In this model **(Fig 6C)**, we posit that piCs form frequently
561 throughout the genome, mostly from recent TE insertions, with certain LTR retrotransposon
562 families making a stronger contribution to seeding new piCs. Newly emerged piCs may increase
563 in frequency and size due to natural selection or drift, depending on factors such as their
564 propensity to trigger genomic autoimmunity (Blumenstiel et al. 2016; Lee and Karpen 2017),
565 ectopic recombination (Petrov et al. 2003; Sentmanat and Elgin 2012) and the establishment of a
566 chromatin environment conducive for piRNA production (le Thomas et al. 2014). Over time, these
567 stabilized clusters may grow by 'trapping' additional TE insertions, which will eventually result in
568 large heterochromatin clusters such as *flamenco* and *42AB*. Due to the host's limited capacity to
569 maintain such piCs without incurring a fitness cost, those clusters that lack piRNAs targeting active
570 TEs may gradually lose activity or become dispensable (Gebert et al. 2021). Further studies are
571 warranted to test this "birth-and-death" model. Our study provides a first in-depth view of piC
572 evolution in *D. melanogaster* that is likely to stimulate other comparative studies of piRNA
573 evolution.

574 Materials and Methods

575 Fly stocks

576 DSPR founder stocks of A1 (b1_paired), A2 (b3841_paired), A4 (b1_3852), A7 (t7_paired), B3
577 (b3864_paired), B6 (t1_paired) and Oregon-R were a gift from Anthony Long (UCI). iso-1

578 reference strain (#2057) was obtained from Bloomington Drosophila Stock Center. All stocks were
579 maintained on standard cornmeal medium at 22°C under a 12-hr day/night cycle.

580 **Small RNA library construction and sequencing**

581 Small RNA libraries were constructed by size fractionation on urea-polyacrylamide gel
582 electrophoresis as described in Ma *et al.* 2021. Libraries were constructed for two biological
583 replicates per strain, from ovaries collected at times separated by 25 weeks. Briefly, ovaries were
584 dissected from 25-30 yeast-fed adult females of 4-6 day old and total RNA was extracted using
585 TRIzol reagent and quantified with NanoDrop. Small RNAs of 17-29 nt length were size
586 fractionated from 10 µg of total RNA on Novex TBE-Urea Gels, 15% (Thermo Fischer
587 EC6885BOX) using ZR small-RNA ladder (ZymoResearch, R1090) as reference. Small RNAs in
588 excised gel fragments were first eluted in 500µL 0.3M NaCl and kept on an agitator for 16 hrs.
589 Small RNAs were then precipitated with 2 volumes of chilled isopropanol and 1 µL of 20 mg/mL
590 glycogen. Small RNA pellet was then washed with chilled 75% ethanol and eluted in 10µL of
591 freshly made 50% (w/v) PEG-8000 to enhance 3′ end ligation efficiency. Library preparation was
592 carried out using half of eluted small RNAs (5 µL) for each replicate with NEB Small RNA library
593 preparation kit (E7300) as per manufacturer's protocol. All libraries were PCR amplified to 14
594 cycles, visualized on a 2% agarose gel, and purified with NEB Monarch PCR & DNA Cleanup Kit
595 (T1030S). All libraries were quantified using Qubit 3.0, pooled into replicate-1 and replicate-2
596 groups, and analyzed on Agilent Bioanalyzer. Single end 75 bp Illumina sequencing was carried
597 out for all libraries on NextSeq500 at Cornell Biotechnology Resource Center and few select
598 libraries were re-sequenced if a minimum of 10 million reads was not obtained in first round.

599 **Processing of small RNA libraries**

600 Reads were first trimmed of the adapter sequences and quality filtered using cutadapt
601 (v3.4)(Martin 2011). Read length distribution, per sequence quality and duplication level was

602 obtained from FastQC (Andrews and others 2010). Reads mapping to annotated miRNAs
603 (Kozomara et al. 2019), other non-coding RNAs like rRNA, snRNA, tRNA, and snoRNA sequences
604 (Hoskins et al. 2015) were removed using bowtie (-v 2 -k 1 -y –un -S) and a combined custom
605 reference of non-coding small RNAs. The remaining 23-29 nt genome-mapping reads were
606 retained as piRNA reads.

