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

ABSTRACT

Centromeres are chromosomal regions essential for coordinating chromosome segregation during cell division. While centromeres are defined by the presence of a centromere-specific histone H3 variant rather than a particular DNA sequence, they are typically embedded in repeat-dense chromosomal genome regions. In many species, centromeres are associated with transposable elements, but it is unclear if these elements are selfish or if they play a role in centromere specification or function. Here we use *Drosophila melanogaste*r as a model to understand the evolution of centromere-associated transposable elements. *G2/Jockey-3* is a non-LTR retroelement in the *Jockey* clade and the only sequence shared by all centromeres. We study the evolution of *G2/Jockey-3* using short and long read population genomic data to infer insertion polymorphisms across the genome. We combine estimates of the age, frequency, and location of insertions to infer the evolutionary processes shaping *G2/Jockey-3* and its association with the centromeres. We find that *G2/Jockey-3* is an active retroelement targeted by the piRNA pathway that is enriched in centromeres at least in part due to an insertion bias. We do not detect signatures of positive selection on any *G2/Jockey-3* insertions that would suggest than individual copies are favored by natural selection. Instead, we infer that most insertions are neutral or weakly deleterious both inside and outside of the centromeres. Therefore, *G2/Jockey-3* evolution is consistent with it being a selfish genetic element that targets centromeres. We propose that targeting centromeres helps active retroelements escape host defenses, as the unique centromeric chromatin may prevent targeting by the host silencing machinery. At the same time, centromeric TEs insertions may be tolerated or even beneficial if they also contribute to the transcriptional and chromatin environment. Thus, we suspect centromere-associated retroelements like *G2/Jockey-3* reflect a balance between conflict and cooperation at the centromeres.

47 **INTRODUCTION**

48

49 Centromeres are important chromosomal structures for kinetochore assembly and proper
50 chromosome segregation during cell division. While centromeres are usually in highly repetitive
51 regions, the role of DNA sequences in centromere function or establishment remains unclear
52 (reviewed in [1]). Across many organisms, centromeres are associated with satellite DNAs
53 (satDNAs), large blocks of tandemly repeated DNA sequences (reviewed in [2, 3]). Centromere-
54 associated tandem repeats are capable of initiating centromere assembly in fission yeast [4] and
55 humans [5]. However, centromeres are not defined by any particular DNA sequences. They are
56 instead defined epigenetically by CENP-A, a specific variant of histone H3 [6-11]. Centromere
57 domains consist of unique chromatin where H3 nucleosomes are interspersed with CENP-A
58 nucleosomes [12-15]. Repeat sequences are not always required for centromere formation, as
59 neocentromeres can form outside of centromeric DNA in the absence of satDNA sequences [16-
60 18] and some species have centromeres devoid of repeats (*e.g.* satellite-free centromeres in the
61 Equus genus [19]). Neocentromeres have the potential to replace older centromeres with
62 implications for karyotype evolution and speciation [20-22].

63

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

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

93 located in centromeres; about 40% of all insertions are found outside of the centromere islands
94 [53]. Polymorphic *G2/Jockey-3* insertions [58] and transcripts from published RNAseq datasets
95 [59-61] suggest this element may still be active in *D. melanogaster* [53]. *G2/Jockey-3* is also
96 found at the centromeres in *D. simulans*, a close relative of *D. melanogaster,* suggesting that its
97 centromere association may be broadly conserved [53].
98
99 It is unclear whether *G2/Jockey-3* is an important centromere component in *D. melanogaster* or
100 if they are instead purely selfish TEs. If *G2/Jockey-3* are important for centromere specification
101 or function, we may detect evidence for positive natural selection on centromeric copies.
102 Alternatively, *G2/Jockey-3* may be a selfish element that targets insertions in centromeres but are
103 otherwise selectively neutral or weakly deleterious. These hypotheses are not necessarily
104 mutually exclusive. Here we investigated the age and frequency of *G2/Jockey-3* insertions across
105 *D. melanogaster* genomes to determine if its evolution is consistent with a role in centromere
106 function or underlying selfish behavior. Our analyses suggest that *G2/Jockey-3* insertions are
107 largely neutral or weakly deleterious both inside and outside centromeres, suggesting that they
108 are selfish genetic elements. Our results also suggest that *G2/Jockey-3* has had a weak insertion
109 bias toward centromeres. Although there is no evidence for positive selection on individual
110 *G2/Jockey-3* centromeric insertions, insertion of active TEs like *G2/Jockey-3* may still contribute
111 to transcription and the unique chromatin environment at centromeres. We propose that there
112 may be a balance between conflict and cooperation between active TEs like *G2/Jockey-3* and
113 centromeres.
114
115

116 **RESULTS**

117

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

119

120 Prior to analyzing the distribution of *G2/Jockey-3* insertion polymorphisms in *D. melanogaster*
121 populations, we first confirmed the consensus annotation. We analyzed all *G2/Jockey-3*
122 insertions in our *D. melanogaster* genome assembly [53] and found two lines of evidence that the
123 annotation needed updating: 1) 25 of the 167 *G2/Jockey-3* sequences in the genome shared up to
124 92 bp of upstream sequence at the 5'-end of the annotation; and 2) our reanalysis of PROseq data
125 [62] suggests that *G2/Jockey-3* transcription starts 92 bp upstream of the published start site
126 (Figure S1A). This upstream region is also enriched for GO terms for RNA polymerase II
127 transcription factors and other transcription factor activity (Figure S1B). We therefore extended
128 the original *G2/Jockey-3* consensus sequence by 92 bases (File S1).

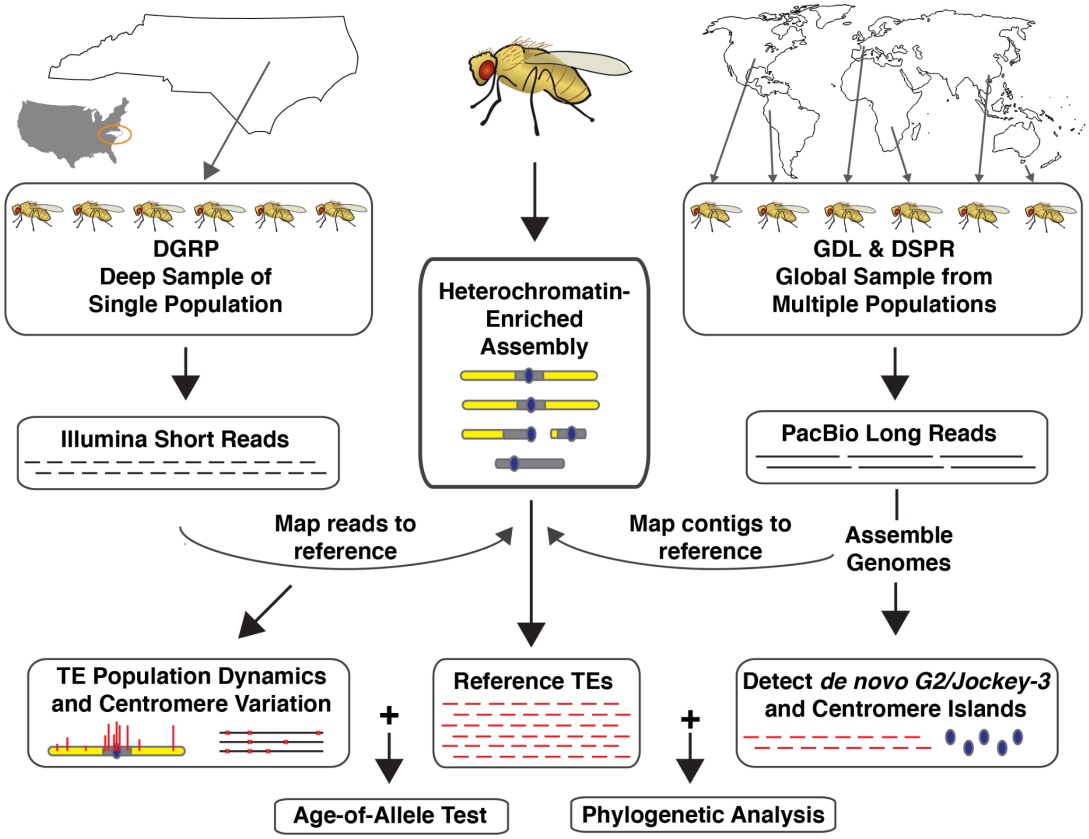
129

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

139 Superfamily Accession cl00490, E-value 7.21E-18) and RT (RT_like Superfamily Accession
140 cd01650, E-value 4.62E-56). *G2/Jockey-3* ends in a poly(A) signal.
141
142 *2. Overview of data sets analyzed*
143
144 We studied the evolutionary dynamics of *G2/Jockey-3* and other TEs by investigating the
145 patterns of insertion polymorphisms both in a single panmictic population, and in individuals
146 from a wide geographic range (Figure 1, Table S1). We conducted a survey of TE dynamics and
147 genetic variation in 135 samples from the *Drosophila melanogaster* Genetic Reference Panel
148 (DGRP) collected from a single population in North Carolina with whole genome Illumina
149 sequence data [64, 65] and a subset of long read data from the same population for validation
150 [66] (see Supplemental Text). For the samples from a wide geographic range, we used both long-
151 and short-read data from 13 founder lines of the *Drosophila* Synthetic Population Resource
152 (DSPR) and Oregon-R [67]. We also analyzed long-read PacBio data from five Global Diversity
153 Lines (GDL) samples from different continents [68]. We generated *de novo* genome assemblies
154 of each line to determine the organization of the centromere islands originally described in the
155 reference genome for comparison (Table S2) and predicted insertions in short read data with the
156 McClintock pipeline [69]. We assessed the reliability of our TE insertion predictions by looking
157 for concordance between short and long read predictions (see Supplemental Text) and through
158 PCR validation of a subset of insertions. We confirmed 64% (67 / 105) of the *G2/Jockey-3*
159 insertions detected by McClintock are found in the long-read samples we examined (File S2).
160 PCR confirmed 80% (28 / 35) of *G2/Jockey-3* insertions predicted in the GDL lines, and ~67%
161 (12 / 18, File S3) of the centromere insertions. Based on our assessment, we limit our analysis to

162 short read datasets with >20X centromere coverage (Table S3) and TEs predicted by more than

163 one TE caller implemented in McClintock (Supplemental Text).



