Centromere-associated retroelement evolution in *Drosophila melanogaster* reveals an underlying conflict

Lucas W. Hemmer[1,2*], Sherif Negm[1,3], Xuewen Geng[1], Cécile Courret[1], Beatriz Navarro-Domínguez[1,4], Iain Speece[1], Xiaolu Wei[1], Eddyson Altidor[1], James Chaffer[1], John S. Sproul[1,5], Amanda M. Larracuente[1*]

Affiliations:

[1] Department of Biology, University of Rochester, Rochester, NY, USA

[2] Invaio Sciences, Cambridge, MA, USA

[3] Department of Human Genetics, University of Chicago, Chicago, IL, USA

[4] Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, 18071, Spain

[5] Department of Biology, University of Nebraska Omaha, Omaha, NE, USA

[*]**Correspondence:** alarracu@bio.rochester.edu, lhemmer08@gmail.com

17 **ABSTRACT**

18

19 Centromeres are chromosomal regions essential for coordinating chromosome segregation during
20 cell division. While centromeres are defined by the presence of a centromere-specific histone H3
21 variant called CENP-A rather than a particular DNA sequence, they are typically embedded in
22 repeat-dense and heterochromatic chromosomal genome regions. In many species, centromeres
23 are associated with transposable elements but it is unclear if these elements are selfish and target
24 centromeres for insertion or if they play a role in centromere specification and function. Here we
25 use *Drosophila melanogaste*r as a model to understand the evolution of centromere-associated
26 transposable elements. *G2/Jockey-3* is a non-LTR retroelement in the *Jockey* clade and the only
27 sequence shared by all centromeres. We study the evolution of *G2/Jockey-3* using short and long
28 read population genomic data to infer insertion polymorphisms across the genome. We combine
29 estimates of the age, frequency, and location of insertions to infer the evolutionary processes
30 shaping *G2/Jockey-3* and its association with the centromeres. We find that *G2/Jockey-3* is an
31 active retroelement that is targeted by the piRNA pathway. Our results suggest that *G2-Jockey-3*
32 is highly enriched in centromeres at least in part due to an insertion bias. We do not detect any
33 signature of positive selection on any *G2/Jockey-3* insertions that would suggest than individual
34 insertions are favored by natural selection. Instead, we infer that most insertions are neutral or
35 weakly deleterious both inside and outside of the centromeres. Therefore, *G2/Jockey-3* evolution
36 is consistent with it being a selfish genetic element that targets centromeres. We suspect targeting
37 centromeres for insertion helps active retroelements like *G2/Jockey-3* escape host defenses, as the
38 unique centromeric chromatin may prevent targeting by the host silencing machinery. On the other
39 hand, centromeric TEs insertions may be tolerated or even beneficial if they also contribute to the

right transcriptional and chromatin environment. Thus, we suspect centromere-associated retroelements like *G2/Jockey-3* reflect a balance between conflict and cooperation at the centromeres.

43 **INTRODUCTION**

44

45 Centromeres are important chromosomal structures for kinetochore assembly and proper
46 chromosome segregation during cell division. While centromeres are usually in highly repetitive
47 regions, the role of DNA sequences in centromere function or establishment remains unclear
48 (reviewed in [1]). Across many organisms, centromeres are associated with satellite DNAs
49 (satDNAs), large blocks of tandemly repeated DNA sequences (reviewed in [2, 3]). Centromere-
50 associated tandem repeats are capable of initiating centromere assembly in fission yeast [4] and
51 humans [5]. However, centromeres are defined epigenetically by CENP-A, a specific variant of
52 histone H3 [6-11] and the centromere domains consist of unique chromatin with H3 nucleosomes
53 interspersed with CENP-A nucleosomes [12-15]. Repeat sequences are not always required for
54 centromere formation, as neocentromeres can form outside of centromeric DNA in the absence of
55 satDNA sequences [16-18] and some species have centromeres devoid of repeats (*e.g.* satellite-
56 free centromeres in the Equus genus [19]). Neocentromeres have the potential to replace older
57 centromeres with implications for karyotype evolution and speciation [20-22].

58

59 Centromeres and pericentromeric heterochromatin are also enriched in transposable elements
60 (TEs)— selfish genetic elements that mobilize in genomes. Evidence for specific TE-centromere
61 associations has emerged in the last two decades, including in grasses [23, 24], rice [25, 26], maize
62 [27], marsupials [28, 29], bats [30], gibbons [31, 32], and humans [18, 33, 34]. However, it is
63 unclear how, or if, TEs contribute to centromere biology for most species [35]. Centromeres may
64 offer a 'safe haven' to TEs that jump there and for the genome [36, 37], as centromere insertions
65 are unlikely to disrupt centromere function [35]. At the same time, centromeric TEs may contribute

66 to important aspects of centromere evolution and biology. Homology between some TEs and
67 centromeric tandem repeats [3, 38] suggests that TEs can contribute to centromere evolution by
68 giving rise to new centromeric satDNAs [39]. TEs may also play roles in centromere function.
69 These roles may be indirect, such as maintaining the constitutive pericentromeric heterochromatin,
70 which is important for centromere function [40-42]. However, more direct roles are also possible.
71 Some transcription of centromeric regions is necessary for recruitment of CENP-A and
72 establishing centromere location [43-47]. Centromeric transcripts may also facilitate recruitment
73 of kinetochores to centromeric DNA [48, 49]. Given that TEs contain sequences important for
74 promoting transcription [50], their presence in centromeres may contribute to centromere
75 specification and function [35]. In most cases it is unclear whether TEs are a part of the CENP-A
76 domain corresponding to functional centromeres or just associated with adjacent, repeat-rich
77 regions of chromosomes in which functional centromeres are often embedded. Advances in single
78 molecule long-read sequencing technology now allows for detailed study of centromere
79 organization at the sequence level to provide insights into centromere structure, composition, and
80 evolution [51-56].
81
82 *Drosophila melanogaster* is an excellent model organism to study centromere evolution with its
83 compact genome and expansive genetic toolkit [57]. A recent study of centromere organization in
84 *D. melanogaster* revealed that CENP-A is bound to TE-rich islands of complex DNA embedded
85 within large arrays of satellite repeats [53]. The composition of each centromere island is unique.
86 The only sequence element shared among all centromeres is a single non-LTR retroelement in the
87 *Jockey* superfamily, *G2/Jockey-3*. However, *G2/Jockey-3* is not exclusively located in
88 centromeres; about 40% of all insertions are found outside of the centromere islands [53].

Polymorphic *G2/Jockey-3* insertions [58] and transcripts from published RNAseq datasets [59-61] suggest this element may still be active in *D. melanogaster* [53]. *G2/Jockey-3* is also found at the centromeres in *D. simulans*, a close relative of *D. melanogaster,* suggesting that its centromere association may be broadly conserved [53].

It is unclear whether *G2/Jockey-3* is an important centromere component in *D. melanogaster* or if they are instead purely selfish TEs. If *G2/Jockey-3* are important for centromere specification or function, we may detect evidence for positive natural selection on centromeric copies. Alternatively, *G2/Jockey-3* may be a selfish element that targets insertions in centromeres but are otherwise selectively neutral or weakly deleterious. These hypotheses are not necessarily mutually exclusive. Here we investigated the age and frequency of *G2/Jockey-3* insertions across *D. melanogaster* genomes to determine if its evolution is consistent with a role in centromere function or underlying selfish behavior. Our analyses suggest that *G2/Jockey-3* insertions are largely neutral or weakly deleterious both inside and outside centromeres, suggesting that these are selfish genetic elements. Our results also suggest that *G2/Jockey-3* has an insertion bias toward centromeres. Although there is no evidence for positive selection on individual *G2/Jockey-3* centromeric insertions, insertion of active TEs like *G2/Jockey-3* may still contribute to transcription and the unique chromatin environment at centromeres. We propose that there may be a balance between conflict and cooperation between active TEs like *G2/Jockey-3* and centromeres.

112 **RESULTS**

113

114 *1. A new consensus sequence for G2/Jockey-3*

115

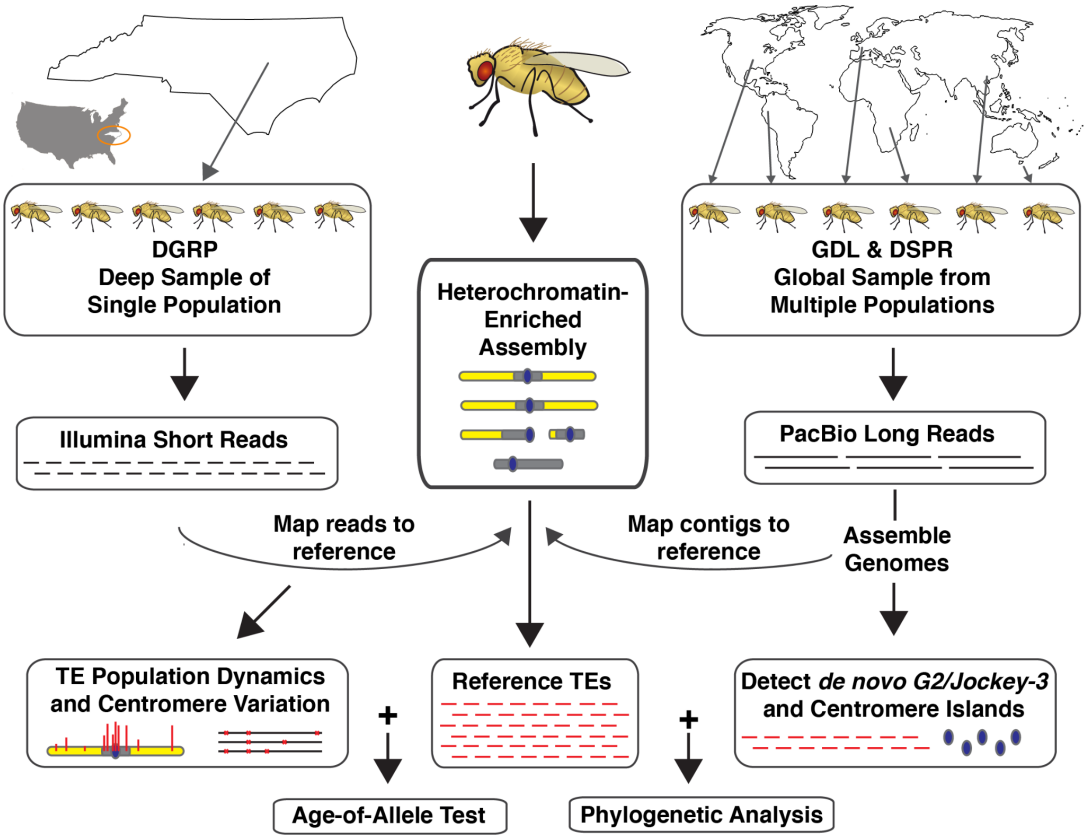
116 Prior to analyzing the distribution of *G2/Jockey-3* in populations, we first confirmed the consensus
117 annotation. We analyzed all *G2/Jockey-3* insertions in our *D. melanogaster* genome assembly [53]
118 and found two lines of evidence that the annotation needed updating: 1) 25 of the 167 *G2/Jockey-*
119 *3* sequences in the genome shared up to 92 bp of upstream sequence at the 5'-end of the annotation;
120 and 2) our reanalysis of PROseq data [62] suggests that *G2/Jockey-3* transcription starts 92 bp
121 upstream of the published start site (Figure S1A). This upstream region is also enriched for GO
122 terms for RNA polymerase II transcription factors and other transcription factor activity (Figure
123 S1B). We therefore extended the original *G2/Jockey-3* consensus sequence by 92 bases (File S1).

124

125 *G2/Jockey-3* is a part of the *Jockey* superfamily of non-LTR retroelements, which typically contain
126 two open reading frames (ORFs) (Figure S1A). While ORF1 is generally less conserved [63],
127 ORF2 is well-conserved across *Jockey* elements and encodes for a reverse transcriptase (RT) and
128 an endonuclease. NCBI ORF finder detected both ORFs within *G2/Jockey-3*. ORF1 is predicted
129 to produce a 427 amino-acid protein approximately 186 bp downstream of the 5' end. The NCBI
130 Conserved Domains identified a PRE_C2GC super family (cl11655, E-value 9.22E-18) within
131 ORF1 and while the function is unknown, it is associated with Cys2His2 type zinc fingers found
132 in putative GAG proteins in other arthropod species. The predicted ORF2 overlaps with the last
133 four bp of ORF1 and contains a domain corresponding to endonuclease (EEP Superfamily

134 Accession cl00490, E-value 7.21E-18) and RT (RT_like Superfamily Accession cd01650, E-value
135 4.62E-56). *G2/Jockey-3* ends in a poly(A) signal.
136
137 *2. Overview of data sets analyzed*
138
139 We studied the evolutionary dynamics of *G2/Jockey-3* and other TEs by investigating the patterns
140 of insertion polymorphisms both in a single panmictic population, and in individuals from a wide
141 geographic range (Figure 1, Table S1). We conducted a survey of TE dynamics and genetic
142 variation in 135 samples from the *Drosophila melanogaster* Genetic Reference Panel (DGRP)
143 collected from a single population in North Carolina with whole genome Illumina sequence data
144 [64, 65] and a subset of long read data from the same population for validation [66] (see
145 Supplemental Text). For the samples from a wide geographic range, we used both long- and short-
146 read data from 13 founder lines of the *Drosophila* Synthetic Population Resource (DSPR) and
147 Oregon-R [67]. We also analyzed long-read PacBio data from five Global Diversity Lines (GDL)
148 samples from different continents [68]. We generated *de novo* genome assemblies of each line to
149 determine the organization of the centromere islands originally described in the reference genome
150 for comparison (Table S2) and predicted insertions in short read data with the McClintock pipeline
151 [69]. We assessed the reliability of our TE insertion predictions by looking for concordance
152 between short and long read predictions (see Supplemental Text) and through PCR validation of a
153 subset of insertions. We confirmed 64% (67 / 105) of the *G2/Jockey-3* insertions detected by
154 McClintock are found in the long-read samples we examined (File S2). PCR confirmed 80% (28 /
155 35) of *G2/Jockey-3* insertions predicted in the GDL lines, and ~67% (12 / 18, File S3) of the
156 centromere insertions. Based on our assessment, we limit our analysis to short read datasets with

157 >20X centromere coverage (Table S3) and TEs predicted by more than one TE caller implemented
158 in McClintock (Supplemental Text).