607 **piC annotation**

608 Active piRNA cluster (piC) annotation was conducted independently for each replicate of each
609 strain using a custom pipeline adapted from previously described methods (Mohn et al. 2014;
610 Rosenkranz and Zischler 2012). Genome mapping 23-29 reads, filtered from annotated non-
611 piRNA small RNA genes, were mapped to respective genome assemblies using bowtie -n 1 -l 12
612 -a -m 1 -y -S and resulting unique alignments were separated from unmapped reads using
613 samtools in bam files (Langmead et al. 2009; Li et al. 2009). bedtools *makewindows* (Quinlan and
614 Hall 2010) was used to create 500 bp bookended windows from 7 DSPR genome assemblies
615 (Chakraborty et al. 2019) and iso-1 reference assembly (Hoskins et al. 2015) followed by bedtools
616 *coverage* to calculate uniquely mapped piRNA reads per million (RPM) per window from bam files.
617 Windows with piRNA expression of 2 RPM or more were merged if located within 10 kb of each
618 other into piRNA expression domains. RPKM values for piRNA expression was calculated for such
619 domains (ranged from 500 bp to 330 kb).

620 piC annotation from merged domains was conducted in two modes - permissive and restrictive.
621 First piRNA reads that uniquely map to selected domains were recovered and separated for each
622 domain using samtools to quantify unique piRNA sequences per domain. Additionally, theoretical
623 mappability scores of 0-1 for 25 nt reads was computed using GEM (Derrien et al. 2012) for
624 bookended 500 bp windows for each genome assembly . For the *restrictive* mode of annotation
625 - domains with at least 8 unique piRNAs per kb per million total piRNA reads were selected and
626 then merged if interrupted by low mappability region of 10 kb or less. Similarly, for the permissive

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

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

634 The same processed small RNA libraries as described above were used for alternative piC
635 annotation using proTRAC-v2.4.4 (Rosenkranz and Zischler 2012). Specifically, for proTRAC
636 analysis, each library was collapsed to include only unique piRNA sequences using TBr2_collapse
637 of NGS-toolbox (Rosenkranz et al. 2015) and mapped to respective genomes using bowtie -n 1 -
638 l 12 -a --best --strata --quiet -y -chunkmbs 1024. proTRAC 2.4.4 was run on sam files generated
639 from bowtie with the following parameters -swsize 500 -swincr 100 -clsize 500 -1Tor10A 0.6 -pimin
640 23 -clhitsn 10 -pdens 0.2 -pti followed by removal of clusters with normalized multi-mapped piRNA
641 coverage of 25 or less. proTRAC piC annotation was extracted from resulting clusters.gtf files.

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

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

651 **Manual curation and collapsed replicate annotations**

652 Recovery of complete piC regions primarily depends on expression and density of uniquely
653 mapped piRNAs along the length of the cluster. Clusters identified independently from two
654 different strains may differ in length due to natural variation in piRNA expression or density along
655 the length of the cluster. To identify homologous clusters across strains despite changed
656 boundaries due to structural variation in piCs, most heterochromatic piCs were examined in the
657 IGV browser. Clusters resulting from merged bins across annotated protein-coding genes were
658 unmerged into separate clusters. Any cluster partitioned into multiple smaller clusters due to lack
659 of any uniquely mapping piRNAs for more than 20 kb were merged to recover the complete cluster.
660 All strain genome assembly-match piC annotations from *restrictive*, *proTRAC* and *master-list* are
661 provided in **Supplementary Table S2,** whereas curated and replicate collapsed annotations are
662 provided in **Supplementary Table S3.**

663 **Structural variation detection and filtering**

664 Raw long reads for the 7 DSPR strains, the iso-1 reference strain, *D. simulans* (wxd1) and *D.*
665 *sechellia* (sech25) were mapped to the *D. melanogaster* iso-1 release 6 (GCA_000001215.4)
666 without the Y chromosome with minimap-2.1 map-pb --N3 and resulting sam file was converted to
667 bam and sorted (Li 2018; Li et al. 2009). Details of raw long reads used are provided in
668 **Supplementary Table S4**. Three structural variant (SV) callers – sniffles-2.0, cuteSV-1.0.13, and
669 svim-2.0 were used for SV detection. All three callers were run with default parameters but mapQ
670 required to be >50 and minimum read support required adjusted for each strain by sequencing
671 coverage i.e., 5 reads per 100X coverage(Jiang et al. 2020; Sedlazeck et al. 2018; Heller and
672 Vingron 2019). Only insertions and deletions >30 bp and duplications and inversions of >10 kb
673 were retained. Next, for each caller, biallelic SVs with precise mapping were selected and merged
674 from 8 samples using Survivor (Jeffares et al. 2017) and only simple SVs (non-complex and
675 unambiguous) were used. Summary of filtered and raw SV calls are in Supplementary Table 3.