164

165 **Figure 1: The summary of data and analyses using one source or a combination of sources**

166 **in this paper.** A heterochromatin-enriched assembly was originally used to discover centromere

167 islands in Chang et al. [53] and identify reference TE insertions. We used short reads from the

168 DGRP to identify reference and *de novo* insertions to study TE dynamics within a single

169 population from North Carolina (map source: R package usmaps https://cran.r-

170 project.org/web/packages/usmap/). We generated genome assemblies for 18 samples from global

171 populations to uncover the centromere island sequences and *de novo* elements (map source:

172 www.outline-world-map.com). We combined the reference sequences with population

173 frequencies to infer selection with an age-of-allele test [70]. We also combined reference TE

174 sequences with *de novo* elements uncovered in our assemblies for phylogenetic analysis. The fly

175 illustrations are modified images from Wikipedia Commons

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

177

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

179

180 Previous studies of TE polymorphism in *D. melanogaster* are consistent with *G2/Jockey-3*

181 (previously annotated as *G2*) being recently active [58, 71, 72]. To determine if *G2/Jockey-3* is

182 active in the DGRP, we detected TE insertion polymorphisms with the McClintock pipeline [69]

183 and classified each insertion based on their presence (a "reference" insertion) or absence (a non-

184 reference or "*de novo*" insertion, Figure 2) in our reference genome [53, 73]. *G2/Jockey-3*

185 showed approximately five *de novo* insertions per line (File S4-S5), which is considerably fewer

186 than the most highly active *Drosophila* TEs. For example, we estimate *Hobo* and *Jockey-1* have

187 approximately 35 *de novo* insertions per line (consistent with [58]). However, the fraction of *de*

188 *novo* insertions as percentage of all insertions (39% or 733 / 1124 for *G2/Jockey3*) is comparable

189 to other highly active elements [72] such as *Doc* (25%; 2944 / 3826), *Jockey* (24% ;1308 /

190 5448), *I* element (27%; 658 / 2364), and *Copia* (88% ;1083 / 1226) (File S4).

191

192 We detected insertions in the centromeres, here defined as the centromere islands identified in

193 Chang et al. [53]. Of the 18 different TE families detected within centromeres, *G2/Jockey-3* is

194 the most common. Within centromeres, *G2/Jockey-3* has the highest percentage of *de novo*

195 (32%; 15/47) and reference (72%; 28 / 39, Figure 2) TE insertions. Only *G2/Jockey-3* and *Doc*

196 have more than 10 unique insertions (File S4). Other *de novo* centromere insertions include

197 active families such as the *F* element, *R1* and *Roo*.

198 Many active TEs are targeted by small, 24-30-nt long Piwi-interacting RNAs (piRNA) in the
199 germline [74, 75]. These piRNAs target TEs for both transcriptional and post-transcriptional
200 silencing. In the *D. melanogaster* germline, piRNAs are produced from precursor transcripts that
201 originate from discrete loci in the genome that contain clusters of nested TEs [76] called piRNA
202 clusters [77]. Subsequent amplification of the piRNA signal comes from the piRNA-guided
203 cleavage of active TEs through the 'ping-pong' cycle [77, 78] generating a signature of 10-nt
204 overlap between sense and antisense piRNAs. Consistent with *G2/Jockey-3* being a recently
205 active element, we detect insertions of *G2/Jockey-3* in the *38C1* (2 copies) and *38C2* (9 copies)
206 piRNA clusters, and *G2/Jockey-3*-derived piRNAs in the germline. *G2/Jockey-3* is ranked in the
207 top 4% and top 18% of TE-derived piRNAs in testes and ovaries, respectively (Table S4). The
208 resulting *G2/Jockey-3* piRNAs show signatures of ping-pong in ovaries (Z-score = 7.11; data
209 from [60, 79, 80]) but not in testes (Z-score = 0.15; data from [81-83]), consistent with other
210 testis piRNAs. Taken together, these results suggest that *G2/Jockey-3* is an active TE enriched in
211 centromere islands.
212
213
214
215
216
217
218
219
220

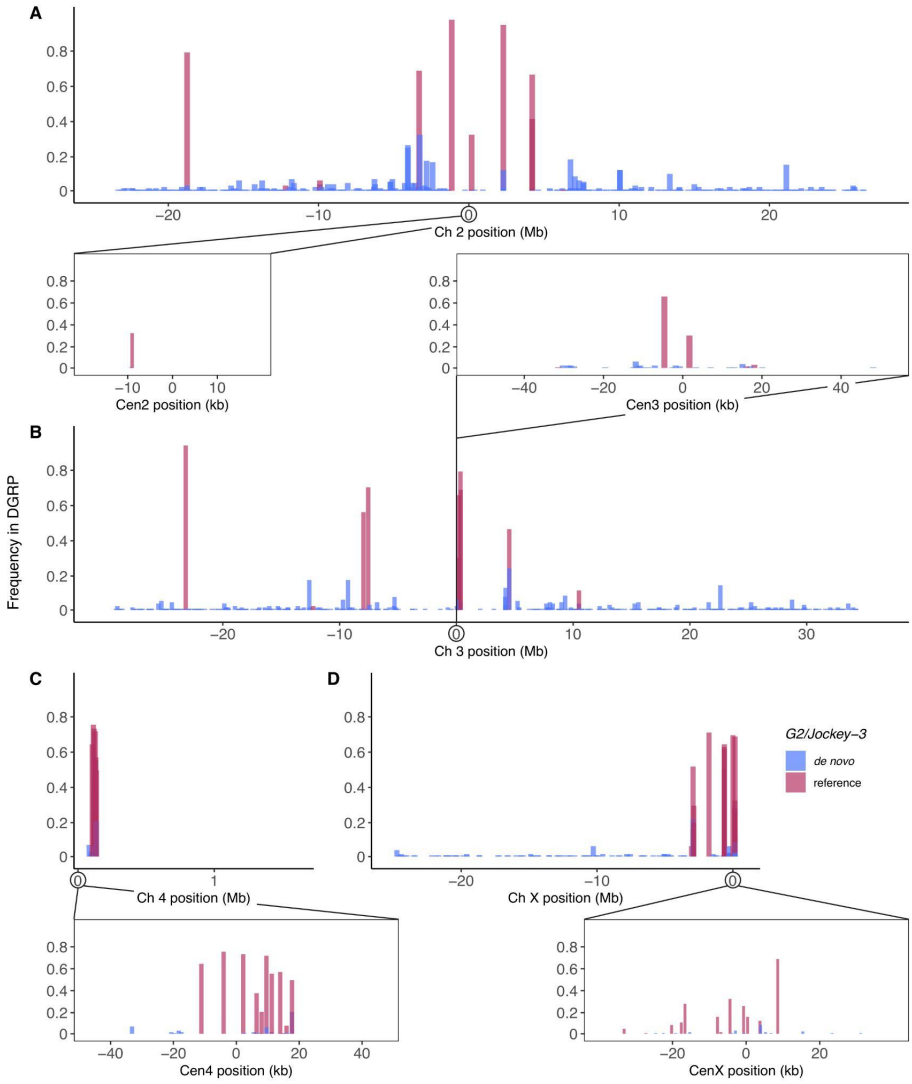


Figure 2: The distribution and frequency of *G2/Jockey-3* insertions in *D. melanogaster*. The height of the bars corresponds to frequency of an insertion detected in the DGRP samples and the color denotes whether an insertion is found in the reference genome or is *de novo*. Each A-D subpanels depict a chromosome (Ch 2, Ch 3, Ch 4, Ch X) from the heterochromatin-enriched assembly of *D. melanogaster* genome [53, 73], with a window zoomed in on the corresponding centromere (*e.g.,* X chromosome centromere = CenX, *etc*).