159

160 **Figure 1: The summary of data and analyses using one source or a combination of sources**
161 **in this paper.** A heterochromatin-enriched assembly was originally used to discover centromere
162 islands in Chang et al. [53] and identify reference TE insertions. We used short-reads from the
163 DGRP to identify reference and *de novo* insertions to study TE dynamics within a single population
164 from North Carolina (map source: R package usmaps https://cran.r-
165 project.org/web/packages/usmap/). We generated genome assemblies for 18 samples from global
166 populations to uncover the centromere island sequences and *de novo* elements (map source:
167 www.outline-world-map.com). We combined the reference sequences with population frequencies
168 to infer selection with an age-of-allele test [70]. We also combined reference TE sequences with
169 *de novo* elements uncovered in our assemblies for phylogenetic analysis. The fly illustrations are

170 modified images from Wikipedia Commons

171 (https://commons.wikimedia.org/wiki/File:Drosophila-drawing.svg).

172 *3. G2/Jockey-3 is active in Drosophila melanogaster populations*

173

174 Previous studies of TE polymorphism in *D. melanogaster* are consistent with *G2/Jockey-3*
175 (previously annotated as *G2*) being recently active [58, 71, 72]. To determine if *G2/Jockey-3* is
176 active in the DGRP, we detected TE insertion polymorphisms with the McClintock pipeline [69]
177 and classified each insertion based on their presence (a "reference" insertion) or absence (a non-
178 reference or "*de novo*" insertion, Figure 2) in our reference genome [53, 73]. *G2/Jockey-3* showed
179 approximately five *de novo* insertions per line (File S4-S5), which is considerably fewer than the
180 most highly active *Drosophila* TEs. For example, we estimate *Hobo* and *Jockey-1* have
181 approximately 35 *de novo* insertions per line (consistent with [58]). However, the fraction of *de*
182 *novo* insertions as percentage of all insertions (39% or 733 / 1124 for *G2/Jockey3*) is comparable
183 to other highly active elements [72] such as *Doc* (25%; 2944 / 3826), *jockey* (24% ;1308 / 5448),
184 *I* element (27%; 658 / 2364), and *Copia* (88% ;1083 / 1226) (File S4).

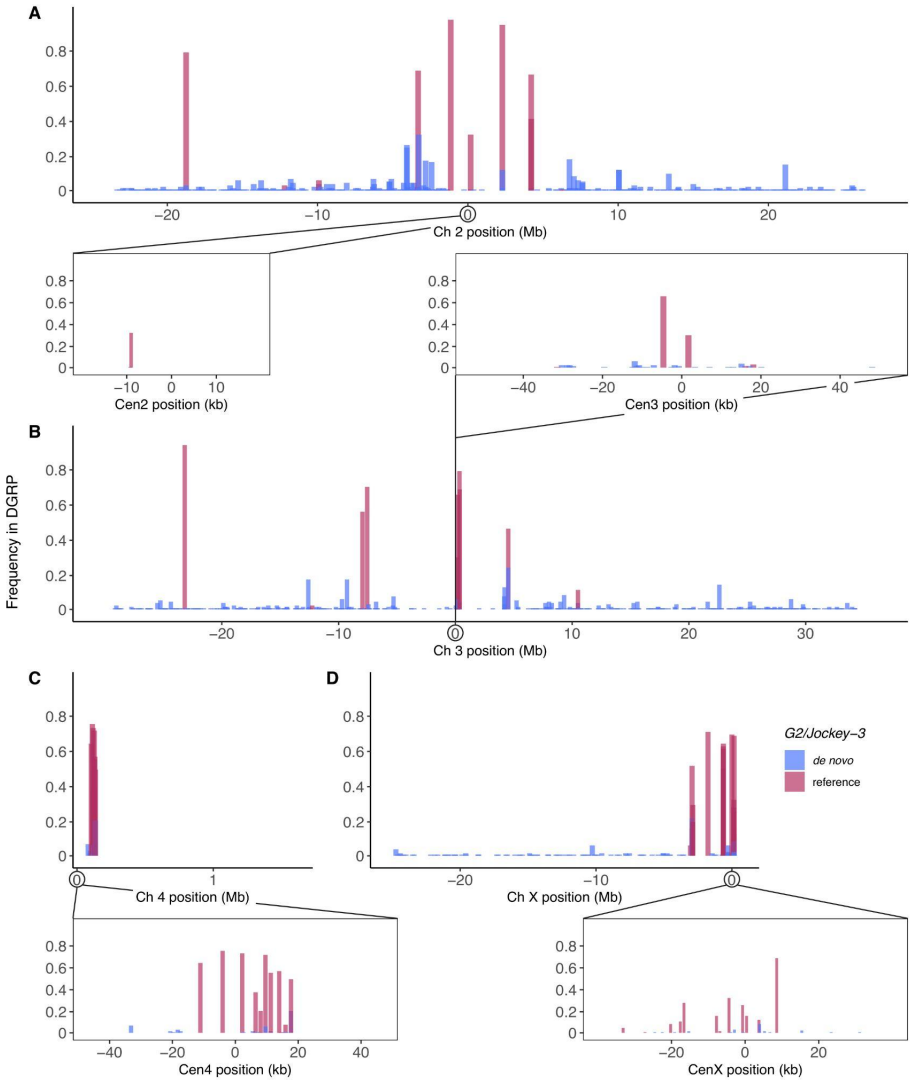
185

186 We detected insertions in the centromeres, here defined as the centromere islands identified in
187 Chang et al. [53]. Of the 18 different TE families detected within centromeres, *G2/Jockey-3* is the
188 most common. Within centromeres, *G2/Jockey-3* has the highest percentage of *de novo* (32%;
189 15/47) and reference (72%; 28 / 39, Figure 2) TE insertions. Only *G2/Jockey-3* and *Doc* have more
190 than 10 unique insertions (File S4). Other *de novo* centromere insertions include active families
191 such as the *F* element, *R1* and *Roo*.

192

193 Many active TEs are targeted by small, 24-30-nt long Piwi-interacting RNAs (piRNA) in the
194 germline [74, 75]. These piRNAs target TEs for both transcriptional and post-transcriptional

195 silencing. In the *D. melanogaster* germline, piRNAs are produced from precursor transcripts that
196 originate from discrete loci in the genome that contain clusters of nested TEs [76] called piRNA
197 clusters [77]. Subsequent amplification of the piRNA signal comes from the piRNA-guided
198 cleavage of active TEs through the 'ping-pong' cycle [77, 78] generating a signature of 10-nt
199 overlap between sense and antisense piRNAs. Consistent with *G2/Jockey-3* being a recently active
200 element, we detect insertions of *G2/Jockey-3* in the *38C1* (2 copies) and *38C2* (9 copies) piRNA
201 clusters, and *G2/Jockey-3*-derived piRNAs in the germline. *G2/Jockey-3* is ranked in the top 4%
202 and top 18% of TE-derived piRNAs in testes and ovaries, respectively (Table S4). The resulting
203 *G2/Jockey-3* piRNAs show signatures of ping-pong in ovaries (Z-score = 7.11; data from [60, 79,
204 80]) but not in testes (Z-score = 0.15; data from [81-83]), consistent with other testis piRNAs.
205 Taken together, these results suggest that *G2/Jockey-3* is an active TE enriched in centromere
206 islands.



207

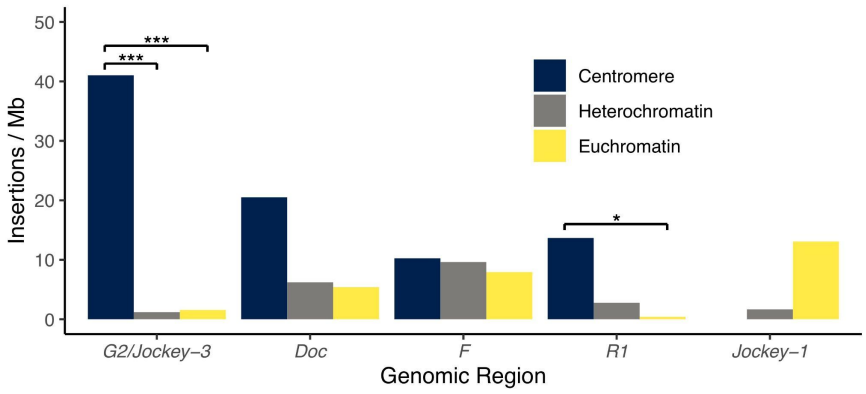
208 **Figure 2: The distribution and frequency of reference and *de novo G2/Jockey-3* insertions**
209 **along the chromosomes and within the centromeres in *D. melanogaster*.** The height of the bars
210 corresponds to frequency of an insertion detected in the DGRP samples and the color denotes
211 whether an insertion is found in the reference genome or *de novo*. Each A-D subpanels depict a
212 chromosome (2,3,4,X) from the heterochromatin-enriched assembly of *D. melanogaster* genome
213 [53, 73]. Corresponding centromere sequences are zoomed in as highlighted (X chromosome
214 centromere = CenX, *etc*).

215 *4. Young G2/Jockey3 insertions are enriched in centromeres*

216

217 To understand if the observed enrichment of *G2/Jockey3* at centromeres is due to insertion bias,
218 we studied insertion frequencies across the genome. We reasoned that an enrichment of putatively
219 young *G2/Jockey-3* at centromeres may reflect an insertion preference. Young recently inserted
220 TEs are expected to occur at lower frequencies than older TEs. We thus examined the distribution
221 of rare insertions (<10% population frequency) of *G2/Jockey-3* and other TE families in the DGRP.
222 We found that rare *G2/Jockey-3* insertions are significantly enriched in centromeres compared to
223 other regions relative to their size (12 insertions; FET, P < 1E-3, Figure 3). *R1* elements are the
224 only other TE enriched in centromeres, but only relative to euchromatin (FET, P = 0.014, Figure
225 3). *G2/Jockey-3* is therefore unique in being the only TE that is significantly enriched in
226 centromeres over the heterochromatin and euchromatin (Figure 3). Among rare insertions,
227 *G2/Jockey-3* is by far the most enriched TE in centromeres (39% of all rare insertions).

228



229

230 **Figure 3: Young *G2/Jockey-3* insertions are highly enriched in centromeres relative to**
231 **heterochromatin and euchromatin.** The distribution of rare insertions (< 10% of the population)
232 per Mb between the centromere islands, heterochromatin, and euchromatin regions for several TE
233 families in DGRP lines. Significant differences are noted as: * P < 0.05, *** P < 0.001.

234

235 We confirmed these patterns using an orthogonal approach with long-read genome assemblies of
236 flies from a wide geographic distribution (*i.e.,* founder DSPR lines, GDL lines, and our reference
237 genome; Table S1). Here we identified putatively young *G2/Jockey-3* insertions based on sequence
238 divergence instead of frequency. We classified young insertions as those with < 1% divergence
239 from the consensus sequence, or with < 2% divergence from the consensus and found within a
240 single genome assembly. Insertions classified as young were found in less than 50% of the
241 assemblies, had fewer unique substitutions, and were more likely to have a complete ORF than
242 older fragments (Figure 4, FET P = 3.93E-12, File S6, File S7). Consistent with our short-read
243 DGRP findings (Figure 3), we found significant relative enrichment in centromeres (Figure 5, FET
244 P = 8.28E-5; Y chromosome excluded P = 0.0142), even with some of the centromere island
245 sequences missing in the long-read assemblies (Table S2).

246

247 These results suggest that *G2/Jockey-3* shows at least some insertion preference for centromeres.
248 As insertion bias can be associated with a target site sequence, we searched for motif enrichment
249 at insertions sites. Although we found an enrichment of AT-rich simple motifs like "TAGTTTT"
250 and rDNA-associated sequences (*e.g.* 18S and non-transcribed spacers; Figure S2) within 100
251 bases of predicted insertion sites, these sequences are generally abundant in repeat-dense regions
252 of the genome and not specifically enriched at *G2/Jockey-3* insertions (File S6, File S8; MEME
253 E-value = 3.8E-16).