676 Genotyping for merged SVs for each caller was performed by cuteSV-1.0.13 with min_support set
677 as 3 reads. Genotyped calls that were supported by two SV callers or more, for which intra-strain
678 allele frequency (AF) could be determined were then filtered and their SV length, AF, and read
679 support averaged from results of SV callers using bedtools *merge* function. Additionally,
680 overlapping SVs of same type with length difference by >20% or 500 bp were treated as
681 independent events, otherwise collapsed as a same SV event. 71% of all simple SVs filtered were
682 detected by at least two callers and 44.5% detected by all three callers. Next, *D. melanogaster*
683 SVs were polarized by comparison to their absolute presence or absence in *D. simulans* and *D.*
684 *sechellia*. Any SVs with conflicted calls between these two sister species were ignored. Filtered
685 and polarized SV calls are reported in **Supplementary Table S5**.

686 **Structural variant parsing and enrichment analyses**

687 Filtered SVs of insertions and deletions class were used for parsing for further analysis. The vast
688 majority of SVs are expected to have extremely low frequency of <0.1, which reflects the general
689 deleterious nature of SVs. Since raw long read data used from published studies were from pooled
690 sequencing of ~200 flies for DSPR strains and 60-80 flies for iso-1, only SVs with intra-strain AF
691 of 0.2 or more were considered, to enrich for germline SVs. SV enrichment analysis was carried
692 out using *poverlap* (https://github.com/brentp/poverlap) with 1000 bootstraps of random shuffling.
693 Mean expected overlap counts against piC coordinates were compared to expected overlap.

694 ***de novo* TE annotation**

695 To create a comprehensive and accurate representative TE library representing the TE insertions
696 contained in the 8 strains, *de novo* TE annotation was conducted using several computational
697 tools. Summary of all major steps is presented in a flow-chart in **Fig. 4A**. Briefly, canonical FlyBase
698 TE consensus sequences were curated for each strain to include only TEs that best represent the
699 TE insertion landscape of each strain using RepeatMasker-4.1.0 results(Smit 1999; Larkin et al.

700 2021). FlyBase TE families with at least 3 copies of >200 bp and <1% divergence was retained.
701 This resulted in a reference TE library for each strain, averaging ~110 TE families. Next,
702 RepeatModeler2 was run on the 7 DSPR genomes (Chakraborty et al. 2019) and reference iso-1
703 strain(Hoskins et al. 2015) followed by removal of non-TE repeats like tRNA, satellites, rRNA etc.,
704 as well as TE sub-families using bash scripts (Flynn et al. 2020). Next, identification of novel TE
705 families absent in the FlyBase TE library was carried out using the 80-80-80 rule (Wicker et al.
706 2007; Quesneville et al. 2005). TE consensus fragments from RepeatModeler2 library that passed
707 the 80-80-80 rule using blastn (Camacho et al. 2009) alignments with reference TEs were
708 removed and remaining likely "novel" ones were used to mask respective genome assemblies
709 with RepeatMasker. From RepeatMasker results, all TE consensus fragments with at least 3
710 copies of >200 bp with <5% divergence from the consensus were retained as potential novel TE
711 family fragments and combined into one compiled DSPR-library for all 8 strains.

712 **TE family and super-family enrichment analyses**

713 RepeatMasker outputs of the comprehensive *D. melanogaster* TE library on respective genome
714 assemblies of 8 strains was generated (Smit 1999). RM.out files were parsed, and insertions
715 defragmented using scripts from https://github.com/4ureliek/Parsing-RepeatMasker-Outputs
716 (Kapusta et al. 2017). Family-level and Superfamily-level enrichment analyses were conducted
717 using the TE_analysis_Shuffle_bed.pl from https://github.com/4ureliek/TEanalysis (Kapusta et al.
718 2013). 1000 bootstraps were performed and only TE families with 10 or more total insertions were
719 considered for enrichment analyses.

720 **Phylogenetic tree construction**

721 Maximum likelihood (ML) trees were constructed for all defragmented iso-1 TE insertions for
722 superfamilies and families reported in **Fig. 6** using the method described below. First, TE insertion
723 sequences were extracted from the reference genome using bedtools *getfasta.* Sequences in the

724 range of 200-500 bp were manually examined for further defragmentation if nearby insertions of
725 the same family were present with non-overlapping sequences when compared to consensus.
726 Sequences were then subjected to multiple sequence alignment using mafft v7.453 with the E-
727 INS-i strategy and following parameters --adjustdirectionaccurately --maxiterate 1000. Next, ML
728 trees were constructed using raxML-HPC with GTRGAMMA model (Stamatakis 2014). ML trees
729 were then uploaded in R and terminal branch lengths calculated and visualized using the tidytree
730 and ggtree R packages (Yu 2022; Yu et al. 2018).