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

231

232 To understand if the observed enrichment of *G2/Jockey3* at centromeres is due to insertion bias,
233 we studied insertion frequencies across the genome. We reasoned that an enrichment of
234 putatively young *G2/Jockey-3* at centromeres may reflect an insertion preference. Young,
235 recently inserted TEs are expected to occur at lower frequencies than older TEs. We thus
236 examined the distribution of rare insertions (<10% population frequency) of *G2/Jockey-3* and
237 other TE families in the DGRP. We found that rare *G2/Jockey-3* insertions are significantly
238 enriched in centromeres compared to other regions relative to their size (19 centromere
239 insertions; Permutation tests: n_{perm} = 10,000, $P < 0.0001$; *Z*-score = 26.736 for whole genome; P
240 < 0.0001; *Z*-score = 23.895, excluding heterochromatin windows; $P < 0.0001$; *Z*-score = 17.354,
241 excluding euchromatin windows; Figure 3). *R1* elements are the only other TE enriched in
242 centromeres (4 centromere insertions; Permutation tests: n_{perm} = 10,000, $P < 0.0001$; *Z*-score =
243 6.892 for whole genome; $P < 0.0001$; *Z*-score = 5.92, excluding heterochromatin windows; P =
244 0.0187; *Z*-score = 2.605, excluding euchromatin windows; Figure 3), but to a lesser degree than
245 *G2/Jockey-3*. Among rare insertions, *G2/Jockey-3* is by far the most enriched TE in centromeres
246 (39% of all rare insertions).

247

248

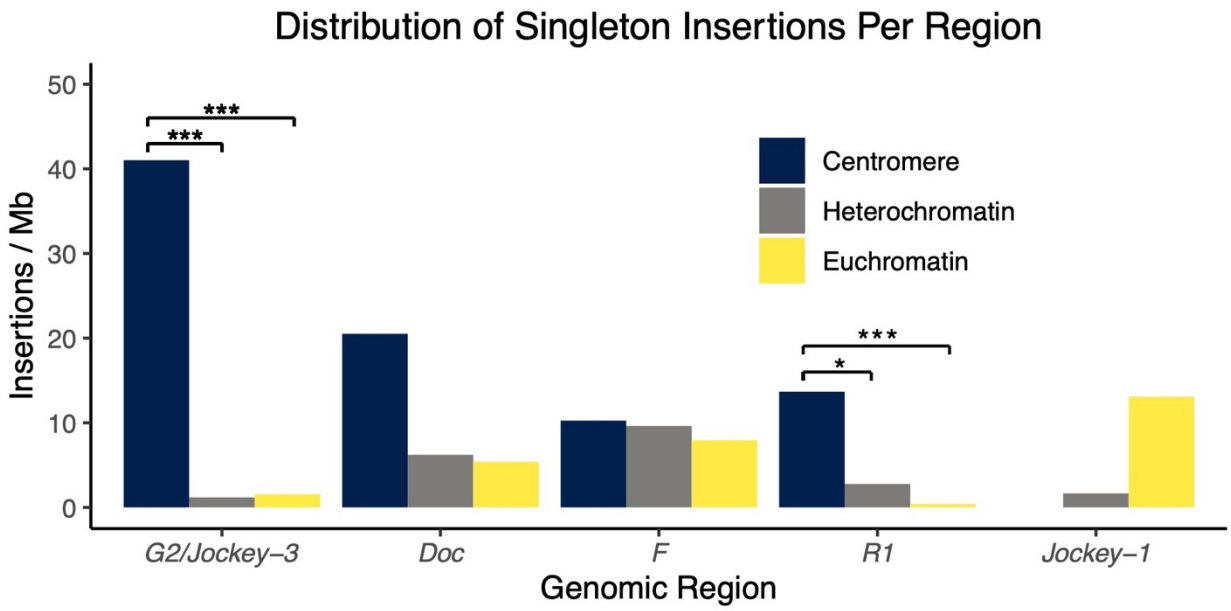
249

250

251

252

253



254

255

256 **Figure 3: Young *G2/Jockey-3* insertions are highly enriched in centromeres relative to**

257 **heterochromatin and euchromatin.** The distribution of rare insertions (< 10% of the

258 population) per Mb between the centromere islands, heterochromatin, and euchromatin regions

259 for several TE families in DGRP lines. Significant differences are noted as: * $P < 0.05$, *** $P <$

260 0.0001 in the permutation test.

261

262 We confirmed these patterns using an orthogonal approach with long-read genome assemblies of

263 flies from a wide geographic distribution (*i.e.,* founder DSPR lines, GDL lines, and our reference

264 genome; Table S1). Here we identified putatively young *G2/Jockey-3* insertions based on

265 sequence divergence instead of frequency. We classified young insertions as those with < 1%

266 divergence from the consensus sequence, or with < 2% divergence from the consensus and found

267 within a single genome assembly. Insertions classified as young were found in less than 50% of

268 the assemblies, had fewer unique substitutions, and were more likely to have a complete ORF

269 than older fragments (Figure 4, FET P = 3.93E-12, File S6, File S7). Consistent with our short-
270 read DGRP findings (Figure 3), we found significant relative enrichment in centromeres
271 (Permutation tests n_{perm} = 10,000: $P < 0.0001$; Z-score = 15.633 for whole genome; $P < 0.0001$;
272 Z-score = 7.505 excluding euchromatin windows), even with some of the centromere island
273 sequences missing in the long-read assemblies (Table S2).
274
275 These results suggest that *G2/Jockey-3* has had at least some insertion preference for
276 centromeres. As insertion bias can be associated with a target site sequence, we searched for
277 motif enrichment at insertions sites. Although we found an enrichment of AT-rich simple motifs
278 like "TAGTTTT" and rDNA-associated sequences (*e.g.* 18S and non-transcribed spacers; Figure
279 S2) within 100 bases of predicted insertion sites, these sequences are generally abundant in
280 repeat-dense regions of the genome and not specifically enriched at *G2/Jockey-3* insertions (File
281 S6, File S8; MEME E-value = 3.8E-16).
282
283
284
285
286
287
288
289
290
291

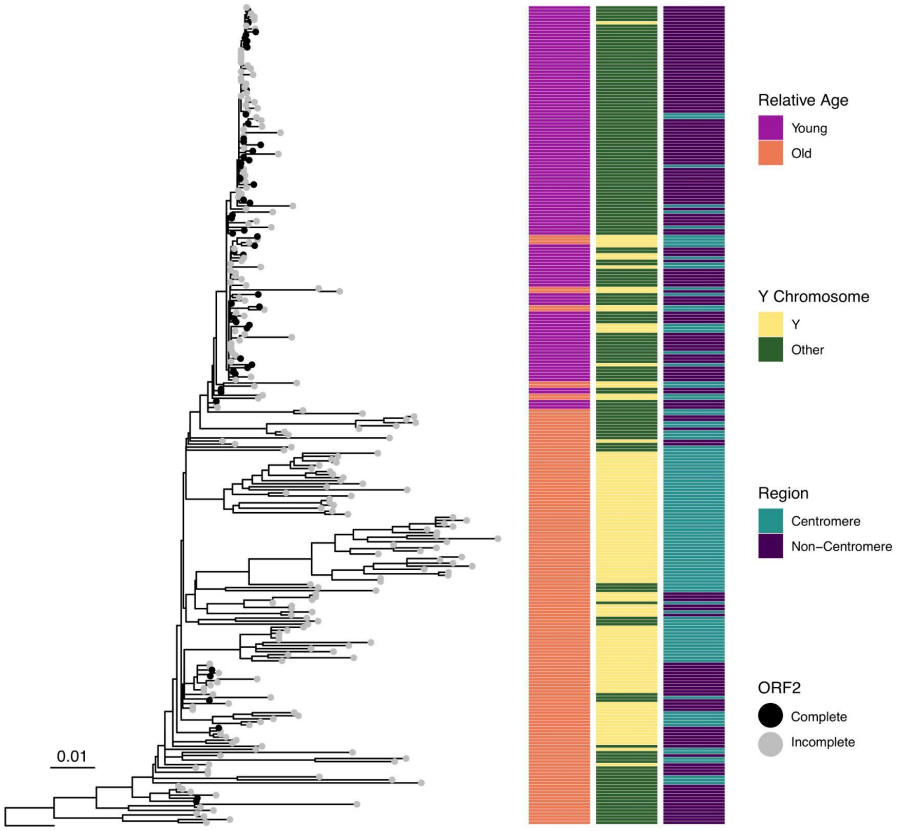


Figure 4: Young *G2/Jockey-3* insertions share a common origin. A maximum likelihood phylogenetic tree with all non-redundant *G2/Jockey-3* sequences larger than 10% of the full-length (400 bp) in the reference genome [53] and long-read assemblies of founder DSPR and GDL lines, with the *D. simulans* 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.