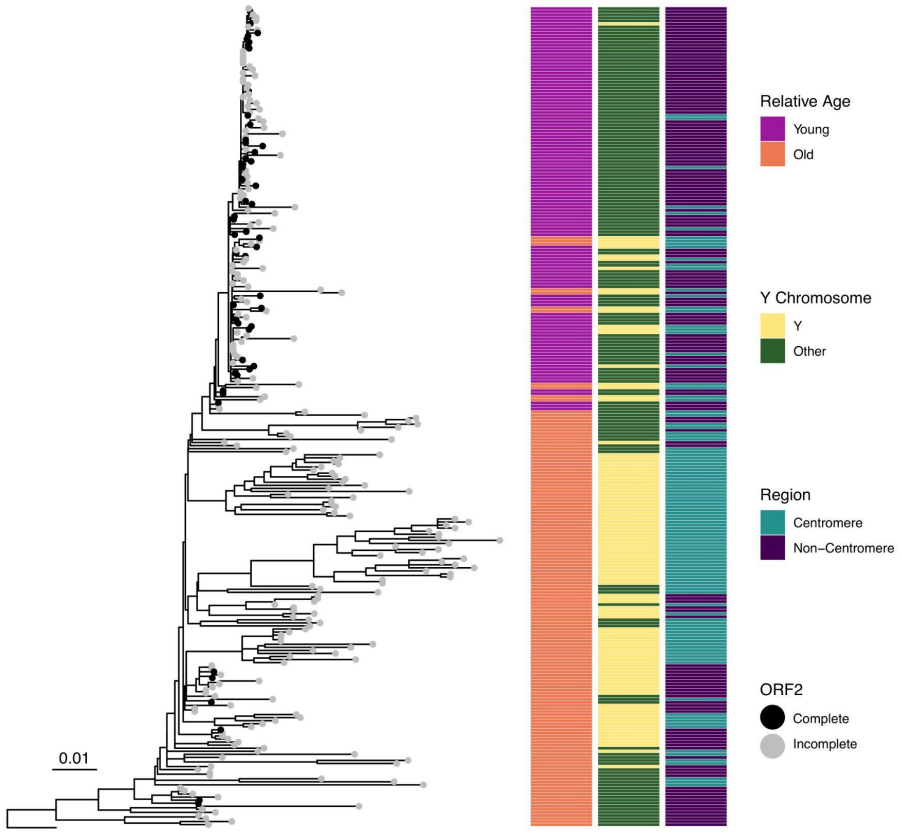


Figure 4: Young *G2/Jockey-3* insertions share a common origin and are enriched in centromeres. A maximum likelihood phylogenetic tree with all non-redundant *G2/Jockey-3* sequences larger than 10% of the full-length (400 bp) in the reference genome [53] and long-read assemblies of founder DSPR and GDL lines, with the *D. simulans* consensus as the outgroup. Black circles denote "complete" *G2/Jockey-3* elements containing a full ORF2; gray circles are incomplete elements. The columns and colors correspond to characteristics of each insertion including chromatin region (centromere or not), relative age, and presence on Y chromosome. Young insertions, determined using an orthogonal approach to the Figure 3 analysis, are those with < 1% divergence from the consensus sequence or < 2% and found within a single genome assembly.

265 *5. Skew toward high frequency variants of G2/Jockey-3 in the centromere*

266

267 To test if *G2/Jockey3* insertions are retained at centromeres by natural selection we studied their
268 insertion frequencies. Because most TE insertions are deleterious in the euchromatin, they are
269 found at low population frequencies [84]. In contrast, TEs in heterochromatin are allowed to reach
270 higher population frequencies because recombination is infrequent, and natural selection is less
271 efficacious (reviewed in [85-87]. To infer differences in the strength and direction of natural
272 selection in different genomic regions, we examined the frequency spectra of a subset of TE
273 families in centromeres, heterochromatin, and euchromatin (Figure 5, Figure S3). If natural
274 selection favors centromeric insertions, we should observe a skew toward high frequency variants
275 in centromeres compared to heterochromatin. We compared the frequency spectra of two TE
276 families which had five or more insertions in the centromeres: *G2/Jockey-3* and *Doc*. As expected,
277 both elements showed an enrichment for rare insertions in the euchromatin, consistent with TEs
278 being generally deleterious in gene-rich regions (Figure 5; FET $P < 0.001$). Similarly, we observed
279 a slight skew toward higher frequency *G2/Jockey-3* and *Doc* insertions (*e.g.*
280 common/intermediate) in heterochromatin and centromeres compared to euchromatin (Figure 5,
281 FET $P < 0.001$), suggesting relaxed constraints or a reduced efficacy of natural selection outside
282 of euchromatic regions. Other TE families found in centromeres are also enriched for low
283 frequency insertions in the euchromatin (Figure S3). However, we did not observe differences
284 between the frequency spectra of the euchromatin and centromeres, suggesting most TEs may be
285 deleterious in these regions (Figure S3).

286

The frequency spectra of *G2/Jockey3* does not appear significantly different between centromeres and heterochromatin. However, *G2/Jockey-3* has more "common" insertions in centromeres relative to other TEs, including *Doc* (Figure 5; Figure S3). This could be due to relaxed constraints allowing *G2/Jockey-3* insertions to rise to higher frequencies over time or positive selection for some *G2/Jockey-3* insertions in centromeres.

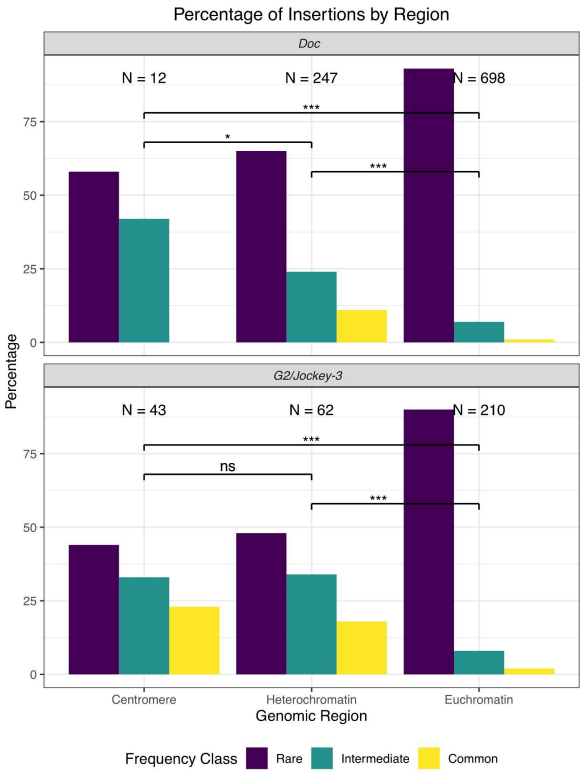


Figure 5: *G2/Jockey-3* has higher frequency variants in centromeres. TE frequencies were binned into rare (less than 10% of all samples), intermediate (10-50% of all samples), and common (> 50% of all samples). The frequency distributions of the four TE families with > 10 unique insertions in the centromeres were compared among the centromeres, heterochromatin, and euchromatin. Fisher's Exact Tests between regions, Bonferroni adjusted two-tailed P-value * $P < 0.05$; *** $P < 0.001$; ns is not significant.

300 *6. G2/Jockey-3 ages suggest a historical burst of activity*

301

302 While young *G2/Jockey-3* insertions appear to be enriched at centromeres (Figure 3), centromeric
303 copies are generally older—within a range of 2-7% divergence from the consensus—than those in
304 heterochromatin (MWU, $P < 0.001$, Figure 6A), but not significantly different than euchromatic
305 copies (MWU, $P > 0.5$; Figure 6A). This pattern suggests that most centromeric copies of
306 *G2/Jockey-3* resulted from a historical burst of activity. This burst of activity is consistent with
307 general patterns observed at other non-LTR retroelements [72, 88]. Other TE families show a
308 similar age distribution between heterochromatin and the centromeres (Figure S4). These copies
309 are also older and highly fragmented. In general, both the lack of recombination and lack of strong
310 purifying selection against TE insertions in centromeres and heterochromatin may allow these
311 insertions to persist over longer evolutionary time periods [89].

312

313 We can combine information about the TE's age and frequency to infer how, and if, its history is
314 consistent with natural selection. In the absence of selection, the frequency of a TE insertion should
315 correlate with its age: *e.g.,* older TEs have more time to drift to higher population frequencies than
316 young TEs and thus accumulate more unique mutations. Natural selection can cause deviations
317 from expected frequencies based on the age. While negative selection might cause TEs to have
318 lower population frequency than expected given their age, positive selection might increase their
319 frequencies. We therefore used an age-of-allele test [70] to detect whether *G2/Jockey-3*
320 frequencies deviate from our expectations. We estimated age as the number of unique substitutions
321 for each TE copy based on the reference genome and we estimated frequency from the 28 high-
322 coverage DGRP samples, controlling for demographic history [70]. For this test, we need to

compare insertions spanning a range of ages to identify individual insertion frequencies that deviate from expectations [70]. We thus examine *G2/Jockey-3* and other TE families that have centromere insertions. We find that *G2/Jockey-3* (Figure 6B, File S9), like the majority of TEs, segregates either at the expected frequencies or lower than expected, given their age (Figure S6, File S9). This suggests that many *G2/Jockey-3* copies, both within and outside the centromere, are either neutral or weakly deleterious. Therefore *G2/Jockey-3* shares similar selective pressures as other presumably weakly deleterious TE families in the genome.

Interestingly, we observed an enrichment of copies with zero unique substitutions for the heterochromatin and CenY insertions (Figure S5). However, we suspect this pattern arises from duplications rather than independent *G2/Jockey-3* insertion events (see below). Genomic duplication events are common in heterochromatin in the Y chromosome [73, 76, 90].

In sum, while we do not see evidence for strong recent positive selection on any individual *G2/Jockey-3* insertions in the centromeres, there may be selection on centromere haplotypes that differ in their composition or structure.

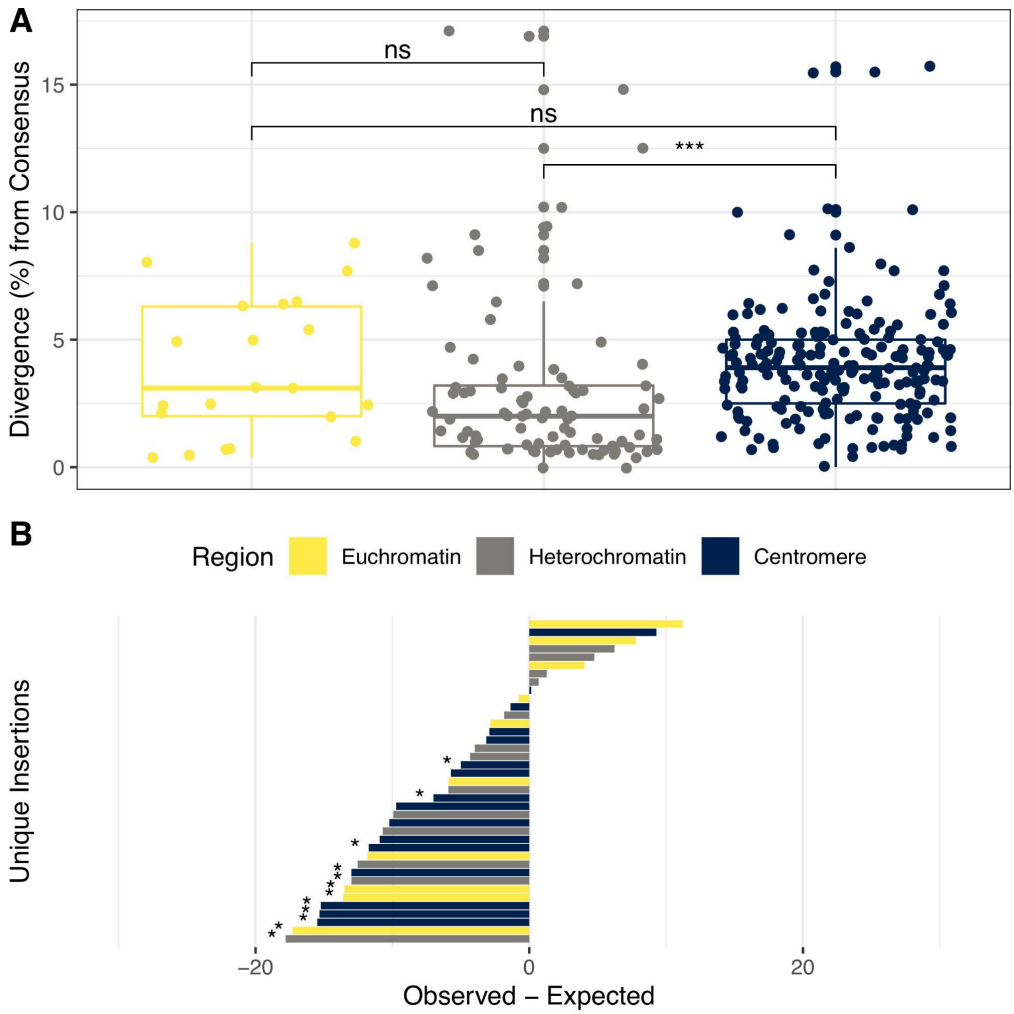


Figure 6: *G2/Jockey-3* copies are either as frequent or less frequent than expected given their age. A) The number of *G2/Jockey-3* insertions grouped by divergence from the consensus sequence split between the euchromatin, heterochromatin, and centromeres. *** $P < 0.001$, ns is not significant. B) The difference between observed and expected frequency of *G2/Jockey-3* copies in the DGRP. The *P*-values are calculated as the sum of the probabilities that an insertion is less frequent than expected (if observed - expected < 0) or more frequent than expected (observed - expected > 0). * $P < 0.1$, a stringent threshold given the reduced power to detect deviations from neutrality as stated in Blumenstiel et al. [70].