731 ***de novo* TE insertion detection from long reads**

732 Raw long reads utilized in SV detection was also used for *de novo* TE insertion analyses. Reads
733 (>1 kb) were mapped to iso-1 reference genome (without Y-linked contigs and all contigs<20 kb)
734 using minimap2.1 default parameters and resulting sam file converted to bam file, sorted and
735 indexed using samtools (Li 2018; Li et al. 2009). TLDR, a *de novo* TE detection program, was run
736 for each strain using comprehensive *D. melanogaster* TE library curated in this study (Ewing et al.
737 2020). Insertion calls with 2-3 different TE families were treated as nested or complex insertion,
738 where each family was treated as an independent insertion. However, they were ignored if all
739 insertions belong to the same superfamily classification to remove false-positive hits resulting from
740 micro-homologies between related families. Nested calls were also ignored if they belonged to
741 more than 3 families. TLDR was run with default parameters except for *--flanksize* 200 --
742 *max_te_len* 15000 *-m* 2 and result insertions table filtered. TLDR calls retained for analyses by
743 filtering for medianMapQ >20, TEmatch>80%, and intrasample frequency of >0.05 (supporting
744 reads/empty reads). TE insertion enrichment analyses were performed using
745 TE_analysis_Shuffle-bed.pl script (Kapusta et al. 2013). Filtered TLDR insertion calls for all 8
746 strains are reported in **Supplementary table S7**.

747

Data availability

Small RNA libraries sequenced in this study are deposited in SRA under the project accession PRJNAXXXXX. Relevant source data for figures is listed in supplementary tables. Raw data files, TE consensus sequences and scripts are available at https://github.com/kerogens101/Dmel_piCs.

Acknowledgements

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. This work was supported by grants from the National Institute of Health, R01-GM119125 to A.G.C and R35-GM122550 to C.F.

Competing interests

All authors declare they have no competing interests.

Supplementary Tables

Supplementary_Table_S1. Summary statistics of sequencing depth and mapping of all small RNA libraries analyzed in this study.

Supplementary_Table_S2. All data downloaded and analyzed from public domain.

Supplementary_Table_S3. piRNA cluster (piC) annotations for each small RNA library per strain from *restrictive*, *proTRAC*, and master-List.

Supplementary_Table_S4. piC population frequency for each pipeline after remap to iso-1 genome (liftOver).

Supplementary_Table_S5. Collapsed, genotyped, filtered, and polarized structural variant calls.

Supplementary_Table_S6. TE classification and activity determined from non-reference TE insertions.

Supplementary_Table_S7. Filtered and collapsed TLDR non-reference TE insertion calls.

775 **References**

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

777 Assis R, Kondrashov AS. 2009. Rapid repetitive element-mediated expansion of piRNA clusters
778 in mammalian evolution. *Proceedings of the National Academy of Sciences* **106**: 7079–
779 7082.

780 Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. 2011. BamTools: a C++ API
781 and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691–1692.

782 Bartolomé C, Maside X, Charlesworth B. 2002. On the Abundance and Distribution of
783 Transposable Elements in the Genome of Drosophila melanogaster. *Mol Biol Evol* **19**: 926–
784 937.

785 Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. 2006. Recurrent insertion and
786 duplication generate networks of transposable element sequences in the *Drosophila*
787 *melanogaster* genome. *Genome Biol* **7**: R112.

788 Bertocchi NA, Torres FP, Deprá M, da Silva Valente VL. 2020. Evolutionary study of the
789 412/mdg1 lineage of the Ty3/gypsy group of
790 LTR retrotransposons in Diptera. *bioRxiv* 2020.09.24.311225.

791 Bingham PM, Chapman CH. 1986. Evidence that white-blood is a novel type of temperature-
792 sensitive mutation resulting from temperature-dependent effects of a transposon insertion
793 on formation of white transcripts. *EMBO J* **5**: 3343–51.

794 Bingham PM, Kidwell MG, Rubin GM. 1982. The molecular basis of P-M hybrid dysgenesis: The
795 role of the P element, a P-strain-specific transposon family. *Cell* **29**: 995–1004.

796 Blumenstiel JP, Erwin AA, Hemmer LW. 2016. What Drives Positive Selection in the Drosophila
797 piRNA Machinery? The Genomic Autoimmunity Hypothesis. *Yale J Biol Med* **89**: 499–512.

798 Blumenstiel JP, Hartl DL, Lozovsky ER. 2002. Patterns of insertion and deletion in contrasting
799 chromatin domains. *Mol Biol Evol* **19**: 2211–25.

800 Bogu GK, Reverter F, Marti-Renom MA, Snyder MP, Guigó R. 2019. Atlas of transcriptionally
801 active transposable elements in human adult tissues. *bioRxiv* 714212.

802 Brand CL, Levine MT. 2021. Functional Diversification of Chromatin on Rapid Evolutionary
803 Timescales. *Annu Rev Genet* **55**: 401–425.

804 Brennecke J, Aravin A a, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ. 2007.
805 Discrete small RNA-generating loci as master regulators of transposon activity in
806 Drosophila. *Cell* **128**: 1089–103.