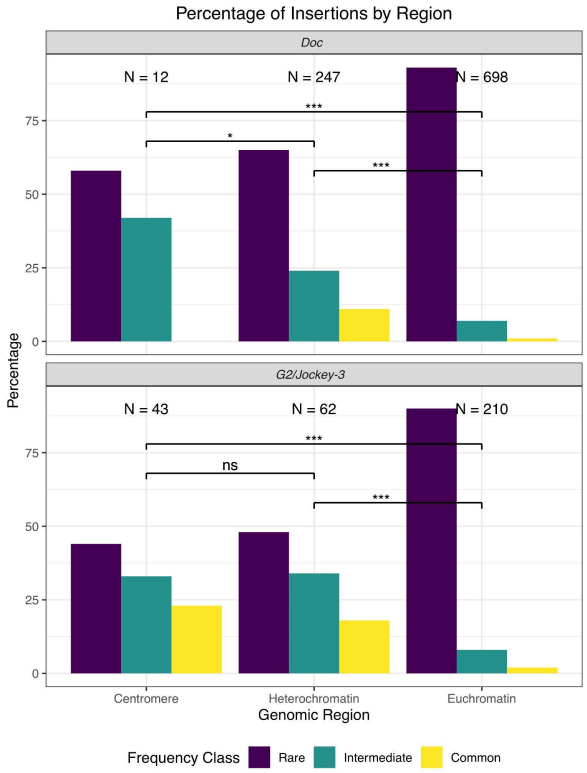
305 *5. Skew toward high frequency copies of G2/Jockey-3 in the centromere*

306

307 To test if *G2/Jockey3* insertions are retained at centromeres by natural selection, we studied their
308 frequencies. Because most TE insertions are deleterious in the euchromatin, they are found at
309 low population frequencies [84]. In contrast, TEs in heterochromatin are allowed to reach higher
310 population frequencies because recombination is infrequent, and natural selection is less
311 efficacious (reviewed in [85-87]). To infer differences in the strength and direction of natural
312 selection in different genomic regions, we examined the frequency spectra of a subset of TE
313 families in centromeres, heterochromatin, and euchromatin (Figure 5, Figure S3). If natural
314 selection favors centromeric insertions, we should observe a skew toward high frequency
315 variants in centromeres compared to heterochromatin. We compared the frequency spectra of
316 two TE families which had five or more insertions in the centromeres: *G2/Jockey-3* and *Doc*. As
317 expected, both elements showed an enrichment for rare insertions in the euchromatin, consistent
318 with TEs being generally deleterious in gene-rich regions (Figure 5; FET $P < 0.001$). Similarly,
319 we observed a slight skew toward higher frequency *G2/Jockey-3* and *Doc* insertions (*e.g.*
320 common/intermediate) in heterochromatin and centromeres compared to euchromatin (Figure 5,
321 FET $P < 0.001$), suggesting relaxed constraints or a reduced efficacy of natural selection outside
322 of euchromatic regions. Other TE families found in centromeres are also enriched for low
323 frequency insertions in the euchromatin (Figure S3). However, we did not observe differences
324 between the frequency spectra of the euchromatin and centromeres, suggesting most TEs may be
325 deleterious in these regions (Figure S3).

326

327 The frequency spectra of *G2/Jockey3* does not appear significantly different between
328 centromeres and heterochromatin. However, *G2/Jockey-3* has more "common" insertions in
329 centromeres relative to other TEs, including *Doc* (Figure 5; Figure S3). This could be due to
330 relaxed constraints allowing *G2/Jockey-3* insertions to rise to higher frequencies over time or
331 positive selection for some *G2/Jockey-3* insertions in centromeres.
332



333

334 **Figure 5: *G2/Jockey-3* has higher frequency variants in centromeres.** TE frequencies were
335 binned into rare (less than 10% of all samples), intermediate (10-50% of all samples), and
336 common (> 50% of all samples). The frequency distributions of the four TE families with > 10
337 unique insertions in the centromeres were compared among the centromeres, heterochromatin,
338 and euchromatin. Fisher's Exact Tests between regions, Bonferroni adjusted two-tailed P-value *
339 $P < 0.05$; *** $P < 0.001$; ns is not significant.

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

341

342 Centromeric *G2/Jockey-3* copies are generally older—within a range of 2-7% divergence from
343 the consensus—than those in heterochromatin (MWU, $P < 0.001$, Figure 6A), but not
344 significantly different than euchromatic copies (MWU, $P > 0.5$; Figure 6A). This pattern
345 suggests that most centromeric copies of *G2/Jockey-3* resulted from a historical burst of activity.
346 This burst of activity is consistent with general patterns observed at other non-LTR retroelements
347 [72, 88]. Other TE families show a similar age distribution between heterochromatin and the
348 centromeres (Figure S4). These copies are also older and highly fragmented. In general, both the
349 lack of recombination and lack of strong purifying selection against TE insertions in centromeres
350 and heterochromatin may allow these insertions to persist over longer evolutionary time periods
351 [89].

352

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

363 For this test, we need to compare insertions spanning a range of ages to identify individual
364 insertion frequencies that deviate from expectations [70]. We thus examine *G2/Jockey-3* and
365 other TE families that have centromere insertions. We find that *G2/Jockey-3* (Figure 6B, File
366 S9), like the majority of TEs, segregates either at the expected frequencies or lower than
367 expected, given their age (Figure S6, File S9). This suggests that many *G2/Jockey-3* copies, both
368 within and outside the centromeres, are either neutral or weakly deleterious. Therefore
369 *G2/Jockey-3* shares similar selective pressures as other presumably weakly deleterious TE
370 families in the genome.
371
372 Interestingly, we observed an enrichment of copies with zero unique substitutions for the
373 heterochromatin and CenY insertions (Figure S5). However, we suspect this pattern arises from
374 duplications rather than independent *G2/Jockey-3* insertion events (see below). Genomic
375 duplication events are common in heterochromatin in the Y chromosome [73, 76, 90].
376
377 While we do not see evidence for strong recent positive selection on any individual *G2/Jockey-3*
378 insertions in the centromeres, there may be selection on centromere haplotypes that differ in their
379 composition or structure.
380
381
382
383
384
385

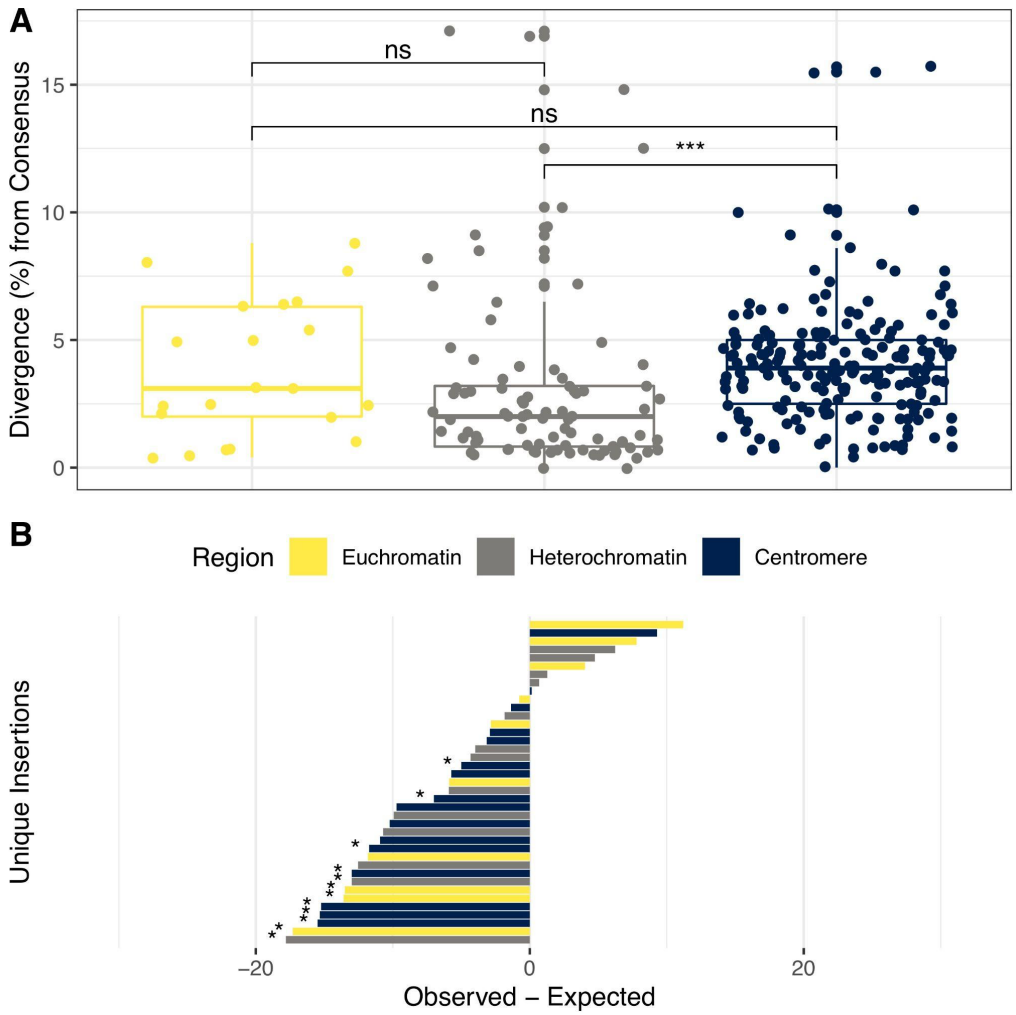


Figure 6: *G2/Jockey-3* copies are either as frequent or less frequent than expected given their age. A) The number of *G2/Jockey-3* insertions grouped by divergence from the consensus sequence split between the euchromatin, heterochromatin, and centromeres. Two-tailed Mann Whitney U tests *** $P < 0.001$, 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].