349 *7. Centromere island haplotype structure not likely driven by G2/Jockey-3 insertions*

350

351 To detect signatures of natural selection on centromere haplotypes and any associations with
352 *G2/Jockey-3* copies, we examined patterns of nucleotide variation. Centromeres and
353 pericentromeric regions in *D. melanogaster* do not recombine [91, 92] therefore the effects of
354 linked selection in non-recombining regions should lead to low levels of nucleotide diversity. We
355 thus expect to see reduced genetic variation in the centromeres and high levels of linkage
356 disequilibrium. We quantified genetic variation in the centromere islands for DGRP samples with
357 high centromere coverage (> 20X, see Methods for more details). As expected, nucleotide diversity
358 is greatly reduced in every centromere compared to the major chromosome arms (MWU $P < 2.0E-$
359 07, Figure S7, Table S5) except for chromosome 4, which does not recombine, and its centromere
360 and chromosome arms are similar (MWU $p = 0.56$, Figure S8, Table S5). We find that Tajima's
361 D, a summary of the site frequency spectrum, is also typically lower in centromere islands (Figure
362 S8), indicating a skew toward rare alleles in centromeres. The centromere island haplotype
363 structures (Figures S9-S11) indicate that the majority of centromeric SNPs are randomly
364 distributed with respect to the *G2/Jockey-3* sequences (FET $P > 0.46$). Linked selection anywhere
365 in or near centromeres can reduce genetic diversity, skew frequency spectra, and create haplotype
366 structure. While centromere island haplotype structure is consistent with a history of linked
367 selection, we do not see evidence that *G2/Jockey-3* insertions themselves are necessarily the targets
368 of selection.

369

370 *8. G2/Jockey-3 copy number amplification on the Y chromosome is driven by duplication*

371

372 The Y chromosome centromere is unique in that it has the largest centromere island containing
373 nearly half of all *G2/Jockey-3* fragments, but no identifiable satellite sequences [53]. We cannot
374 use the genomic re-sequencing data to study polymorphic insertions on the Y chromosome, as the
375 data are either female-only or mixed sex and thus lower coverage. However, our high-quality
376 reference assembly provides an opportunity to infer the organization and evolutionary history of
377 Y-linked *G2/Jockey-3*. We inferred a phylogenetic tree of all Y-linked copies of *G2/Jockey-3.*
378 Based on this phylogeny, we classified centromeric Y-linked *G2/Jockey-3* copies into four broad
379 classes (Figure S12). Class 1 *G2/Jockey-3* correspond to young and mostly long insertions, class
380 2 to short (< 160 bp) 3' end fragments, class 3 to old internal fragments missing the 5' and 3'
381 terminal ends, and class 4 to generally long and old 3' end fragments divided into two subclasses
382 (4 and 4A) based on position (Figure 9, Figure S12). We mapped these classifications onto our Y
383 chromosome assembly and estimated pairwise divergence between all *G2/Jockey-3* copies in the
384 Y centromere (Figure 7) and elsewhere on the Y chromosome (Figure S13). Each class seems to
385 undergo independent duplication events, however at a larger scale we do not identify higher order
386 repeat domains (Figure 7). This structure differs from other known centromeres (*e.g.* primates),
387 which are organized into higher order repeats, the youngest repeats are found in the center and the
388 older, more divergent repeats, are at the outer edges of the array [51, 52]. By contrast, in the *D.*
389 *melanogaster* Y centromere we observe that regions enriched with young *G2/Jockey-3* copies
390 punctuate regions rich with older copies (Figure 9), suggesting that recent, local rearrangements
391 affect centromere organization.

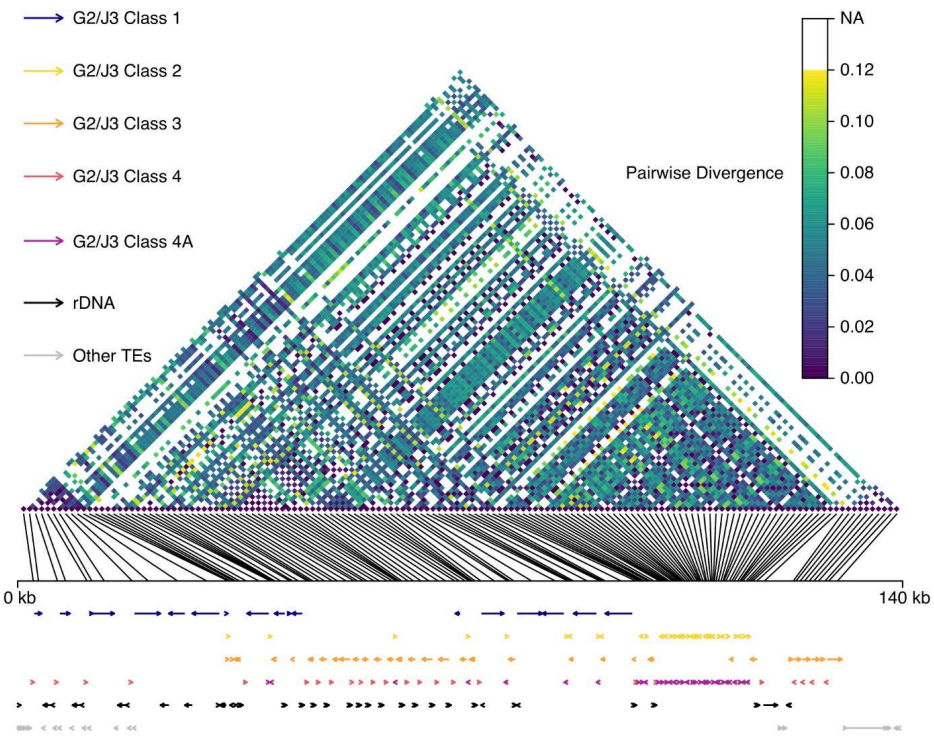


Figure 7: *G2/Jockey-3* evolution in the Y chromosome centromere is shaped by expansion via duplication, gene conversion and rearrangements. The colored pairwise divergences between individual *G2/Jockey-3* copies on the Y centromere are shown on top with their relative positions on the line graph below. White spaces represent *G2/Jockey-3* copies which do not overlap, so pairwise divergence cannot be calculated. Below the graph is a visual representation of the broad categories of repeats on the Y chromosome centromere, labeled by color with length and direction illustrated with arrowed lines. The repeats appear arranged into roughly four regions with overlapping borders based on the types of repeats and pairwise divergence. The classes of *G2/Jockey-3* are as follows: Class 1 younger insertions, Class 2 short (< 160 bp) 3' end fragments, Class 3 older internal fragments missing the 5' and 3' terminal ends, Class 4 generally longer and older 3' end fragments. Class 4 contains a subclade (4A) with highly similar sequences almost exclusively from the single region of CenY.

405

406 We estimated gene conversion rates of *G2/Jockey-3* on the Y centromere with respect to their
407 classes (Supplemental Text). Our estimates for gene conversion are on the order of 1E-6 per site
408 per generation (Table S6), ~10E3 times higher than the other chromosomes in *Drosophila* [91-93]
409 and an order of magnitude less than previously estimated gene conversion rates between duplicated
410 gene families on the Y chromosome [73]. Moreover, we do not see complete sequence
411 homogenization within classes. We observe similar patterns of duplication and expansion outside
412 the centromere on the Y chromosome (Figure S13), as 32 of 47 non-centromere Y-linked copies
413 appear to be duplicated; they are of similar fragment size, orientation, and have few no zero unique
414 substitutions (File S6). Therefore, we conclude that *G2/Jockey-3* evolution on the Y chromosome
415 is driven by duplication of historical insertions with some gene conversion among copies.
416 Population genomic data from the Y chromosome will help distinguish between types of
417 recombination affecting Y centromere structure.

418

419 Outside of the Y chromosome, putative duplications involving *G2/Jockey-3* are less common
420 (Figure S13). Many copies of *G2/Jockey-3* within the centromeres are fragmented and therefore
421 share no homologous sequences for pairwise calculations (Figure S13). Pairwise divergence
422 between centromeric *G2/Jockey-3* fragments is elevated in comparison to fragments in both the
423 euchromatin and non-centromeric heterochromatin (Figure S14, MWU, $P < 0.01$), and lowest in
424 non-centromeric heterochromatin, especially on the Y chromosome. Therefore, the majority of
425 *G2/Jockey-3* within the non-Y centromeres appears to be the result of recurrent past insertions.

426 CONCLUSIONS

427

428 *G2/Jockey-3* is a unique TE in that it is the only sequence shared among centromeres of *D.*
429 *melanogaster* [53]. Here we show that *G2/Jockey-3* is a recently active non-LTR retroelement that
430 may have an insertion preference for centromeres (Figures 3-4). While the centromere enrichment
431 and skew toward high frequency centromere copies are expected for insertions favored by natural
432 selection, we do not find evidence for positive selection on individual *G2/Jockey-3* insertions.
433 Instead, most *G2/Jockey-3* insertions both inside and outside of the centromeres are at the expected
434 frequency or significantly rarer than expected given their age, suggesting they may be neutral or
435 slightly deleterious and accumulating due to relaxed constraints against centromeric TE insertions
436 (Figure 6B). The majority of *G2/Jockey-3* copies within the centromeres appear to be the result of
437 recurrent past insertions, except for on the Y chromosome. Some Y-linked copies are likely a result
438 of duplications rather than *de novo* insertions, similar to patterns reported in other species (*e.g.*
439 [28]).

440

441 We propose that most *G2/Jockey-3* insertions in the centromeres resulted from a burst of activity,
442 which is consistent with the evolutionary history of other non-LTR retroelements [88]. When a TE
443 first invades, it can proliferate rapidly until it comes under the surveillance of germline defense
444 mechanisms like the piRNA pathway for example, by inserting into a piRNA cluster [94-96].
445 These dynamics are consistent with the invasion and control [94, 97] of the *P* element in *D.*
446 *melanogaster* [98] and its close relative, *D. simulans* [99]. Once targeted by the piRNA pathway
447 for transcriptional and posttranscriptional silencing, the TE may eventually become dormant or go

448 extinct. Here we show that *G2/Jockey-3* had a burst of activity and is now inside piRNA clusters
449 and actively targeted by the piRNA pathway (Table S4).
450
451 Therefore, the *G2/Jockey-3* insertion polymorphism patterns and piRNA observations are most
452 consistent with some underlying selfish behavior typical among TE families, except that this
453 element can target centromeres. Although we are unsure of the precise targeting mechanism, we
454 suspect the target is the centromere chromatin rather than any particular DNA sequence. Such
455 centromere-specific targeting was reported for LTR retroelements in other species such as maize
456 [100, 101] and *Arabidopsis* [102, 103]. For example, when the *Arabidopsis lyrata Tal-1 Copia-*
457 type retroelement [103] is introduced into *A. thaliana,* the element targets the centromeres despite
458 divergence in centromeric DNA sequences between these species [102].
459
460 TEs may accumulate in heterochromatic regions where they experience relaxed selection
461 compared to euchromatic insertions, which have potential to disrupt functional elements, risk
462 ectopic exchange and resulting genomic rearrangements [85-87], and spread silent chromatin into
463 nearby functional elements [85-87, 104, 105]. However, centromeres are not typical
464 heterochromatin [106]. Here we propose that centromeres offer a unique environment for TEs
465 (Figure 8A). First, insertions are unlikely to be strongly disfavored, as the function of centromeres
466 is unlikely to be disrupted by a single insertion event. Second, centromeric insertions may escape
467 host silencing mechanisms. Small RNAs target TEs to limit potentially harmful TE activity by
468 inducing H3K9 methylation and thus transcriptional silencing (reviewed in [107-109]). However,
469 CENP-A structure and posttranslational modifications differ from H3 [106], and the host silencing
470 machinery is not known to target CENP-A. Therefore, centromeres may offer a 'safe haven' for

471 some TEs not just because they can avoid removal by purifying selection, but also because TEs
472 may temporarily escape targeting by small RNAs (*e.g.* piRNAs, siRNAs) and alternative host
473 defenses (*e.g.* DNA methylation in pericentromeres [110] hypomethylated DNA in humans [52,
474 111]).
475
476 While most TEs are considered to be parasitic elements that are neutral or mildly deleterious [112,
477 113], some TE insertions, including those in non-recombining and/or gene poor heterochromatic
478 environments, might contribute to host function [114]. For example, TEs can adopt commensal
479 and/or mutualistic strategies to propagate without direct confrontation by cooperating with their
480 host (reviewed in [115]). One extreme outcome is TE domestication, where the elements no longer
481 pose a threat to the host and instead now perform some important genome function. One often
482 cited example is at *Drosophila* telomeres, where three non-LTR retroelements in the *Jockey* clade
483 (*HeT-A*, *TAHRE*, and *TART* collectively referred to as HTT) have specialized roles [116, 117]. The
484 transposition activity of HTT is limited to telomeres, where they directly contribute to telomere
485 length and integrity [116-119]. However, detailed study of the evolutionary genetic history of the
486 retroelements across species suggests instead that there is still an underlying conflict between the
487 HTT and the genome, as there is rapid turnover of HTT between species [120]. Alternatively,
488 given the specialization of HTT, the dynamic and unstable telomeric environment may necessitate
489 rapid HTT evolution [121].
490
491 While centromeric TEs do not currently show evidence for domestication in *D. melanogaster*, they
492 may be tolerated, or even beneficial for three reasons. First, TEs may contribute to centromere
493 transcription (reviewed in [35]), as some transcription at centromeres is important for CENP-A

494 assembly and function [14, 46, 122-126] and centromere function [14, 45, 127]. Second, TEs might
495 contribute to the unique chromatin at centromeres. H3 nucleosomes are typically interspersed with
496 CENP-A within centromere domains [12, 14] [15], including a limited amount of H3K9me2/3
497 [13]. We propose that insertion of active TEs like *G2/Jockey-3* may contribute to this unique
498 CENP-A/H3 chromatin at the centromeres (Figure 8A). While landing in CENP-A nucleosomes
499 may allow the TE to temporarily escape silencing, if an active TE lands in an H3 nucleosome
500 region of the centromere, it may be targeted for silencing through H3K9 methylation (*e.g.* piRNA-
501 or siRNA-mediated silencing) and thus contribute to H3K9me2/3 domains within the centromere.
502 Third, TEs may help mitigate other conflicts that happen at centromeres. Some centromeric
503 sequences have the capacity to be selfish and bias their transmission through female meiosis in a
504 process known as meiotic drive [128]. Driving centromeres generally have expansions of satellite
505 DNA sequences that can recruit more kinetochore proteins and cause biased segregation toward
506 the egg pole [129, 130] (Figure 8B). Meiotic drive at the centromeres can lead to rapid evolution
507 of both centromere DNA and proteins that work to restore fair segregation [131, 132]. There are
508 parallel pathways to suppress meiotic drive involving rebalancing kinetochore proteins and
509 spreading heterochromatin to limit the CENP-A domain [133]. While centromere drive and
510 suppression models predict that the flanking heterochromatin domains are the main participants in
511 the conflict [133], the H3K9me2/3 domains within the centromeres might also contribute. Active
512 centromeric TEs that are targeted by small RNA pathways, like *G2/Jockey-3,* could lead to
513 spreading of H3K9me2/3 heterochromatin from within the centromere (Figure 8C). In this way,
514 the centromeric TEs and 'internal' centromeric H3K9 domains may also be a substrate for
515 suppressors of centromere drive to stop the unchecked spread of CENP-A caused by meiotic
516 drivers.