807 Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. 2008. An
808 Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. *Science (1979)*
809 **322**: 1387–1392.

810 Calvi BR, Gelbart WM. 1994. The basis for germline specificity of the hobo transposable element
811 in Drosophila melanogaster. *EMBO J* **13**: 1636.

812 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
813 BLAST+: Architecture and applications. *BMC Bioinformatics* **10**: 1–9.

814 Carr M, Bensasson D, Bergman CM. 2012. Evolutionary Genomics of Transposable Elements in
815 Saccharomyces cerevisiae. *PLoS One* **7**: e50978.

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

819 Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit
820 widespread allelic heterogeneity and shape variation in complex traits. *Nat Commun* **10**: 1–
821 11.

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

825 Charlesworth B, Jarne P, Assimacopoulos S. 1994. The distribution of transposable elements
826 within and between chromosomes in a population of Drosophila melanogaster. III. Element
827 abundances in heterochromatin. *Genet Res (Camb)* **64**: 183–197.

828 Charlesworth B, Langley CH. 1989. The population genetics of Drosophila transposable
829 elements. *Annu Rev Genet* **23**: 251–287.

830 Chen P, Kotov AA, Godneeva BK, Bazylev SS, Olenina L V., Aravin AA. 2021. piRNA-mediated
831 gene regulation and adaptation to sex-specific transposon expression in D. melanogaster
832 male germline. *Genes Dev* **38**.

833 Chirn G-W, Rahman R, Sytnikova YA, Matts JA, Zeng M, Gerlach D, Yu M, Berger B, Naramura
834 M, Kile BT, et al. 2015. Conserved piRNA Expression from a Distinct Set of piRNA Cluster
835 Loci in Eutherian Mammals. *PLoS Genet* **11**: e1005652.

836 Coronado-Zamora M, Salces-Ortiz J, González J. 2023. DrosOmics: A Browser to Explore -
837 omics Variation Across High-Quality Reference Genomes From Natural Populations of
838 *Drosophila melanogaster* ed. J. Parsch. *Mol Biol Evol* **40**: 2022.07.22.501088.

839 Cosby RL, Chang N-C, Feschotte C. 2019. Host–transposon interactions: conflict, cooperation,
840 and cooption. *Genes Dev* **33**: 1098–1116.

841 Costas J, Valadé E, Naveira H. 2001. Structural Features of the mdg1 Lineage of the Ty3/gypsy
842 Group of LTR Retrotransposons Inferred from the Phylogenetic Analyses of Its Open
843 Reading Frames. *J Mol Evol* **53**: 165–171.

844 Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ. 2018. piRNA-
845 Guided Genome Defense: From Biogenesis to Silencing. *Annu Rev Genet* **52**: 131–157.

846 de Vanssay A, Bougé A-L, Boivin A, Hermant C, Teysset L, Delmarre V, Antoniewski C,
847 Ronsseray S. 2012. Paramutation in Drosophila linked to emergence of a piRNA-producing
848 locus. *Nature* **490**: 112–115.

849 di Nocera PP, Graziani F, Lavorgna G. 1986. Genomic and structural organization of Drosophila
850 melanogaster G elements. *Nucleic Acids Res* **14**: 675–91.

851 Dolgin ES, Charlesworth B. 2006. The fate of transposable elements in asexual populations.
852 *Genetics* **174**: 817–827.

853 Dopman EB, Hartl DL. 2007. A portrait of copy-number polymorphism in Drosophila
854 melanogaster. *Proceedings of the National Academy of Sciences* **104**: 19920–19925.

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

858 Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, Cheetham SW,
859 Faulkner GJ. 2020. Nanopore Sequencing Enables Comprehensive Transposable Element
860 Epigenomic Profiling. *Mol Cell* **80**: 915-928.e5.

861 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
862 RepeatModeler2 for automated genomic discovery of transposable element families.
863 *Proceedings of the National Academy of Sciences* **117**: 9451–9457.

864 Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. 2021. Large Drosophila
865 germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation.
866 *Mol Cell* **81**: 3965-3978.e5.

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

870 Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar
871 DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs
872 in animals. *Nature 2008 455:7217* **455**: 1193–1197.

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

876 Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand
877 breaks, and the maintenance of genome integrity. *Mutat Res* **616**: 46–59.

878 Heller D, Vingron M. 2019. SVIM: structural variant identification using mapped long reads.
879 *Bioinformatics* **35**: 2907–2915.

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

883 Houwing S, Kamminga LM, Berezikov E, Cronembold D, Girard A, van den Elst H, Filippov D v.,
884 Blaser H, Raz E, Moens CB, et al. 2007. A Role for Piwi and piRNAs in Germ Cell
885 Maintenance and Transposon Silencing in Zebrafish. *Cell* **129**: 69–82.