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

397

398 To detect signatures of natural selection on centromere haplotypes and any associations with
399 *G2/Jockey-3* copies, we examined patterns of nucleotide variation. Centromeres and
400 pericentromeric regions in *D. melanogaster* generally do not recombine by crossing over [91, 92]
401 therefore the effects of linked selection in non-recombining regions should lead to low levels of
402 nucleotide diversity. We thus expect to see reduced genetic variation in the centromeres and high
403 levels of linkage disequilibrium. We quantified genetic variation in the centromere islands for
404 DGRP samples with high centromere coverage (> 20X, see Methods). As expected, nucleotide
405 diversity is greatly reduced in every centromere compared to the major chromosome arms
406 (MWU $P < 2.0E-07$; Figure S7, Table S5) except for chromosome 4, which does not recombine,
407 and its centromere and chromosome arms are similar (MWU $p = 0.56$; Figure S8, Table S5). We
408 find that Tajima's *D*, a summary of the site frequency spectrum, is also typically lower in
409 centromere islands (Figure S8), indicating a skew toward rare alleles in centromeres. The
410 centromere island haplotype structures (Figures S9-S11) indicate that the majority of centromeric
411 SNPs are randomly distributed with respect to the *G2/Jockey-3* sequences (FET $P > 0.46$).
412 Linked selection anywhere in or near centromeres can reduce genetic diversity, skew frequency
413 spectra, and create haplotype structure. While centromere island haplotype structure is consistent
414 with a history of linked selection, we do not detect evidence that *G2/Jockey-3* insertions
415 themselves are necessarily the targets of selection.

416

417

418

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

420

421 The Y chromosome centromere is unique in that it has the largest centromere island containing
422 nearly half of all *G2/Jockey-3* fragments, but no identifiable satellite sequences [53]. We cannot
423 use the genomic re-sequencing data to study polymorphic insertions on the Y chromosome, as
424 the data are either female-only or mixed sex and thus lower coverage. However, our high-quality
425 reference assembly provides an opportunity to infer the organization and evolutionary history of
426 Y-linked *G2/Jockey-3*. We inferred a phylogenetic tree of all Y-linked copies of *G2/Jockey-3.*
427 Based on this phylogeny, we classified centromeric Y-linked *G2/Jockey-3* copies into four broad
428 classes (Figure S12). Class 1 *G2/Jockey-3* correspond to young and mostly long insertions, class
429 2 to short (< 160 bp) 3' end fragments, class 3 to old internal fragments missing the 5' and 3'
430 terminal ends, and class 4 to generally long and old 3' end fragments divided into two subclasses
431 (4 and 4A) based on position (Figure 9, Figure S12). We mapped these classifications onto our Y
432 chromosome assembly and estimated pairwise divergence between all *G2/Jockey-3* copies in the
433 Y centromere (Figure 7) and elsewhere on the Y chromosome (Figure S13). Each class seems to
434 undergo independent duplication events, however at a larger scale we do not identify higher
435 order repeat domains (Figure 7). This structure differs from other known centromeres (*e.g.*
436 primates), which are organized into higher order repeats, the youngest repeats are found in the
437 center and the older, more divergent repeats, are at the outer edges of the array [51, 52]. By
438 contrast, in the *D. melanogaster* Y centromere we observe that regions enriched with young
439 *G2/Jockey-3* copies punctuate regions rich with older copies (Figure 9), suggesting that recent,
440 local rearrangements affect centromere organization.

441

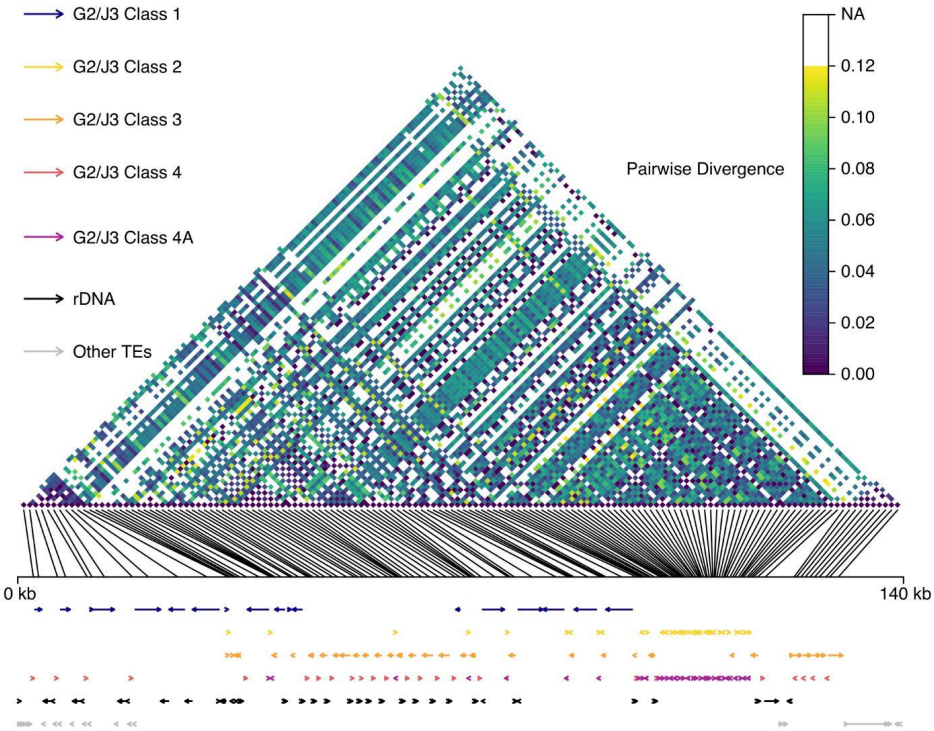


Figure 7: *G2/Jockey-3* evolution in the Y chromosome centromere is shaped by expansion via duplication, gene conversion and rearrangements. The colored pairwise divergences between individual *G2/Jockey-3* copies on the Y centromere are shown on top with their relative positions on the line graph below. 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.

455 We estimated gene conversion rates of *G2/Jockey-3* on the Y centromere with respect to their
456 classes (Supplemental Text). Our estimates for gene conversion are on the order of 1E-6 per site
457 per generation (Table S6), ~10E3 times higher than the other chromosomes in *Drosophila* [91-
458 93] and an order of magnitude less than previously estimated gene conversion rates between
459 duplicated gene families on the Y chromosome [73]. Moreover, we do not see complete
460 sequence homogenization within classes. We observe similar patterns of duplication and
461 expansion outside the centromere on the Y chromosome (Figure S13), as 32 of 47 non-
462 centromere Y-linked copies appear to be duplicated; they are of similar fragment size,
463 orientation, and have few no zero unique substitutions (File S6). Therefore, we conclude that
464 *G2/Jockey-3* evolution on the Y chromosome is driven by duplication of historical insertions
465 with some gene conversion among copies. Population genomic data from the Y chromosome will
466 help distinguish between types of recombination affecting Y centromere structure.
467
468 Outside of the Y chromosome, putative duplications involving *G2/Jockey-3* are less common
469 (Figure S13). Many copies of *G2/Jockey-3* within the centromeres are fragmented and therefore
470 share no homologous sequences for pairwise calculations (Figure S13). Pairwise divergence
471 between centromeric *G2/Jockey-3* fragments is elevated in comparison to fragments in both the
472 euchromatin and non-centromeric heterochromatin (Figure S14, MWU, $P < 0.01$), and lowest in
473 non-centromeric heterochromatin, especially on the Y chromosome. Therefore, the majority of
474 *G2/Jockey-3* within the non-Y centromeres appears to be the result of recurrent past insertions.
475
476
477

478 CONCLUSIONS

479

480 *G2/Jockey-3* is a unique TE in that it is the only sequence shared among centromeres of *D.*
481 *melanogaster* [53]. Here we show that *G2/Jockey-3* is a recently active non-LTR retroelement
482 that may have had an insertion preference for centromeres, given their relatively small size
483 compared to the rest of the genome (Figures 3). While the centromere enrichment and skew
484 toward high frequency centromere copies are expected for insertions favored by natural
485 selection, we do not find evidence for positive selection on individual *G2/Jockey-3* insertions.
486 Instead, most *G2/Jockey-3* insertions both inside and outside of the centromeres are at the
487 expected frequency or significantly rarer than expected given their age, suggesting they may be
488 neutral or slightly deleterious and accumulating due to relaxed constraints against centromeric
489 TE insertions (Figure 6B). The majority of *G2/Jockey-3* copies within the centromeres appear to
490 be the result of recurrent past insertions, except for on the Y chromosome. Some Y-linked copies
491 are likely a result of duplications rather than *de novo* insertions, similar to patterns reported in
492 other species (*e.g*. [28]).