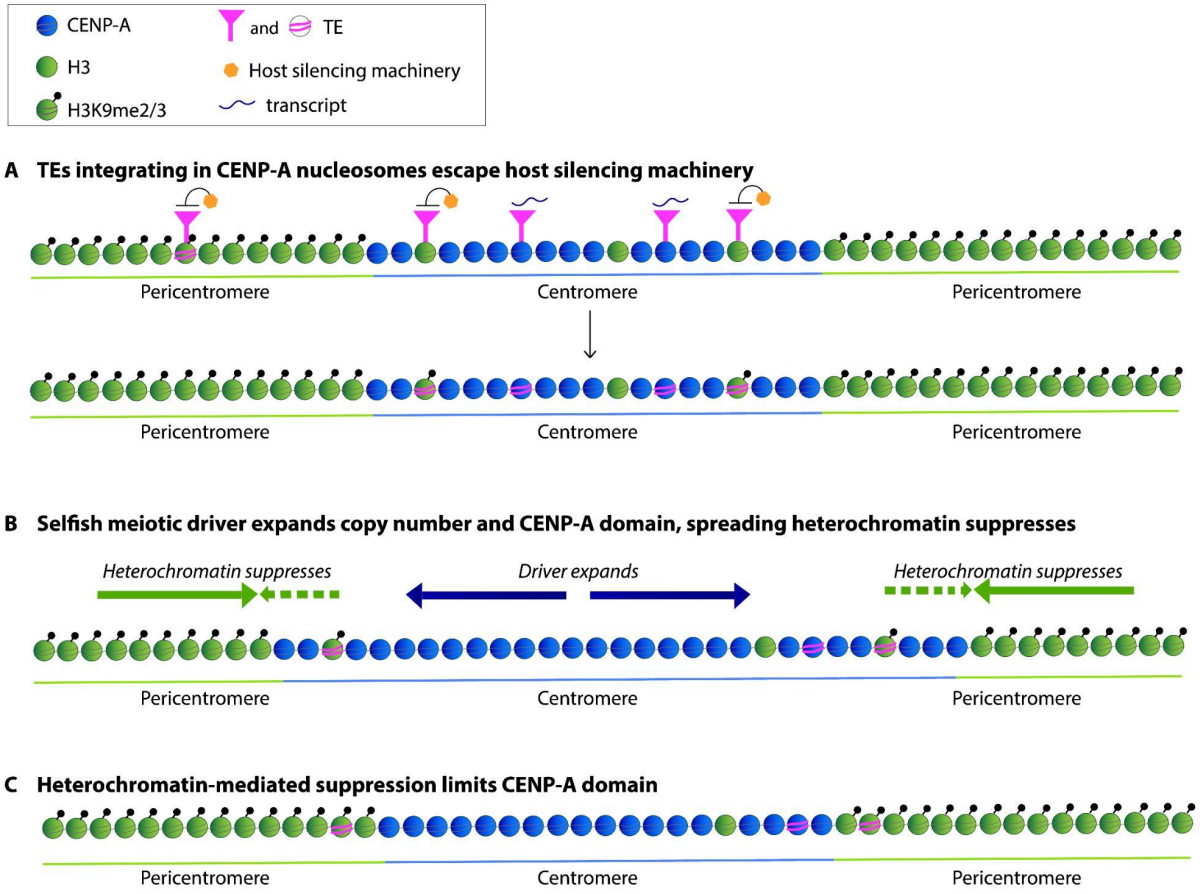
517

518 Even if TEs cooperate with centromeres, this may be an uneasy alliance as there could still be an
519 underlying conflict. For example, the invasion of H3K9me2/3 heterochromatin in centromeres is
520 potentially dangerous, as it may evict CENP-A and silence a centromere [5]. Therefore, the
521 centromeric chromatin, and the relative abundance of CENP-A and H3 nucleosomes, may mediate
522 the balance between conflict and cooperation between active TEs and centromeres.

523

524 In many species across the tree of life, centromeres are associated with TEs [18, 23-34, 53]. The
525 connection between TEs and centromeres is supported not just by TE sequences embedded in and
526 around present-day centromeres, but also in the origins of some proteins that function in
527 centromere biology (*e.g.* similarity between centromere protein CENP-B and Pogo family
528 transposons [134]). This relationship may represent a delicate balance between conflict and
529 cooperation between the centromeres and the repetitive selfish genetic elements that occupy these
530 genome regions. We expect this balance to vary depending on genomic context, level of TE
531 activity, and presence and frequency of centromere meiotic drive. If TEs contribute positively to
532 centromere function, many types of elements may be sufficient to provide the right 'environment'.
533 In this case, we would not necessarily expect to see a signature of selection on any one insertion
534 or type of insertion. While *G2/Jockey-3* is currently the principal element occupying the
535 centromere niche in *D. melanogaster*, it may be replaced over time by another type of repeat. In
536 our model, any repeat that can be transcribed and can recruit H3K9me2/3 (*e.g*. is targeted by small
537 RNAs) can contribute to centromere function. Detecting the mode, tempo, and target of selection
538 in non-recombining and dynamic environments like the centromeres is difficult. Exploring the

539 evolution of centromere composition in closely related species will provide additional insights
540 necessary to test these hypotheses.



541

542 **Figure 8. Model for the accumulation of centromere-associated retroelements.** Centromeres
543 have unique chromatin with interspersed CENP-A and H3 domains, including some H3K9me2/3
544 [12-15]. The diagrams represent nucleosomes, either CENP-A (blue) or H3 (green), with
545 H3K9me2/3 modifications indicated with a black dot. A. Active TEs like *G2/Jockey-3* jump in
546 centromeric chromatin (magenta triangles and lines in nucleosomes) and could contribute to
547 transcription (black wavy line) or be silenced by host machinery (orange hexagon). Centromeres
548 may offer a safe haven for TE insertion because insertions there are unlikely to be strongly
549 disfavored and because insertions in CENP-A nucleosomes may escape host silencing machinery

550 (*e.g.* piRNA pathway). Centromeric TE insertions might be tolerated or even favored by the host
551 if they contribute to centromere transcription or a low level of H3K9me2/3 inside centromeres. B.
552 Selfish centromeric DNA sequences can expand their copy number and bias their transmission to
553 the next generation through female meiosis [128]. C. One way to suppress cheating centromeres
554 is by spreading heterochromatin to limit the CENP-A domain [133]. TE insertions in H3
555 nucleosomes within the centromeres may be targeted for silencing by host silencing machinery
556 (green circle with magenta lines and black circle). These H3K9me2/3 nucleosomes within
557 centromeres may help create a more efficient response to driving centromeres by merging
558 spreading pericentric heterochromatin with internal H3K9me2/3.

559 **MATERIALS AND METHODS**

560

561 *G2/Jockey-3 re-annotation and expression analysis*

562

563 Chang et. al. [53] generated the *G2/Jockey-3* consensus sequence after a phylogenetic analysis
564 confirmed previously annotated *G2* and *Jockey-3* were the same element [53]. We annotated a
565 heterochromatin-enriched long-read genome assembly of *D. melanogaster* [73] with
566 Repeatmasker v4.1.0 using the previous *G2/Jockey-3* consensus sequence along with a *Drosophila*
567 repeat library [135]. We extracted *G2/Jockey-3* and the surrounding sequences using BEDTools
568 v2.29.2 [136], aligned the sequences to the previous consensus sequence, manually inspected the
569 alignment, and generated the new consensus sequence with Geneious v8.1.6 [137] (File S1). We
570 identified Open Reading Frames (ORFs) with NCBI ORFfinder
571 (https://www.ncbi.nlm.nih.gov/orffinder/) and screened the ORF candidates by identifying
572 conserved domains common in non-LTR retrotransposons with the NCBI Conserved Domains
573 search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). We defined the 5' end by
574 identifying transcription along the length of *G2/Jockey-3* by mapping publicly-available single-
575 end Precision nuclear Run-On Sequencing (PROseq) downloaded from NCBI SRA under
576 BioProject PRJNA451204 and accessions SRR7050623 - SRR7050630 [62]. We removed
577 adapters and reads < 16bp using cutadapt v2.1. Reads were reverse complemented using the fastX
578 toolkit v0.0.13 (fastx_reverse_complement). We mapped reads to the heterochromatin-enriched
579 genome assembly of *D. melanogaster* using bowtie2 v2.3.5.1 (--no-discordant --dovetail --very-
580 sensitive options; [138]). We extracted reads > 16 bp in length mapping to *G2/Jockey-3*
581 annotations in the genome with samtools v1.7 [139] and remapped the reads to the *G2/Jockey-3*

582 consensus with bowtie2 v2.3.5.1 [138]. We separated reads based on their strand using samtools
583 (v1.7). We calculated read depth at each position along the *G2/Jockey-3* consensus and normalized
584 coverage by the total number of genome mapped reads (RPM) using bedtools genomecov [136].
585 We used alignment depth of all *G2/Jockey-3* genomic fragments to identify where transcription
586 occurs along the length of *G2/Jockey-3*. We submitted the sequence in 5' end motifs under
587 transcription peaks for a GoMo analysis [140] to identify motifs as possible promoters.
588
589 To detect piRNAs corresponding to *G2/Jockey-3,* we trimmed the small RNA-seq reads for
590 adaptors and excluded the reads that align to rRNA or tRNA. We aligned the remaining reads to
591 the heterochromatin-enriched genome assembly [53, 73] using Bowtie [141], and then used a
592 customized python script as described in Wei et al. [83] to count reads that mapped to each TE
593 feature (htseq_bam_count_proportional.py;
594 https://github.com/LarracuenteLab/Dmelanogaster_satDNA_regulation; last accessed 11/2022,
595 https://doi.org/10.5061/dryad.hdr7sqvj3). We reported the RPM (Reads Per Million) values in a
596 summary table. We calculated the 10nt overlap Z-score of piRNAs using piPipes [142].
597
598 *Short-read analysis*
599
600 We detected TE insertions from publicly available FASTQ files from 135 Drosophila
601 melanogaster Genetic Reference Panel (DGRP) lines from BioProject PRJNA36679 [64, 65]
602 downloaded from the European Nucleotide Archive last accessed September 20, 2019, 13 of the
603 founder lines of the Drosophila Synthetic Population Resource (DSPR) BioProject
604 PRJNA156883[143], and 85 samples from the Global Diversity Lines BioProject PRJNA268111

605 (Table S3)[144]. We detected TE insertions in each sample with the McClintock pipeline [69] (last
606 downloaded and updated, August 20, 2020) combining several split-read and discordant read
607 mapping programs including the following: ngs_te_mapper, popoolationTE, popoolationTE2,
608 Retroseq, TE-locate, and TEMP (scripts and files for running McClintock at
609 https://github.com/LarracuenteLab/Dmel_Jockey-3_Evolution). We concatenated the outputs to
610 calculate the number of TEs detected by each program and sample and analyzed the outputs with
611 Python and R scripts available at https://github.com/LarracuenteLab/Dmel_Jockey-3_Evolution.
612 We repeated our analyses of TEs with and without insertions in known piRNA clusters and
613 removed fragments of Gypsy6_I_Dmoj and Gypsy1_I_Dmoj which resemble satellite sequences
614 in *D. melanogaster*. Depth of each short-read sample was calculated from the sorted BAM outputs
615 from the McClintock pipeline using Mosdepth [145] (Table S3). We retained samples which had
616 a centromere coverage greater than 20X for our analyses (see Supplemental Text for more
617 information).
618
619 We used the same short-read FASTQ files from the DGRP to quantify genetic diversity between
620 the centromere islands and the rest of the genome using a best practices GATK pipeline v4.1.9.0
621 to detect variant SNPs. We mapped reads from high coverage samples to our reference genome
622 with BWA v1.7.15 and samtools 1.7 to generate sorted BAM files and used GATK
623 MarkDuplicates, BaseRecalibrator, ApplyBQSR, AnalyzeCovariates, and HaplotypeCaller to call
624 and filter SNPs for each sample into a VCF file. The VCFs for all samples were concatenated with
625 GenotypeGVCFs and relicabrated with VariantRecalibrator and ApplyVQSR. We quality filtered
626 SNPs in the resulting GVCF file with the following criteria; QD < 2.0, FS > 60.0, MQ < 40.0,
627 MQRankSum < -12.5, and ReadPosRankSum < -8.0. We calculated π and Tajima's D with the