886 Huang CRL, Burns KH, Boeke JD. 2012. Active transposition in genomes. *Annu Rev Genet* **46**:
887 651–75.

888 Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D,
889 Lyman RF, et al. 2014. Natural variation in genome architecture among 205 Drosophila
890 melanogaster Genetic Reference Panel lines. *Genome Res* **24**: 1193–1208.

891 Huang Y, Shukla H, Lee YCG. 2022. Species-specific chromatin landscape determines how
892 transposable elements shape genome evolution. *Elife* **11**.

893 Hung YH, Slotkin RK. 2021. The initiation of RNA interference (RNAi) in plants. *Curr Opin Plant*
894 *Biol* **61**.

895 Inouye S, Hattori K, Yuki S, Saigo K. 1986. Structural variations in the *Drosophila*
896 retrotransposon, *17.6*. *Nucleic Acids Res* **14**: 4765–4778.

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

900 Jiang T, Liu Y, Jiang Y, Li J, Gao Y, Cui Z, Liu Y, Liu B, Wang Y. 2020. Long-read-based human
901 genomic structural variation detection with cuteSV. *Genome Biol* **21**: 1–24.

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

905 Kapitonov V V., Jurka J. 2003. Molecular paleontology of transposable elements in the
906 *Drosophila melanogaster* genome. *Proceedings of the National Academy of Sciences* **100**:
907 6569–6574.

908 Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay LA, Bourque G, Yandell M, Feschotte C.
909 2013. Transposable Elements Are Major Contributors to the Origin, Diversification, and
910 Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genet* **9**: e1003470.

911 Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and
912 mammals. *Proceedings of the National Academy of Sciences* **114**.

913 Kelleher ES, Barbash DA. 2013. Analysis of piRNA-mediated silencing of active TEs in
914 Drosophila melanogaster suggests limits on the evolution of host genome defense. *Mol Biol*
915 *Evol* **30**: 1816–29.

916 Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD,
917 Weng Z, Theurkauf WE. 2011. Adaptation to P element transposon invasion in drosophila
918 melanogaster. *Cell* **147**: 1551–1563.

919 Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS,
920 Nowosielska A, et al. 2009. The Drosophila HP1 homolog Rhino is required for transposon
921 silencing and piRNA production by dual-strand clusters. *Cell* **138**: 1137–49.

922 Kofler R. 2020. piRNA Clusters Need a Minimum Size to Control Transposable Element
923 Invasions. *Genome Biol Evol* **12**: 736–749.

924 Kofler R, Nolte V, Schlötterer C. 2015. Tempo and Mode of Transposable Element Activity in
925 Drosophila. *PLoS Genet* **11**: e1005406.

Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C. 2018. Molecular dissection of a natural transposable element invasion. *Genome Res* **28**: 824–835.

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. MiRBase: From microRNA sequences to function. *Nucleic Acids Res* **47**: D155–D162.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25.

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.

Laski FA, Rio DC, Rubin GM. 1986. Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell* **44**: 7–19.

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**: 363–7.

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**: 1667–80.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *Elife* **6**.

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

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

Li XZ, Roy CK, Dong X, Bolcun-Filas E, Wang J, Han BW, Xu J, Moore MJ, Schimenti JC, Weng Z, et al. 2013. An Ancient Transcription Factor Initiates the Burst of piRNA Production during Early Meiosis in Mouse Testes. *Mol Cell* **50**: 67–81.

Luo S, Zhang H, Duan Y, Yao X, Clark AG, Lu J. 2020. The evolutionary arms race between transposable elements and piRNAs in Drosophila melanogaster. *BMC Evol Biol* **20**: 14.

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.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, Hannon GJ. 2009. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. *Cell* **137**: 522–535.

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

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

966 Mohamed M, Dang NT-M, Ogyama Y, Burlet N, Mugat B, Boulesteix M, Mérel V, Veber P,
967 Salces-Ortiz J, Severac D, et al. 2020. A Transposon Story: From TE Content to TE
968 Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing
969 Technology from Oxford Nanopore. *Cells* **9**.

970 Mohn F, Sienski G, Handler D, Brennecke J. 2014. The Rhino-Deadlock-Cutoff Complex
971 Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell*
972 **157**: 1364–1379.

973 Montgomery EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by
974 ectopic recombination in Drosophila melanogaster: genome structure and evolution.
975 *Genetics* **129**: 1085–98.

976 Moon S, Cassani M, Lin YA, Wang L, Dou K, Zhang ZZ. 2018. A Robust Transposon-
977 Endogenizing Response from Germline Stem Cells. *Dev Cell* **47**: 660-671.e3.