493

494 We propose that most *G2/Jockey-3* insertions in the centromeres resulted from a burst of activity,
495 which is consistent with the evolutionary history of other non-LTR retroelements [88]. When a
496 TE first invades, it can proliferate rapidly until it comes under the surveillance of germline
497 defense mechanisms like the piRNA pathway, for example by inserting into a piRNA cluster
498 [94-96]. These dynamics are consistent with the invasion and control [94, 97] of the *P* element in
499 *D. melanogaster* [98] and its close relative, *D. simulans* [99]. Once targeted by the piRNA
500 pathway for transcriptional and posttranscriptional silencing, the TE may eventually become

501 dormant or go extinct. Here we show that *G2/Jockey-3* had a burst of activity and is now inside
502 piRNA clusters and actively targeted by the piRNA pathway (Table S4).
503
504 The *G2/Jockey-3* insertion polymorphism patterns and piRNA observations are most consistent
505 with some underlying selfish behavior typical among TE families, except that this element can
506 target centromeres. Although we are unsure of the precise targeting mechanism, we suspect the
507 target is the centromere chromatin rather than any particular DNA sequence. Such centromere
508 targeting was reported for LTR retroelements in other species such as maize [100, 101] and
509 *Arabidopsis* [102, 103]. For example, when the *Arabidopsis lyrata Tal-1 Copia*-type
510 retroelement [103] is introduced into *A. thaliana,* the element targets the centromeres despite
511 divergence in centromeric DNA sequences between these species [102].
512
513 TEs may accumulate in heterochromatic regions where they experience relaxed selection
514 compared to euchromatic insertions, which have potential to disrupt functional elements, risk
515 ectopic exchange and resulting genomic rearrangements [85, 86, 104], and spread silent
516 chromatin or DNA methylation into nearby functional elements [*e.g.*, 105-108]. However,
517 centromeres are not typical heterochromatin [109]. Here we propose that centromeres offer a
518 unique environment for TEs (Figure 8A). First, insertions are unlikely to be strongly disfavored,
519 as the function of centromeres is unlikely to be disrupted by a single insertion event. Second,
520 centromeric insertions may escape host silencing mechanisms. Small RNAs target TEs to limit
521 potentially harmful TE activity by inducing H3K9 methylation and thus transcriptional silencing
522 (reviewed in [110-112]). However, CENP-A structure and posttranslational modifications differ
523 from H3 [109], and the host silencing machinery is not known to target CENP-A. Therefore,

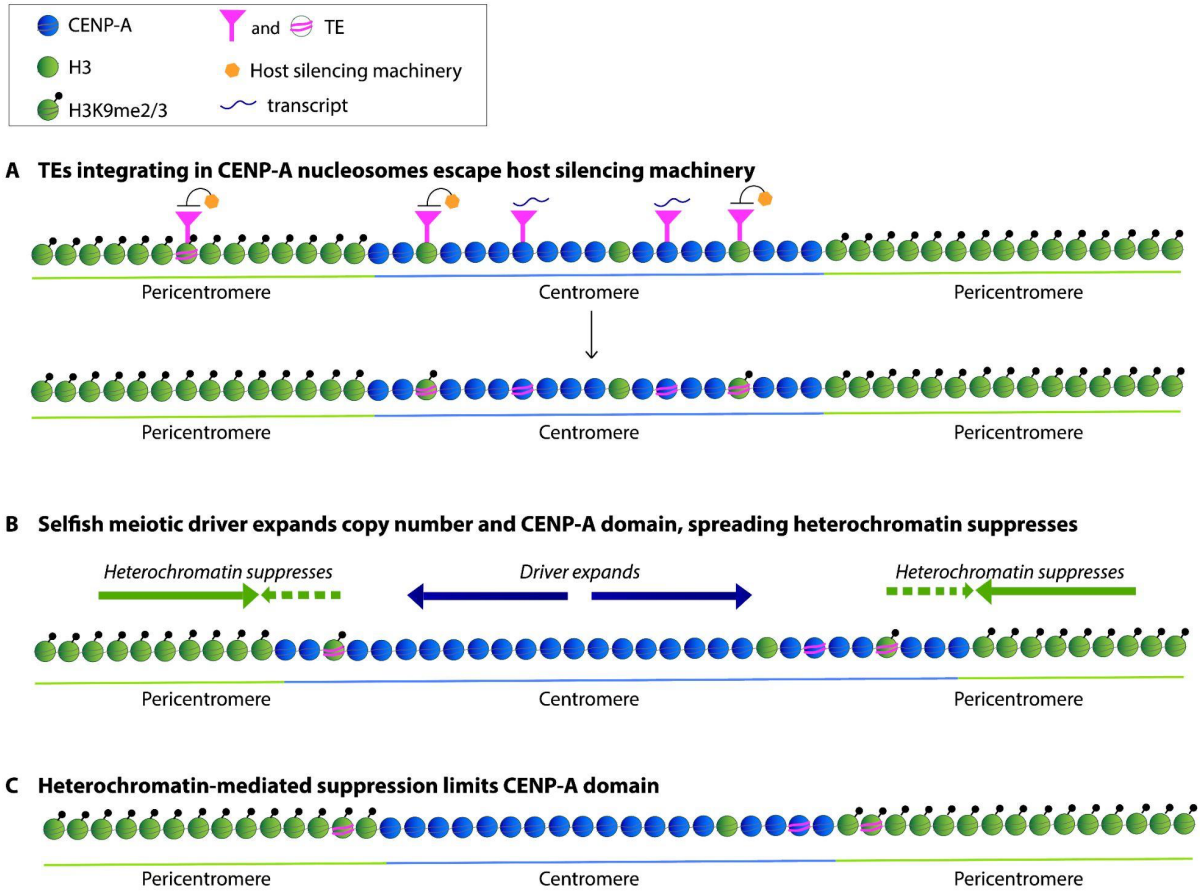
524 centromeres may offer a 'safe haven' for some TEs not just because they can avoid removal by
525 purifying selection, but also because TEs may temporarily escape targeting by small RNAs (*e.g.*
526 piRNAs, siRNAs) and alternative host defenses (*e.g.* DNA methylation in pericentromeres [113],
527 hypomethylated DNA in humans [52, 114]).
528
529 While most TEs are considered to be parasitic elements that are neutral or mildly deleterious
530 [115, 116], some TE insertions, including those in non-recombining and/or gene poor
531 heterochromatic environments, might contribute to host function [117]. For example, TEs can
532 adopt commensal and/or mutualistic strategies to propagate without direct confrontation by
533 cooperating with their host (reviewed in [118]). One extreme outcome is TE domestication,
534 where the elements no longer pose a threat to the host and instead now perform some important
535 genome function. One often cited example is at *Drosophila* telomeres, where three non-LTR
536 retroelements in the *Jockey* clade (*HeT-A*, *TAHRE*, and *TART* collectively referred to as HTT)
537 have specialized roles [119, 120]. The transposition activity of HTT is limited to telomeres,
538 where they directly contribute to telomere length and integrity [119-122]. However, detailed
539 study of the evolutionary genetic history of the retroelements across species suggests instead that
540 there is still an underlying conflict between the HTT and the genome, as there is rapid turnover
541 of HTT between species [123]. Alternatively, given the specialization of HTT, the dynamic and
542 unstable telomeric environment may necessitate rapid HTT evolution [124].
543
544 While centromeric TEs do not currently show evidence for domestication in *D. melanogaster*,
545 they may be tolerated, or even beneficial for three reasons. First, TEs may contribute to
546 centromere transcription (reviewed in [35]), as some transcription at centromeres is important for

547 CENP-A assembly and function [14, 46, 125-129] and centromere function [14, 45, 130].
548 Second, TEs might contribute to the unique chromatin at centromeres. H3 nucleosomes are
549 typically interspersed with CENP-A within centromere domains [12, 14] [15], including a
550 limited amount of H3K9me2/3 (*e.g.*, [13]). We propose that insertion of active TEs like
551 *G2/Jockey-3* may contribute to this unique CENP-A/H3 chromatin at the centromeres (Figure
552 8A). While landing in CENP-A nucleosomes may allow the TE to temporarily escape silencing,
553 if an active TE lands in an H3 nucleosome region of the centromere, it may be targeted for
554 silencing through H3K9 methylation (*e.g.* piRNA- or siRNA-mediated silencing) and thus
555 contribute to H3K9me2/3 domains within the centromere. Third, TEs may help mitigate other
556 conflicts that happen at centromeres. Some centromeric sequences have the capacity to be selfish
557 and bias their transmission through female meiosis in a process known as meiotic drive [131].
558 Driving centromeres generally have expansions of satellite DNA sequences that can recruit more
559 kinetochore proteins and cause biased segregation toward the egg pole [132, 133] (Figure 8B).
560 Meiotic drive at the centromeres can lead to rapid evolution of both centromere DNA and
561 proteins that work to restore fair segregation [134, 135]. There are parallel pathways to suppress
562 meiotic drive involving rebalancing kinetochore proteins and spreading heterochromatin to limit
563 the CENP-A domain [136]. While centromere drive and suppression models predict that the
564 flanking heterochromatin domains are the main participants in the conflict [136], the
565 H3K9me2/3 domains within the centromeres might also contribute. Active centromeric TEs that
566 are targeted by small RNA pathways, like *G2/Jockey-3,* could lead to spreading of H3K9me2/3
567 heterochromatin from within the centromere (Figure 8C). In this way, the centromeric TEs and
568 'internal' centromeric H3K9 domains may also be a substrate for suppressors of centromere
569 drive to stop the unchecked spread of CENP-A caused by meiotic drivers.