628 SimplifyVCF_Basic.v2.pl and Windows_Basic.v2.pl Perl scripts on the filtered SNPs (available
629 at https://github.com/bnavarrodominguez/SD-Population-genomics, [146]. We further analyzed
630 variants within the centromeres with VCFtools [147] and the vcfR R package [148]. Additional
631 details of our analysis with short-read data and the results can be found in the Supplemental Text
632 section.
633
634
635 *Long-read analysis*
636
637 We assembled genomes for several *D. melanogaster* lines following the heterochromatin-enriched
638 assembly protocol from Chang & Larracuente [73]. We downloaded raw PacBio reads for five of
639 the Global Diversity Lines (GDL) from the NBCI BioProject PRJNA451471 [68], eight
640 Drosophila Synthetic Population Resources (DSPR) founder lines and Oregon-R with raw Pacbio
641 reads from BioProject PRJNA418342 [67], and five DGRP lines from Bioproject PRJNA559813
642 [66]. We assembled draft genomes using all reads with Canu v1.7.1 [149] with the relaxed
643 corrected error rate and the parameters "genomeSize=160m useGrid=false
644 stopOnReadQuality=false corMinCoverage=0 corOutCoverage=100 correctedErrorRate=0.100".
645 We then mapped PacBio reads to the reference genome with minimap2 (downloaded August 28,
646 2018) with the "-x map-bp" parameter setting [150]. We then used unmapped reads and reads
647 mapping to heterochromatic regions to generate an additional "heterochromatin-enriched"
648 assembly. We merged the two assemblies into a single assembly with quickmerge [151] and
649 polished the assemblies with Racon v1.3.1 [152] by mapping the PacBio reads three times to the
650 assembly using minimap2 [150]. We also polished the GDL assemblies twice with Illumina reads

651 from BioProject PRJNA268111 [144], the DSPR assemblies with reads from
652 https://wfitch.bio.uci.edu/~dspr/index.html (last accessed September 6, 2021 [153]), BioProject
653 PRJNA252961 for Oregon-R, and the DGRP assemblies with reads from BioProject PRJNA36679
654 [64, 65]. We annotated all assemblies with Repeatmasker v4.1.0 and our *Drosophila* TE library
655 with the updated *G2/Jockey-3* consensus sequence [135] (File S10). We identified centromere
656 island sequences in assemblies using a combination of methods to find sequences corresponding
657 to the centromere islands identified in our reference genome [53]. These include aligning
658 assemblies to the centromere islands with MUMMER v.3.23 nucmer [154], identification of
659 similar sequences with BLAST 2.10.0+ [155], and identifying contigs enriched in *G2/Jockey-3* or
660 satellite DNA repeats enriched around the centromere islands [53]. We also used MEME [156] to
661 detect motifs enriched in the flanking regions of young insertions in our assemblies. Additional
662 details of our analysis with long-read data, the results, and the MEME motif analysis are in the
663 Supplemental Text.
664
665 *Insertion polymorphism validation*
666
667 We validated TE insertion predictions in three ways: 1) we used short reads to confirm TE
668 predictions in long read assemblies (see Supplemental Text); 2) we used long read assemblies to
669 confirm TE predictions in short read data (see Supplemental Text); and 3) we selected a subset to
670 validate with PCR (File S3). We used the results from these analyses to establish criteria for
671 including a predicted insertion in our analyses. For PCR, we selected a subset of 59 *G2/Jockey-3*
672 copies to confirm in four GDL lines; B59, I23, T29A, and ZH26. We included copies from each
673 chromosome (X, 2, 3, 4), each chromatin region (euchromatin, heterochromatin, and centromere),

674 and both reference and non-reference copies. We were able to detect 33 copies including 13 copies
675 in the centromere and seven *de novo* insertions (File S3).
676
677 *G2/Jockey-3 sequence and phylogenetic analysis in D. melanogaster*
678
679 We extracted all *G2/Jockey-3* sequences from the annotated assemblies using BEDTools v2.29.2
680 [136] and classified the insertions as reference or non-reference (*de novo*). We classified insertions
681 as homologous to those in our reference if they were both fragments of similar size (+/- 10 bp) and
682 within +100 bp of the same location in the reference genome. We identified corresponding
683 locations by aligning the assemblies to the reference genome with nucmer within MUMMER
684 v.3.23 [154]. Insertions were classified as *de novo* if they differed in fragment size length and
685 position of reference insertions, found outside the reference fragment position range, and located
686 within breaks in the nucmer alignment to the reference genome.
687
688 We performed phylogenetic analyses for candidate *G2/Jockey-3* sequences identified in our
689 genome assemblies. We used *G2/Jockey-3* sequences from the reference and the non-reference
690 insertions discovered in the DSPR and GDL assemblies. If a non-reference insertion was found in
691 multiple assemblies, we randomly selected an assembly from which to extract that insertion to
692 include in our analysis. Prior to phylogenetic analysis, we filtered sequences to exclude fragments
693 shorter than 400 bp or ~10% of the length of a full *G2/Jockey-3* element. We aligned the sequences
694 with MAFFT [157] with the *D. simulans G2/Jockey-3* consensus as the outgroup and inferred a
695 maximum likelihood phylogenetic tree with RAxML v8.2.11 [158]. To study the evolution of
696 *G2/Jockey-3* on the Y chromosome, we inferred a maximum likelihood phylogeny of all fragments

exclusive to the Y chromosome regardless of size. We determined the relative age of individual insertions using divergence from the consensus from the Repeatmasker output and the number of substitutions unique to a fragment divided by fragment length. "Young" insertions were either < 1% diverged from the consensus sequence or *de novo* insertions exclusive to a single genome assembly and < 2% diverged from the consensus sequence. We calculated pairwise divergence between fragments with MEGA X [159] using the Kimura 2-parameter model assuming uniform rates among all sites. We estimated gene conversion rates between elements using the same method outlined from Chang & Larracuente [73] based on Ohta's α equation [160] and Rozen's equation [161]. We used the following equation for our calculation of gene conversion:

$$c = \frac{2u(n-1)}{\pi}$$

We determined the rate at which a number of sequences (n) homogenize each other per generation (c) based on the point mutation rate (u) and divergence among the sequences (π). We grouped reference *G2/Jockey-3* sequences located within the same contig except for insertions on the Y centromere. Fragments from the Y centromere were grouped based on the phylogenetically-defined classes. We aligned the sequences with MAFFT [157] and calculated divergence with MEGA X [159]. We normalized c based on gene conversion tract length of a single event (c_g) based on previously estimated 400 bp tract length of a single event in *D. melanogaster* [92, 93]:

$$c_g = \frac{c}{400\ x\ 2}$$

Age-of-allele test

Natural selection, transposition, and frequency of insertion influence TE evolution within populations. We used an age-of-allele test [70] to determine departure from neutral evolution of individual insertions based on population frequency, age, and demographic history in *D.*

720 *melanogaster*. The age-of-allele test determines the probability distribution of an insertion
721 frequency within a population under neutrality conditioned on the age of the insertion. The age is
722 inferred by the number of unique substitutions within a TE sequence accumulated since the time
723 of the insertion. To determine the number of unique substitutions, we aligned insertions with
724 MAFFT [157] for *G2/Jockey-3*, *BS*, *G5*, *Doc*, and *Doc2* detected as reference insertions in the
725 McClintock output. Unique substitutions were quantified from each individual insertion using
726 Biostrings [162]. We calculated the probability distribution of the frequency of each unique
727 insertion with the age-of-allele test script from Blumenstiel et al. [70] in Mathematica v12.0.0.0.
728 The probability distribution is also conditioned on population demography of North American *D.*
729 *melanogaster* populations outlined in the original paper [70]. North American *D. melanogaster*
730 populations experienced two bottlenecks in their migration from Africa [163, 164] that led to non-
731 equilibrium TE evolution [70]. We compared the probability distribution to the frequency of
732 insertions detected in the 28 DGRP samples with >25X average coverage of the centromere
733 islands. We excluded insertions in the Y chromosome given the high rate of duplication. The
734 expected frequency of each insertion in the population is the probability of observing an insertion
735 as frequent or more frequent within the population multiplied by the observed number of insertions
736 in our population sample. We calculated significant deviations from neutrality as probability of
737 observing as many or fewer, or as many or greater, copies < 0.1.
738
739
740
741
742

743 **DATA AVAILABILITY STATEMENT**

744

745 All scripts for data analysis and data for figure generation are available on Github at
746 https://github.com/LarracuenteLab/Dmel_Jockey-3_Evolution. The genome assemblies, genome
747 annotations, and large datasets including the McClintock outputs will be available on Dryad (DOI
748 forthcoming).

749

750 **ACKNOWLEDGEMENTS**

751

752 This work was funded by the National Science Foundation (NSF MCB 1844693) and University
753 of Rochester funds to AML. We thank the Mellone lab, Dr. Justin Blumenstiel, and the Larracuente
754 lab for helpful discussion. We also thank the University of Rochester Center for Integrated
755 Resource Computing for access to computing cluster resources.

756

757

758 REFERENCES

759 1. Allshire RC, Karpen GH: **Epigenetic regulation of centromeric chromatin: old dogs,**
760 **new tricks?** *Nat Rev Genet* 2008, **9:**923-937.
761 2. Garrido-Ramos MA: **Satellite DNA: An Evolving Topic.** *Genes* 2017, **8**.
762 3. Hartley G, O'Neill RJ: **Centromere Repeats: Hidden Gems of the Genome.** *Genes*
763 2019, **10**.
764 4. Pidoux AL, Allshire RC: **Kinetochore and heterochromatin domains of the fission**
765 **yeast centromere.** *Chromosome Res* 2004, **12:**521-534.
766 5. Ohzeki J-I, Bergmann JH, Kouprina N, Noskov VN, Nakano M, Kimura H, Earnshaw
767 WC, Larionov V, Masumoto H: **Breaking the HAC Barrier: histone H3K9**
768 **acetyl/methyl balance regulates CENP-A assembly.** *EMBO J* 2012, **31:**2391-2402.
769 6. Choo KH: **Centromerization.** *Trends Cell Biol* 2000, **10:**182-188.
770 7. Henikoff S, Dalal Y: **Centromeric chromatin: what makes it unique?** *Curr Opin*
771 *Genet Dev* 2005, **15:**177-184.

8. Henikoff S, Furuyama T: **Epigenetic inheritance of centromeres.** *Cold Spring Harb Symp Quant Biol* 2010, **75:**51-60.
9. Black BE, Cleveland DW: **Epigenetic centromere propagation and the nature of CENP-a nucleosomes.** *Cell* 2011, **144:**471-479.
10. Mendiburo MJ, Padeken J, Fülöp S, Schepers A, Heun P: **Drosophila CENH3 is sufficient for centromere formation.** *Science* 2011, **334:**686-690.
11. McKinley KL, Cheeseman IM: **The molecular basis for centromere identity and function.** *Nat Rev Mol Cell Biol* 2016, **17:**16-29.
12. Blower MD, Sullivan BA, Karpen GH: **Conserved Organization of Centromeric Chromatin in Flies and Humans.** *Developmental Cell* 2002, **2:**319-330.
13. Ribeiro SA, Vagnarelli P, Dong Y, Hori T, McEwen BF, Fukagawa T, Flors C, Earnshaw WC: **A super-resolution map of the vertebrate kinetochore.** *Proc Natl Acad Sci U S A* 2010, **107:**10484-10489.
14. Bergmann JH, Rodríguez MG, Martins NMC, Kimura H, Kelly DA, Masumoto H, Larionov V, Jansen LET, Earnshaw WC: **Epigenetic engineering shows H3K4me2 is required for HJURP targeting and CENP-A assembly on a synthetic human kinetochore.** *EMBO J* 2011, **30:**328-340.
15. Martins NMC, Bergmann JH, Shono N, Kimura H, Larionov V, Masumoto H, Earnshaw WC: **Epigenetic engineering shows that a human centromere resists silencing mediated by H3K27me3/K9me3.** *Mol Biol Cell* 2016, **27:**177-196.
16. Lo AW, Magliano DJ, Sibson MC, Kalitsis P, Craig JM, Choo KH: **A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA.** *Genome Res* 2001, **11:**448-457.
17. Alonso A, Fritz B, Hasson D, Abrusan G, Cheung F, Yoda K, Radlwimmer B, Ladurner AG, Warburton PE: **Co-localization of CENP-C and CENP-H to discontinuous domains of CENP-A chromatin at human neocentromeres.** *Genome Biol* 2007, **8:**R148.
18. Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE, Black BE: **The octamer is the major form of CENP-A nucleosomes at human centromeres.** *Nat Struct Mol Biol* 2013, **20:**687-695.
19. Nergadze SG, Piras FM, Gamba R, Corbo M, Cerutti F, McCarter JGW, Cappelletti E, Gozzo F, Harman RM, Antczak DF, et al: **Birth, evolution, and transmission of satellite-free mammalian centromeric domains.** *Genome Res* 2018, **28:**789-799.
20. Kasai F, Garcia C, Arruga MV, Ferguson-Smith MA: **Chromosome homology between chicken (*Gallus gallus domesticus*) and the red-legged partridge (Alectoris rufa); evidence of the occurrence of a neocentromere during evolution.** *Cytogenet Genome Res* 2003, **102:**326-330.
21. Marshall OJ, Chueh AC, Wong LH, Choo KHA: **Neocentromeres: new insights into centromere structure, disease development, and karyotype evolution.** *Am J Hum Genet* 2008, **82:**261-282.
22. Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R: **Centromere repositioning in mammals.** *Heredity* 2012, **108:**59-67.
23. Miller JT, Dong F, Jackson SA, Song J, Jiang J: **Retrotransposon-related DNA sequences in the centromeres of grass chromosomes.** *Genetics* 1998, **150:**1615-1623.