978 Muerdter F, Olovnikov I, Molaro a., Rozhkov N V., Czech B, Gordon a., Hannon GJ, Aravin a. a.
979 2012. Production of artificial piRNAs in flies and mice. *Rna* **18**: 42–52.

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

981 Nefedova LN, Kim AI. 2009. Molecular phylogeny and systematics of drosophila
982 retrotransposons and retroviruses. *Mol Biol* **43**: 747–756.

983 Olovnikov I, Ryazansky S, Shpiz S, Lavrov S, Abramov Y, Vaury C, Jensen S, Kalmykova A.
984 2013. De novo piRNA cluster formation in the Drosophila germ line triggered by transgenes
985 containing a transcribed transposon fragment. *Nucleic Acids Res* **41**: 5757–5768.

986 Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD. 2019. PIWI-interacting RNAs: small
987 RNAs with big functions. *Nat Rev Genet* **20**: 89–108.

988 Palmer WH, Hadfield JD, Obbard DJ. 2018. RNA-Interference Pathways Display High Rates of
989 Adaptive Protein Evolution in Multiple Invertebrates. *Genetics* **208**: 1585–1599.

990 Parhad SS, Theurkauf WE. 2019. Rapid evolution and conserved function of the piRNA
991 pathway. *Open Biol* **9**: 180181.

992 Parhad SS, Yu T, Zhang G, Rice NP, Weng Z, Theurkauf WE. 2020. Adaptive Evolution Targets
993 a piRNA Precursor Transcription Network. *Cell Rep* **30**: 2672-2685.e5.

994 Petrov D a., Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR
995 retrotransposable elements and ectopic recombination in drosophila. *Mol Biol Evol* **20**: 880–
996 892.

997 Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, Anxolabehere D.
998 2005. Combined Evidence Annotation of Transposable Elements in Genome Sequences.
999 *PLoS Comput Biol* **1**: e22.

1000 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features.
1001 *Bioinformatics* **26**: 841–842.

1002 Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V,
1003 Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers
1004 transposable elements associated with gene expression variation and adaptive signatures
1005 in Drosophila. *Nat Commun* **13**: 1948.

1006 Riddle NC, Minoda A, Kharchenko P v., Alekseyenko AA, Schwartz YB, Tolstorukov MY,
1007 Gorchakov AA, Jaffe JD, Kennedy C, Linder-Basso D, et al. 2011. Plasticity in patterns of
1008 histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome*
1009 *Res* **21**: 147–163.

1010 Robine N, Lau NC, Balla S, Jin Z, Okamura K, Kuramochi-Miyagawa S, Blower MD, Lai EC.
1011 2009. A Broadly Conserved Pathway Generates 3′UTR-Directed Primary piRNAs. *Current*
1012 *Biology* **19**: 2066–2076.

1013 Ronsseray S, Lehmann M, Anxolabéhère D. 1991. The maternally inherited regulation of P
1014 elements in Drosophila melanogaster can be elicited by two P copies at cytological site 1A
1015 on the X chromosome. *Genetics* **129**: 501–12.

1016 Roovers EF, Rosenkranz D, Mahdipour M, Han CT, He N, de Sousa Lopes SMC, van der
1017 Westerlaken LAJ, Zischler H, Butter F, Roelen BAJ, et al. 2015. Piwi Proteins and piRNAs
1018 in Mammalian Oocytes and Early Embryos. *Cell Rep* **10**: 2069–2082.

1019 Rosenkranz D, Han C-T, Roovers EF, Zischler H, Ketting RF. 2015. Piwi proteins and piRNAs in
1020 mammalian oocytes and early embryos: From sample to sequence. *Genom Data* **5**: 309.

1021 Rosenkranz D, Zischler H. 2012. proTRAC--a software for probabilistic piRNA cluster detection,
1022 visualization and analysis. *BMC Bioinformatics* **13**: 5.

1023 Russo J, Harrington AW, Steiniger M. 2016. Antisense Transcription of Retrotransposons in
1024 Drosophila: An Origin of Endogenous Small Interfering RNA Precursors. *Genetics* **202**:
1025 107–21.

1026 Ryazansky S, Radion E, Mironova A, Akulenko N, Abramov Y, Morgunova V, Kordyukova MY,
1027 Olovnikov I, Kalmykova A. 2017. Natural variation of piRNA expression affects immunity to
1028 transposable elements. *PLoS Genet* **13**: e1006731.

1029 Said I, McGurk MP, Clark AG, Barbash DA. 2022. Patterns of piRNA Regulation in *Drosophila*
1030 Revealed through Transposable Element Clade Inference. *Mol Biol Evol* **39**.1

1031 Saito K, Siomi MC. 2010. Small RNA-Mediated Quiescence of Transposable Elements in
1032 Animals. *Dev Cell* **19**: 687–697.