570 Even if TEs cooperate with centromeres, this may be an uneasy alliance, as there could still be an
571 underlying conflict. For example, the invasion of H3K9me2/3 heterochromatin in centromeres is
572 potentially dangerous, as it may evict too much CENP-A and silence a centromere [5].
573 Therefore, the centromeric chromatin, and the relative abundance of CENP-A and H3
574 nucleosomes, may mediate the balance between conflict and cooperation between active TEs and
575 centromeres.
576
577 In many species across the tree of life, centromeres are associated with TEs [18, 23-34, 53]. The
578 connection between TEs and centromeres is supported not just by TE sequences embedded in
579 and around present-day centromeres, but also in the origins of some proteins that function in
580 centromere biology (*e.g.* similarity between centromere protein CENP-B and Pogo family
581 transposons [137]). This relationship may represent a delicate balance between conflict and
582 cooperation between the centromeres and the repetitive selfish genetic elements that occupy
583 these genome regions. We expect this balance to vary depending on genomic context, level of
584 TE activity, and presence and frequency of centromere meiotic drive. If TEs contribute
585 positively to centromere function, many types of elements may be sufficient to provide the right
586 'environment'. In this case, we would not necessarily expect to see a signature of selection on
587 any one insertion or type of insertion. While *G2/Jockey-3* is currently the principal element
588 occupying the centromere niche in *D. melanogaster*, it may be replaced over time by another
589 type of repeat. In our model, any repeat that can be transcribed and can recruit H3K9me2/3 (*e.g.*
590 is targeted by small RNAs) can contribute to centromere function. Detecting the mode, tempo,
591 and target of selection in non-recombining and dynamic environments like the centromeres is

592 difficult. Exploring the evolution of centromere composition in closely related species will

593 provide additional insights necessary to test these hypotheses.



594

595 **Figure 8. Model for the accumulation of centromere-associated retroelements.** Centromeres

596 have unique chromatin with interspersed CENP-A and H3 domains, including some H3K9me2/3

597 [12-15]. The diagrams represent nucleosomes, either CENP-A (blue) or H3 (green), with

598 H3K9me2/3 modifications indicated with a black dot. The nucleosome distributions are not

599 drawn to scale. A. Active TEs like *G2/Jockey-3* jump in centromeric chromatin (magenta

600 triangles and lines in nucleosomes) and could contribute to transcription (black wavy line) or be

601 silenced by host machinery (orange hexagon). Centromeres may offer a safe haven for TE

602 insertion because insertions there are unlikely to be strongly disfavored and because insertions in

603 CENP-A nucleosomes may escape host silencing machinery (*e.g.* piRNA pathway). Centromeric
604 TE insertions might be tolerated or even favored by the host if they contribute to centromere
605 transcription or a low level of H3K9me2/3 inside centromeres. B. Selfish centromeric DNA
606 sequences can expand their copy number and bias their transmission to the next generation
607 through female meiosis [131]. C. One way to suppress cheating centromeres is by spreading
608 heterochromatin to limit the CENP-A domain [136]. TE insertions in H3 nucleosomes within the
609 centromeres may be targeted for silencing by host silencing machinery (green circle with
610 magenta lines and black circle). These H3K9me2/3 nucleosomes within centromeres may help
611 create a more efficient response to driving centromeres by merging spreading pericentric
612 heterochromatin with internal H3K9me2/3 domains.
613

614 **MATERIALS AND METHODS**

615

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

617

618 Chang et. al. [53] generated the *G2/Jockey-3* consensus sequence after a phylogenetic analysis
619 confirmed previously annotated *G2* and *Jockey-3* were the same element [53]. We annotated a
620 heterochromatin-enriched long-read genome assembly of *D. melanogaster* [73] with
621 Repeatmasker v4.1.0 using the previous *G2/Jockey-3* consensus sequence along with a
622 *Drosophila* repeat library [138]. We extracted *G2/Jockey-3* and the surrounding sequences using
623 BEDTools v2.29.2 [139], aligned the sequences to the previous consensus sequence, manually
624 inspected the alignment, and generated the new consensus sequence with Geneious v8.1.6 [140]
625 (File S1). We identified Open Reading Frames (ORFs) with NCBI ORFfinder

626 (https://www.ncbi.nlm.nih.gov/orffinder/) and screened the ORF candidates by identifying
627 conserved domains common in non-LTR retrotransposons with the NCBI Conserved Domains
628 search tool (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). We defined the 5' end by
629 identifying transcription along the length of *G2/Jockey-3* by mapping publicly-available single-
630 end Precision nuclear Run-On Sequencing (PROseq) downloaded from NCBI SRA under
631 BioProject PRJNA451204 and accessions SRR7050623 - SRR7050630 [62]. We removed
632 adapters and reads < 16bp using cutadapt v2.1. Reads were reverse complemented using the
633 fastX toolkit v0.0.13 (fastx_reverse_complement). We mapped reads to the heterochromatin-
634 enriched genome assembly of *D. melanogaster* using bowtie2 v2.3.5.1 (--no-discordant --
635 dovetail --very-sensitive options; [141]). We extracted reads > 16 bp in length mapping to
636 *G2/Jockey-3* annotations in the genome with samtools v1.7 [142] and remapped the reads to the
637 *G2/Jockey-3* consensus with bowtie2 v2.3.5.1 [141]. We separated reads based on their strand
638 using samtools (v1.7). We calculated read depth at each position along the *G2/Jockey-3*
639 consensus and normalized coverage by the total number of genome mapped reads (RPM) using
640 bedtools genomecov [139]. We used alignment depth of all *G2/Jockey-3* genomic fragments to
641 identify where transcription occurs along the length of *G2/Jockey-3*. We submitted the sequence
642 in 5' end motifs under transcription peaks for a GoMo analysis [143] to identify motifs as
643 possible promoters.
644
645 To detect piRNAs corresponding to *G2/Jockey-3,* we trimmed the small RNA-seq reads for
646 adaptors and excluded the reads that align to rRNA or tRNA. We aligned the remaining reads to
647 the heterochromatin-enriched genome assembly [53, 73] using Bowtie [144], and then used a
648 customized python script as described in Wei et al. [83] to count reads that mapped to each TE

649 feature (htseq_bam_count_proportional.py;
650 https://github.com/LarracuenteLab/Dmelanogaster_satDNA_regulation; last accessed 11/2022,
651 https://doi.org/10.5061/dryad.hdr7sqvj3). We reported the RPM (Reads Per Million) values in a
652 summary table. We calculated the 10nt overlap Z-score of piRNAs using piPipes [145].
653
654 *Short-read analysis*
655
656 We detected TE insertions from publicly available FASTQ files from 135 *Drosophila*
657 *melanogaster* Genetic Reference Panel (DGRP) lines from BioProject PRJNA36679 [64, 65]
658 downloaded from the European Nucleotide Archive last accessed September 20, 2019, 13 of the
659 founder lines of the Drosophila Synthetic Population Resource (DSPR) BioProject
660 PRJNA156883[146], and 85 samples from the Global Diversity Lines BioProject PRJNA268111
661 (Table S3)[147]. We detected TE insertions in each sample with the McClintock pipeline [69]
662 (last downloaded and updated, August 20, 2020) combining several split-read and discordant
663 read mapping programs including the following: ngs_te_mapper, popoolationTE,
664 popoolationTE2, Retroseq, TE-locate, and TEMP (scripts and files for running McClintock at
665 https://github.com/LarracuenteLab/Dmel_Jockey-3_Evolution). We concatenated the outputs to
666 calculate the number of TEs detected by each program and sample and analyzed the outputs with
667 Python and R scripts available at https://github.com/LarracuenteLab/Dmel_Jockey-3_Evolution.
668 We repeated our analyses of TEs with and without insertions in known piRNA clusters and
669 removed fragments of Gypsy6_I_Dmoj and Gypsy1_I_Dmoj which resemble satellite sequences
670 in *D. melanogaster*. Depth of each short-read sample was calculated from the sorted BAM
671 outputs from the McClintock pipeline using Mosdepth [148] (Table S3). We retained samples