816 24. Presting GG, Malysheva L, Fuchs J, Schubert I: **A Ty3/gypsy retrotransposon-like**
817 **sequence localizes to the centromeric regions of cereal chromosomes.** *Plant J* 1998,
818 **16:**721-728.
819 25. Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J:
820 **Functional rice centromeres are marked by a satellite repeat and a centromere-**
821 **specific retrotransposon.** *Plant Cell* 2002, **14:**1691-1704.
822 26. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR,
823 Jiang J: **Sequencing of a rice centromere uncovers active genes.** *Nat Genet* 2004,
824 **36:**138-145.
825 27. Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones
826 KM, Dawe RK, et al: **Molecular and cytological analyses of large tracks of**
827 **centromeric DNA reveal the structure and evolutionary dynamics of maize**
828 **centromeres.** *Genetics* 2003, **163:**759-770.
829 28. Ferreri GC, Brown JD, Obergfell C, Jue N, Finn CE, O'Neill MJ, O'Neill RJ: **Recent**
830 **amplification of the kangaroo endogenous retrovirus, KERV, limited to the**
831 **centromere.** *J Virol* 2011, **85:**4761-4771.
832 29. Ferreri GC, Marzelli M, Rens W, O'Neill RJ: **A centromere-specific retroviral element**
833 **associated with breaks of synteny in macropodine marsupials.** *Cytogenet Genome Res*
834 2004, **107:**115-118.
835 30. de Sotero-Caio CG, Cabral-de-Mello DC, Calixto MdS, Valente GT, Martins C, Loreto
836 V, de Souza MJ, Santos N: **Centromeric enrichment of LINE-1 retrotransposons and**
837 **its significance for the chromosome evolution of Phyllostomid bats.** *Chromosome Res*
838 2017, **25:**313-325.
839 31. Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer
840 TJ, Herrero J, Roos C, Aken B, et al: **Gibbon genome and the fast karyotype evolution**
841 **of small apes.** *Nature* 2014, **513:**195-201.
842 32. Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DIK, Rocchi M, Capozzi
843 O, Archidiacono N, Konkel MK, Walker JA, et al: **Centromere remodeling in *Hoolock***
844 ***leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons.**
845 *Genome Biol Evol* 2012, **4:**648-658.
846 33. Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez
847 MJ, Kappes F, Dube D, Chan SM, Robinson D, Meng F, et al: **HIV infection reveals**
848 **widespread expansion of novel centromeric human endogenous retroviruses.**
849 *Genome Res* 2013, **23:**1505-1513.
850 34. Zahn J, Kaplan MH, Fischer S, Dai M, Meng F, Saha AK, Cervantes P, Chan SM, Dube
851 D, Omenn GS, et al: **Expansion of a novel endogenous retrovirus throughout the**
852 **pericentromeres of modern humans.** *Genome Biol* 2015, **16:**74.
853 35. Klein SJ, O'Neill RJ: **Transposable elements: genome innovation, chromosome**
854 **diversity, and centromere conflict.** *Chromosome Res* 2018, **26:**5-23.
855 36. Birchler JA, Presting GG: **Retrotransposon insertion targeting: a mechanism for**
856 **homogenization of centromere sequences on nonhomologous chromosomes.** *Genes*
857 *Dev* 2012, **26:**638-640.
858 37. Sultana T, Zamborlini A, Cristofari G, Lesage P: **Integration site selection by**
859 **retroviruses and transposable elements in eukaryotes.** *Nature Reviews Genetics* 2017,
860 **18:**292-308.

38. Wong LH, Choo KHA: **Evolutionary dynamics of transposable elements at the centromere.** *Trends Genet* 2004, **20:**611-616.
39. Sharma A, Wolfgruber TK, Presting GG: **Tandem repeats derived from centromeric retrotransposons.** *BMC Genomics* 2013, **14**.
40. Bourc'his D, Bestor TH: **Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L.** *Nature* 2004, **431:**96-99.
41. Kato M, Takashima K, Kakutani T: **Epigenetic control of CACTA transposon mobility in Arabidopsis thaliana.** *Genetics* 2004, **168:**961-969.
42. Ariumi Y: **Guardian of the Human Genome: Host Defense Mechanisms against LINE-1 Retrotransposition.** *Front Chem* 2016, **4:**28.
43. Chan FL, Wong LH: **Transcription in the maintenance of centromere chromatin identity.** *Nucleic Acids Res* 2012, **40:**11178-11188.
44. Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL, Hannan R, Saffery R, Shaw ML, Williams E, Choo KHA: **Centromere RNA is a key component for the assembly of nucleoproteins at the nucleolus and centromere.** *Genome Res* 2007, **17:**1146-1160.
45. Quénet D, Dalal Y: **A long non-coding RNA is required for targeting centromeric protein A to the human centromere.** *Elife* 2014, **3:**e03254.
46. Catania S, Pidoux AL, Allshire RC: **Sequence features and transcriptional stalling within centromere DNA promote establishment of CENP-A chromatin.** *PLoS Genet* 2015, **11:**e1004986.
47. Chen C-C, Bowers S, Lipinszki Z, Palladino J, Trusiak S, Bettini E, Rosin L, Przewloka MR, Glover DM, O'Neill RJ, Mellone BG: **Establishment of Centromeric Chromatin by the CENP-A Assembly Factor CAL1 Requires FACT-Mediated Transcription.** *Dev Cell* 2015, **34:**73-84.
48. Hall LE, Mitchell SE, O'Neill RJ: **Pericentric and centromeric transcription: a perfect balance required.** *Chromosome Research* 2012, **20:**535-546.
49. Talbert PB, Henikoff S: **Transcribing Centromeres: Noncoding RNAs and Kinetochore Assembly.** *Trends in Genetics* 2018, **34:**587-599.
50. Chuong EB, Elde NC, Feschotte C: **Regulatory activities of transposable elements: from conflicts to benefits.** *Nat Rev Genet* 2017, **18:**71-86.
51. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF, Akeson M, Miga KH: **Linear assembly of a human centromere on the Y chromosome.** *Nature Biotechnology* 2018, **36:**321-323.
52. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, et al: **The structure, function and evolution of a complete human chromosome 8.** *Nature* 2021, **593:**101-107.
53. Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen C-C, Erceg J, Beliveau BJ, Wu C-T, et al: **Islands of retroelements are major components of Drosophila centromeres.** *PLoS Biol* 2019, **17:**e3000241.
54. Belser C, Baurens F-C, Noel B, Martin G, Cruaud C, Istace B, Yahiaoui N, Labadie K, Hřibová E, Doležel J, et al: **Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing.** *Commun Biol* 2021, **4:**1047.
55. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al: **High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long Reads.** *Genomics Proteomics Bioinformatics* 2021.

56. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A, Mandakova T, Jamge B, Lambing C, Kuo P, et al: **The genetic and epigenetic landscape of the Arabidopsis centromeres.** *Science* 2021, **374:**eabi7489.
57. Hales KG, Korey CA, Larracuente AM, Roberts DM: **Genetics on the Fly: A Primer on the Drosophila Model System.** *Genetics* 2015, **201:**815-842.
58. Rahman R, Chirn G-W, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC: **Unique transposon landscapes are pervasive across Drosophila melanogaster genomes.** *Nucleic Acids Res* 2015, **43:**10655-10672.
59. Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, et al: **Comparative analysis of the transcriptome across distant species.** *Nature* 2014, **512:**445-448.
60. Andersen PR, Tirian L, Vunjak M, Brennecke J: **A heterochromatin-dependent transcription machinery drives piRNA expression.** *Nature* 2017, **549:**54-59.
61. Laktionov PP, Maksimov DA, Romanov SE, Antoshina PA, Posukh OV, White-Cooper H, Koryakov DE, Belyakin SN: **Genome-wide analysis of gene regulation mechanisms during Drosophila spermatogenesis.** *Epigenetics Chromatin* 2018, **11:**14.
62. Tettey TT, Gao X, Shao W, Li H, Story BA, Chitsazan AD, Glaser RL, Goode ZH, Seidel CW, Conaway RC, et al: **A Role for FACT in RNA Polymerase II Promoter-Proximal Pausing.** *Cell Rep* 2019, **27:**3770-3779.e3777.
63. Metcalfe CJ, Casane D: **Modular organization and reticulate evolution of the ORF1 of Jockey superfamily transposable elements.** *Mobile DNA* 2014, **5:**19.
64. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF, et al: **Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines.** *Genome Res* 2014, **24:**1193-1208.
65. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al: **The *Drosophila melanogaster* Genetic Reference Panel.** *Nature* 2012, **482:**173-178.
66. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J: **Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila.** *Nat Commun* 2022, **13:**1948.
67. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD: **Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits.** *Nat Commun* 2019, **10:**4872.
68. Long E, Evans C, Chaston J, Udall JA: **Genomic Structural Variations Within Five Continental Populations of *Drosophila melanogaster*.** *G3: Genes|Genomes|Genetics* 2018, **8:**3247-3253.
69. Nelson MG, Linheiro RS, Bergman CM: **McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data.** *G3* 2017, **7:**2763-2778.
70. Blumenstiel JP, Chen X, He M, Bergman CM: **An age-of-allele test of neutrality for transposable element insertions.** *Genetics* 2014, **196:**523-538.
71. Kofler R, Betancourt AJ, Schlötterer C: **Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*.** *PLoS Genet* 2012, **8:**e1002487.

953 72. Kofler R, Nolte V, Schlötterer C: **Tempo and Mode of Transposable Element Activity**
954 **in Drosophila.** *PLoS Genet* 2015, **11:**e1005406.
955 73. Chang C-H, Larracuente AM: **Heterochromatin-Enriched Assemblies Reveal the**
956 **Sequence and Organization of the Y Chromosome.** *Genetics* 2019, **211:**333-348.
957 74. Ghildiyal M, Zamore PD: **Small silencing RNAs: an expanding universe.** *Nat Rev*
958 *Genet* 2009, **10:**94-108.
959 75. Ernst C, Odom DT, Kutter C: **The emergence of piRNAs against transposon invasion**
960 **to preserve mammalian genome integrity.** *Nature Communications* 2017, **8**.
961 76. Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M: **Recurrent insertion and**
962 **duplication generate networks of transposable element sequences in the *Drosophila***
963 ***melanogaster* genome.** *Genome Biol* 2006, **7:**R112.
964 77. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ:
965 **Discrete small RNA-generating loci as master regulators of transposon activity in**
966 **Drosophila.** *Cell* 2007, **128:**1089-1103.
967 78. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H,
968 Siomi MC: **A slicer-mediated mechanism for repeat-associated siRNA 5' end**
969 **formation in Drosophila.** *Science* 2007, **315:**1587-1590.
970 79. Mohn F, Sienski G, Handler D, Brennecke J: **The rhino-deadlock-cutoff complex**
971 **licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila.**
972 *Cell* 2014, **157:**1364-1379.
973 80. Parhad SS, Tu S, Weng Z, Theurkauf WE: **Adaptive Evolution Leads to Cross-Species**
974 **Incompatibility in the piRNA Transposon Silencing Machinery.** *Dev Cell* 2017,
975 **43:**60-70.e65.
976 81. Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie
977 WR, Hannon GJ, Evgen'ev MB: **Small RNA-based silencing strategies for transposons**
978 **in the process of invading Drosophila species.** *RNA* 2010, **16:**1634-1645.
979 82. Quénerch'du E, Anand A, Kai T: **The piRNA pathway is developmentally regulated**
980 **during spermatogenesis in Drosophila.** *RNA* 2016, **22:**1044-1054.
981 83. Wei X, Eickbush DG, Speece I, Larracuente AM: **Heterochromatin-dependent**
982 **transcription of satellite DNAs in the female germline.** *Elife* 2021, **10**.
983 84. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J: **Population genomics**
984 **of transposable elements in *Drosophila melanogaster*.** *Mol Biol Evol* 2011, **28:**1633-
985 1644.
986 85. Montgomery E, Charlesworth B, Langley CH: **A test for the role of natural selection in**
987 **the stabilization of transposable element copy number in a population of *Drosophila***
988 ***melanogaster.*** *Genetical Research* 1987, **49:**31-41.
989 86. Charlesworth B, Campos JL: **The relations between recombination rate and patterns**
990 **of molecular variation and evolution in Drosophila.** *Annu Rev Genet* 2014, **48:**383-
991 403.
992 87. Kent TV, Uzunović J, Wright SI: **Coevolution between transposable elements and**
993 **recombination.** *Philos Trans R Soc Lond B Biol Sci* 2017, **372**.
994 88. Bergman CM, Bensasson D: **Recent LTR retrotransposon insertion contrasts with**
995 **waves of non-LTR insertion since speciation in *Drosophila melanogaster*.** *Proc Natl*
996 *Acad Sci U S A* 2007, **104:**11340-11345.
997 89. Blumenstiel JP, Hartl DL, Lozovsky ER: **Patterns of insertion and deletion in**
998 **contrasting chromatin domains.** *Mol Biol Evol* 2002, **19:**2211-2225.