1033 Schrider DR, Houle D, Lynch M, Hahn MW. 2013. Rates and genomic consequences of
1034 spontaneous mutational events in Drosophila melanogaster. *Genetics* **194**: 937–54.

1035 Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC.
1036 2018. Accurate detection of complex structural variations using single-molecule
1037 sequencing. *Nature Methods 2018 15:6* **15**: 461–468.

1038 Sentmanat MF, Elgin SCR. 2012. Ectopic assembly of heterochromatin in Drosophila
1039 melanogaster triggered by transposable elements. *Proc Natl Acad Sci U S A* **109**: 14104–
1040 14109.

1041 Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. 2014. Euchromatic Transposon
1042 Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the
1043 Drosophila Germline. *PLoS Genet* **10**: e1004138.

1044 Simkin A, Wong A, Poh Y-P, Theurkauf WE, Jensen JD. 2013. Recurrent and recent selective
1045 sweeps in the piRNA pathway. *Evolution* **67**: 1081.

1046 Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in
1047 mammalian genomes. *Curr Opin Genet Dev* **9**: 657–663.

1048 Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, Emerson JJ, Hawley RS.
1049 2018. Rapid Low-Cost Assembly of the *Drosophila melanogaster* Reference Genome Using
1050 Low-Coverage, Long-Read Sequencing. *G3 Genes|Genomes|Genetics* **8**: 3143–3154.

1051 Srivastav SP, Rahman R, Ma Q, Pierre J, Bandyopadhyay S, Lau NC. 2019b. Har-P, a short P-
1052 element variant, weaponizes P-transposase to severely impair Drosophila development ed.
1053 M.B. Eisen. *Elife* **8**: e49948.

1054 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
1055 phylogenies. *Bioinformatics* **30**: 1312–3.

1056 Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses
1057 and transposable elements in eukaryotes. *Nature Reviews Genetics 2017 18:5* **18**: 292–
1058 308.

1059 Vermaak D, Henikoff S, Malik HS. 2005. Positive Selection Drives the Evolution of rhino, a
1060 Member of the Heterochromatin Protein 1 Family in Drosophila. *PLoS Genet* **1**: e9.

1061 Wang L, Barbash DA, Kelleher ES. 2020. Adaptive evolution among cytoplasmic piRNA proteins
1062 leads to decreased genomic auto-immunity. *PLoS Genet* **16**: e1008861.

1063 Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H,
1064 Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed
1065 dsRNAs regulate transcripts in mouse oocytes. *Nature 2008 453:7194* **453**: 539–543.

1066 Wells JN, Feschotte C. 2020. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev*
1067 *Genet* **54**: 539.

1068 Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P,
1069 Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic
1070 transposable elements. *Nat Rev Genet* **8**: 973–982.

1071 Wierzbicki F, Kofler R, Signor S. 2023. Evolutionary dynamics of piRNA clusters in Drosophila.
1072 *Mol Ecol* **32**: 1306–1322.

1073 Yi M, Chen F, Luo M, Cheng Y, Zhao H, Cheng H, Zhou R. 2014. Rapid Evolution of piRNA
1074 Pathway in the Teleost Fish: Implication for an Adaptation to Transposon Diversity.
1075 *Genome Biol Evol* **6**: 1393–1407.

1076 Yoth M, Gueguen N, Jensen S, Brasset E. 2022. Germline piRNAs counteract endogenous
1077 retrovirus invasion from somatic cells. *bioRxiv* 2022.08.29.505639.

1078 Yu G. 2022. *Data Integration, Manipulation and Visualization of Phylogenetic Trees*. CRC Press.

1079 Yu G, Lam TT-Y, Zhu H, Guan Y. 2018. Two Methods for Mapping and Visualizing Associated
1080 Data on Phylogeny Using Ggtree ed. F.U. Battistuzzi. *Mol Biol Evol* **35**: 3041–3043.

1081 Yu T, Koppetsch BS, Pagliarani S, Johnston S, Silverstein NJ, Luban J, Chappell K, Weng Z,
1082 Theurkauf WE. 2019. The piRNA Response to Retroviral Invasion of the Koala Genome.
1083 *Cell* **179**: 632-643.e12.

1084 Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S. 2013.
1085 Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the
1086 regulatory properties of piRNA clusters. *Proc Natl Acad Sci U S A* **110**: 19842–7.

1087 Zhang S, Pointer B, Kelleher ES. 2020. Rapid evolution of piRNA-mediated silencing of an
1088 invading transposable element was driven by abundant de novo mutations. *Genome Res*
1089 **30**: 566–575.

1090 Zichner T, Garfield DA, Rausch T, Stütz AM, Cannavó E, Braun M, Furlong EEM, Korbel JO.
1091 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on
1092 population-scale sequencing. *Genome Res* **23**: 568–579.

1093