90. Chang CH, Gregory LE, Gordon KE, Meiklejohn CD, Larracuente AM: **Unique structure and positive selection promote the rapid divergence of Drosophila Y chromosomes.** *Elife* 2022, **11**.
91. Comeron JM, Ratnappan R, Bailin S: **The many landscapes of recombination in Drosophila melanogaster.** *PLoS Genet* 2012, **8:**e1002905.
92. Miller DE, Smith CB, Kazemi NY, Cockrell AJ, Arvanitakis AV, Blumenstiel JP, Jaspersen SL, Hawley RS: **Whole-Genome Analysis of Individual Meiotic Events in *Drosophila melanogaster* Reveals That Noncrossover Gene Conversions Are Insensitive to Interference and the Centromere Effect.** *Genetics* 2016, **203:**159-171.
93. Miller DE, Takeo S, Nandanan K, Paulson A, Gogol MM, Noll AC, Perera AG, Walton KN, Gilliland WD, Li H, et al: **A Whole-Chromosome Analysis of Meiotic Recombination in *Drosophila melanogaster*.** *G3* 2012, **2:**249-260.
94. Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore PD, Weng Z, Theurkauf WE: **Adaptation to P Element Transposon Invasion in *Drosophila melanogaster*.** *Cell* 2011, **147:**1551-1563.
95. Kofler R: **Dynamics of Transposable Element Invasions with piRNA Clusters.** *Mol Biol Evol* 2019, **36:**1457-1472.
96. Said I, McGurk MP, Clark AG, Barbash DA: **Patterns of piRNA Regulation in Drosophila Revealed through Transposable Element Clade Inference.** *Mol Biol Evol* 2022, **39**.
97. Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C: **Molecular dissection of a natural transposable element invasion.** *Genome Res* 2018, **28:**824-835.
98. Anxolabéhère D, Kidwell MG, Periquet G: **Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile P elements.** *Mol Biol Evol* 1988, **5:**252-269.
99. Hill T, Schlotterer C, Betancourt AJ: **Hybrid Dysgenesis in *Drosophila simulans* Associated with a Rapid Invasion of the P-Element.** *PLoS Genet* 2016, **12:**e1005920.
100. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee H, Xu R, et al: **Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons.** *PLoS Genet* 2009, **5:**e1000743.
101. Schneider KL, Xie Z, Wolfgruber TK, Presting GG: **Inbreeding drives maize centromere evolution.** *Proc Natl Acad Sci U S A* 2016, **113:**E987-996.
102. Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A, Tarutani Y, Kakutani T: **Centromere-targeted de novo integrations of an LTR retrotransposon of *Arabidopsis lyrata*.** *Genes Dev* 2012, **26:**705-713.
103. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T: **Bursts of retrotransposition reproduced in Arabidopsis.** *Nature* 2009, **461:**423-426.
104. Hollister JD, Gaut BS: **Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression.** *Genome Res* 2009, **19:**1419-1428.
105. Lee YCG: **The Role of piRNA-Mediated Epigenetic Silencing in the Population Dynamics of Transposable Elements in *Drosophila melanogaster*.** *PLoS Genet* 2015, **11:**e1005269.
106. Srivastava S, Foltz DR: **Posttranslational modifications of CENP-A: marks of distinction.** *Chromosoma* 2018, **127:**279-290.

1045 107. Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ:
1046 **piRNA-Guided Genome Defense: From Biogenesis to Silencing.** *Annu Rev Genet*
1047 2018, **52:**131-157.
1048 108. Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD: **PIWI-interacting RNAs:**
1049 **small RNAs with big functions.** *Nat Rev Genet* 2019, **20:**89-108.
1050 109. Onishi R, Yamanaka S, Siomi MC: **piRNA- and siRNA-mediated transcriptional**
1051 **repression in Drosophila, mice, and yeast: new insights and biodiversity.** *EMBO Rep*
1052 2021, **22:**e53062.
1053 110. Scelfo A, Fachinetti D: **Keeping the Centromere under Control: A Promising Role**
1054 **for DNA Methylation.** *Cells* 2019, **8**.
1055 111. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ,
1056 Uralsky L, Ryabov FD, Shew CJ, et al: **Complete genomic and epigenetic maps of**
1057 **human centromeres.** *Science* 2022, **376:**eabl4178.
1058 112. Doolittle WF, Ford Doolittle W, Sapienza C: **Selfish genes, the phenotype paradigm**
1059 **and genome evolution.** *Nature* 1980, **284:**601-603.
1060 113. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284:**604-607.
1061 114. Dimitri P, Junakovic N: **Revising the selfish DNA hypothesis: new evidence on**
1062 **accumulation of transposable elements in heterochromatin.** *Trends Genet* 1999,
1063 **15:**123-124.
1064 115. Cosby RL, Chang N-C, Feschotte C: **Host-transposon interactions: conflict,**
1065 **cooperation, and cooption.** *Genes Dev* 2019, **33:**1098-1116.
1066 116. Danilevskaya ON, Slot F, Traverse KL, Hogan NC, Pardue ML: **Drosophila telomere**
1067 **transposon HeT-A produces a transcript with tightly bound protein.** *Proc Natl Acad*
1068 *Sci U S A* 1994, **91:**6679-6682.
1069 117. Mason JM, Biessmann H: **The unusual telomeres of Drosophila.** *Trends Genet* 1995,
1070 **11:**58-62.
1071 118. Pardue M-L, DeBaryshe PG: **Drosophila telomeres: A variation on the telomerase**
1072 **theme.** *Fly* 2008, **2:**101-110.
1073 119. Pardue ML, Debaryshe PG: **Drosophila telomere transposons: genetically active**
1074 **elements in heterochromatin.** *Genetica* 2000, **109:**45-52.
1075 120. Saint-Leandre B, Nguyen SC, Levine MT: **Diversification and collapse of a telomere**
1076 **elongation mechanism.** *Genome Res* 2019, **29:**920-931.
1077 121. McGurk MP, Dion-Côté A-M, Barbash DA: **Rapid evolution at the Drosophila**
1078 **telomere: transposable element dynamics at an intrinsically unstable locus.** *Genetics*
1079 2021, **217**.
1080 122. Moazed D, Bühler M, Buker SM, Colmenares SU, Gerace EL, Gerber SA, Hong EJE,
1081 Motamedi MR, Verdel A, Villén J, Gygi SP: **Studies on the mechanism of RNAi-**
1082 **dependent heterochromatin assembly.** *Cold Spring Harb Symp Quant Biol* 2006,
1083 **71:**461-471.
1084 123. Choi ES, Strålfors A, Castillo AG, Durand-Dubief M, Ekwall K, Allshire RC:
1085 **Identification of noncoding transcripts from within CENP-A chromatin at fission**
1086 **yeast centromeres.** *J Biol Chem* 2011, **286:**23600-23607.
1087 124. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA, Wong LH: **Active**
1088 **transcription and essential role of RNA polymerase II at the centromere during**
1089 **mitosis.** *Proc Natl Acad Sci U S A* 2012, **109:**1979-1984.

1090 125. Rošić S, Köhler F, Erhardt S: **Repetitive centromeric satellite RNA is essential for**
1091 **kinetochore formation and cell division.** *J Cell Biol* 2014, **207:**335-349.
1092 126. Bobkov GOM, Gilbert N, Heun P: **Centromere transcription allows CENP-A to**
1093 **transit from chromatin association to stable incorporation.** *J Cell Biol* 2018,
1094 **217:**1957-1972.
1095 127. Ohkuni K, Kitagawa K: **Endogenous transcription at the centromere facilitates**
1096 **centromere activity in budding yeast.** *Curr Biol* 2011, **21:**1695-1703.
1097 128. Sandler L, Novitski E: **Meiotic Drive as an Evolutionary Force.** *The American*
1098 *Naturalist* 1957, **91:**105-110.
1099 129. Akera T, Chmátal L, Trimm E, Yang K, Aonbangkhen C, Chenoweth DM, Janke C,
1100 Schultz RM, Lampson MA: **Spindle asymmetry drives non-Mendelian chromosome**
1101 **segregation.** *Science* 2017, **358:**668-672.
1102 130. Akera T, Trimm E, Lampson MA: **Molecular Strategies of Meiotic Cheating by Selfish**
1103 **Centromeres.** *Cell* 2019, **178:**1132-1144.e1110.
1104 131. Henikoff S, Ahmad K, Malik HS: **The centromere paradox: stable inheritance with**
1105 **rapidly evolving DNA.** *Science* 2001, **293:**1098-1102.
1106 132. Malik HS, Henikoff S: **Adaptive Evolution of Cid, a Centromere-Specific Histone in**
1107 **Drosophila.** *Genetics* 2001, **157:**1293-1298.
1108 133. Kumon T, Ma J, Akins RB, Stefanik D, Nordgren CE, Kim J, Levine MT, Lampson MA:
1109 **Parallel pathways for recruiting effector proteins determine centromere drive and**
1110 **suppression.** *Cell* 2021, **184:**4904-4918.e4911.
1111 134. Kipling D, Warburton PE: **Centromeres, CENP-B and Tigger too.** *Trends Genet* 1997,
1112 **13:**141-145.
1113 135. Khost DE, Eickbush DG, Larracuente AM: **Single-molecule sequencing resolves the**
1114 **detailed structure of complex satellite DNA loci in *Drosophila melanogaster*.** *Genome*
1115 *Res* 2017, **27:**709-721.
1116 136. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
1117 **features.** *Bioinformatics* 2010, **26:**841-842.
1118 137. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper
1119 A, Markowitz S, Duran C, et al: **Geneious Basic: an integrated and extendable**
1120 **desktop software platform for the organization and analysis of sequence data.**
1121 *Bioinformatics* 2012, **28:**1647-1649.
1122 138. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods*
1123 2012, **9:**357-359.
1124 139. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
1125 Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format**
1126 **and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
1127 140. Buske FA, Bodén M, Bauer DC, Bailey TL: **Assigning roles to DNA regulatory motifs**
1128 **using comparative genomics.** *Bioinformatics* 2010, **26:**860-866.
1129 141. Langmead B: **Aligning short sequencing reads with Bowtie.** *Curr Protoc*
1130 *Bioinformatics* 2010, **Chapter 11:**Unit 11.17.
1131 142. Han BW, Wang W, Zamore PD, Weng Z: **piPipes: a set of pipelines for piRNA and**
1132 **transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq,**
1133 **ChIP-seq and genomic DNA sequencing.** *Bioinformatics* 2015, **31:**593-595.
1134 143. Long AD, Macdonald SJ, King EG: **Dissecting complex traits using the Drosophila**
1135 **Synthetic Population Resource.** *Trends Genet* 2014, **30:**488-495.

144. Grenier JK, Roman Arguello J, Moreira MC, Gottipati S, Mohammed J, Hackett SR, Boughton R, Greenberg AJ, Clark AG: **Global Diversity Lines–A Five-Continent Reference Panel of Sequenced *Drosophila melanogaster* Strains.** *G3: Genes|Genomes|Genetics* 2015, **5:**593-603.
145. Pedersen BS, Quinlan AR: **Mosdepth: quick coverage calculation for genomes and exomes.** *Bioinformatics* 2018, **34:**867-868.
146. Navarro-Dominguez B, Chang C-H, Brand CL, Muirhead CA, Presgraves DC, Larracuente AM: **Epistatic selection on a selfish supergene - drive, recombination, and genetic load.** *Elife* 2022, **11**.
147. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.** *Bioinformatics* 2011, **27:**2156-2158.
148. Knaus BJ, Grünwald NJ: **vcfr: a package to manipulate and visualize variant call format data in R.** *Mol Ecol Resour* 2017, **17:**44-53.
149. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive -mer weighting and repeat separation.** *Genome Res* 2017, **27:**722-736.
150. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.
151. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ: **Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage.** *Nucleic Acids Res* 2016, **44:**e147.
152. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from long uncorrected reads.** *Genome Res* 2017, **27:**737-746.
153. King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD, Macdonald SJ: **Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource.** *Genome Res* 2012, **22:**1558-1566.
154. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
155. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
156. Bailey TL, Johnson J, Grant CE, Noble WS: **The MEME Suite.** *Nucleic Acids Research* 2015, **43:**W39-W49.
157. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.
158. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.
159. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: **MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.** *Mol Biol Evol* 2018, **35:**1547-1549.
160. Ohta T: **Allelic and nonallelic homology of a supergene family.** *Proc Natl Acad Sci U S A* 1982, **79:**3251-3254.
161. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC: **Abundant gene conversion between arms of palindromes in human and ape Y chromosomes.** *Nature* 2003, **423:**873-876.
162. **Biostrings: Efficient manipulation of biological strings.** [https://bioconductor.org/packages/Biostrings]

1182 163. Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S: **Demographic inference reveals**
1183 **African and European admixture in the North American *Drosophila melanogaster***
1184 **population.** *Genetics* 2013, **193:**291-301.
1185 164. Li H, Stephan W: **Inferring the demographic history and rate of adaptive**
1186 **substitution in Drosophila.** *PLoS Genet* 2006, **2:**e166.
1187