672 which had a centromere coverage greater than 20X for our analyses (see Supplemental Text for
673 more information).
674
675 We used the same short-read FASTQ files from the DGRP to quantify genetic diversity between
676 the centromere islands and the rest of the genome using a best practices GATK pipeline v4.1.9.0
677 to detect variant SNPs. We mapped reads from high coverage samples to our reference genome
678 with BWA v1.7.15 and samtools 1.7 to generate sorted BAM files and used GATK
679 MarkDuplicates, BaseRecalibrator, ApplyBQSR, AnalyzeCovariates, and HaplotypeCaller to
680 call and filter SNPs for each sample into a VCF file. The VCFs for all samples were
681 concatenated with GenotypeGVCFs and relicabrated with VariantRecalibrator and ApplyVQSR.
682 We quality filtered SNPs in the resulting GVCF file with the following criteria; QD < 2.0, FS >
683 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. We calculated π and
684 Tajima's *D* with the SimplifyVCF_Basic.v2.pl and Windows_Basic.v2.pl Perl scripts on the
685 filtered SNPs (available at https://github.com/bnavarrodominguez/SD-Population-genomics,
686 [149]). We further analyzed variants within the centromeres with VCFtools [150] and the vcfR R
687 package [151]. Additional details of our analysis with short-read data and the results can be
688 found in the Supplemental Text section.
689
690 *Long-read analysis*
691
692 We assembled genomes for several *D. melanogaster* lines following the heterochromatin-
693 enriched assembly protocol from Chang & Larracuente [73]. We downloaded raw PacBio reads
694 for five of the Global Diversity Lines (GDL) from the NBCI BioProject PRJNA451471 [68],

695 eight Drosophila Synthetic Population Resources (DSPR) founder lines and Oregon-R with raw
696 Pacbio reads from BioProject PRJNA418342 [67], and five DGRP lines from Bioproject
697 PRJNA559813 [66]. We assembled draft genomes using all reads with Canu v1.7.1 [152] with
698 the relaxed corrected error rate and the parameters "genomeSize=160m useGrid=false
699 stopOnReadQuality=false corMinCoverage=0 corOutCoverage=100 correctedErrorRate=0.100".
700 We then mapped PacBio reads to the reference genome with minimap2 (downloaded August 28,
701 2018) with the "-x map-bp" parameter setting [153]. We then used unmapped reads and reads
702 mapping to heterochromatic regions to generate an additional "heterochromatin-enriched"
703 assembly. We merged the two assemblies into a single assembly with quickmerge [154] and
704 polished the assemblies with Racon v1.3.1 [155] by mapping the PacBio reads three times to the
705 assembly using minimap2 [153]. We also polished the GDL assemblies twice with Illumina
706 reads from BioProject PRJNA268111 [147], the DSPR assemblies with reads from
707 https://wfitch.bio.uci.edu/~dspr/index.html (last accessed September 6, 2021 [156]), BioProject
708 PRJNA252961 for Oregon-R, and the DGRP assemblies with reads from BioProject
709 PRJNA36679 [64, 65]. We annotated all assemblies with Repeatmasker v4.1.0 and our
710 *Drosophila* TE library with the updated *G2/Jockey-3* consensus sequence [138] (File S10). We
711 identified centromere island sequences in assemblies using a combination of methods to find
712 sequences corresponding to the centromere islands identified in our reference genome [53].
713 These include aligning assemblies to the centromere islands with MUMMER v.3.23 nucmer
714 [157], identification of similar sequences with BLAST 2.10.0+ [158], and identifying contigs
715 enriched in *G2/Jockey-3* or satellite DNA repeats enriched around the centromere islands [53].
716 We also used MEME [159] to detect motifs enriched in the flanking regions of young insertions

717 in our assemblies. Additional details of our analysis with long-read data, the results, and the
718 MEME motif analysis are in the Supplemental Text.
719
720 *Insertion polymorphism validation*
721
722 We validated TE insertion predictions in three ways: 1) we used short reads to confirm TE
723 predictions in long read assemblies (see Supplemental Text); 2) we used long read assemblies to
724 confirm TE predictions in short read data (see Supplemental Text); and 3) we selected a subset to
725 validate with PCR (File S3). We used the results from these analyses to establish criteria for
726 including a predicted insertion in our analyses. For PCR, we selected a subset of 59 *G2/Jockey-3*
727 copies to confirm in four GDL lines; B59, I23, T29A, and ZH26. We included copies from each
728 chromosome (X, 2, 3, 4), each chromatin region (euchromatin, heterochromatin, and
729 centromere), and both reference and non-reference copies. We were able to detect 33 copies
730 including 13 copies in the centromere and seven *de novo* insertions (File S3).
731 We used permutation tests to test for significant differences in density of recent *G2/Jockey-3*
732 insertions between genome regions. For these tests, the genome was divided into 90-kb windows
733 (average size of a centromere, excluding the Y chromosome and 2nd chromosome) and
734 categorized as centromere, heterochromatin, or euchromatin. We estimated *G2/Jockey-3* counts
735 in each window based on final insertion polymorphisms detected by short-read and long-read
736 datasets separately. We excluded the Y chromosome from these analyses. To test for enrichment
737 of *G2/Jockey-3* counts within categories, we permuted the datasets 10,000 times using regioneR
738 1.18.1 ([160] ntimes=10000, randomize.function=resampleRegions,

739 evaluate.function=numOverlaps). Results were similar for permutation tests with the whole
740 genome, excluding heterochromatin windows, and excluding euchromatin windows.
741
742 *G2/Jockey-3 sequence and phylogenetic analysis in D. melanogaster*
743
744 We extracted all *G2/Jockey-3* sequences from the annotated assemblies using BEDTools v2.29.2
745 [139] and classified the insertions as reference or non-reference (*de novo*). We classified
746 insertions as homologous to those in our reference if they were both fragments of similar size
747 (+/- 10 bp) and within +100 bp of the same location in the reference genome. We identified
748 corresponding locations by aligning the assemblies to the reference genome with nucmer within
749 MUMMER v.3.23 [157]. Insertions were classified as *de novo* if they differed in fragment size
750 length and position of reference insertions, found outside the reference fragment position range,
751 and located within breaks in the nucmer alignment to the reference genome.
752
753 We performed phylogenetic analyses for candidate *G2/Jockey-3* sequences identified in our
754 genome assemblies. We used *G2/Jockey-3* sequences from the reference and the non-reference
755 insertions discovered in the DSPR and GDL assemblies. If a non-reference insertion was found
756 in multiple assemblies, we randomly selected an assembly from which to extract that insertion to
757 include in our analysis. Prior to phylogenetic analysis, we filtered sequences to exclude
758 fragments shorter than 400 bp or ~10% of the length of a full *G2/Jockey-3* element. We aligned
759 the sequences with MAFFT [161] with the *D. simulans G2/Jockey-3* consensus as the outgroup
760 and inferred a maximum likelihood phylogenetic tree with RAxML v8.2.11 [162]. To study the
761 evolution of *G2/Jockey-3* on the Y chromosome, we inferred a maximum likelihood phylogeny

762 of all fragments exclusive to the Y chromosome regardless of size. We determined the relative
763 age of individual insertions using divergence from the consensus from the Repeatmasker output
764 and the number of substitutions unique to a fragment divided by fragment length. "Young"
765 insertions were either < 1% diverged from the consensus sequence or *de novo* insertions
766 exclusive to a single genome assembly and < 2% diverged from the consensus sequence. We
767 calculated pairwise divergence between fragments with MEGA X [163] using the Kimura 2-
768 parameter model assuming uniform rates among all sites. We estimated gene conversion rates
769 between elements using the same method outlined from Chang & Larracuente [73] based on
770 Ohta's α equation [164] and Rozen's equation [165]. We used the following equation for our
771 calculation of gene conversion:

772 $$c = \frac{2u(n-1)}{\pi}$$

773 We determined the rate at which a number of sequences (n) homogenize each other per
774 generation (c) based on the point mutation rate (u) and divergence among the sequences (π). We
775 grouped reference *G2/Jockey-3* sequences located within the same contig except for insertions on
776 the Y centromere. Fragments from the Y centromere were grouped based on the
777 phylogenetically-defined classes. We aligned the sequences with MAFFT [161] and calculated
778 divergence with MEGA X [163]. We normalized c based on gene conversion tract length of a
779 single event (c_g) based on previously estimated 400 bp tract length of a single event in *D.*
780 *melanogaster* [92, 93]:

781 $$c_g = \frac{c}{400\ x\ 2}$$

782

783

784

785 *Age-of-allele test*

786

787 Natural selection, transposition, and frequency of insertion influence TE evolution within
788 populations. We used an age-of-allele test [70] to determine departure from neutral evolution of
789 individual insertions based on population frequency, age, and demographic history in *D.*
790 *melanogaster*. The age-of-allele test determines the probability distribution of an insertion
791 frequency within a population under neutrality conditioned on the age of the insertion. The age is
792 inferred by the number of unique substitutions within a TE sequence accumulated since the time
793 of the insertion. To determine the number of unique substitutions, we aligned insertions with
794 MAFFT [161] for *G2/Jockey-3*, *BS*, *G5*, *Doc*, and *Doc2* detected as reference insertions in the
795 McClintock output. Unique substitutions were quantified from each individual insertion using
796 Biostrings [166]. We calculated the probability distribution of the frequency of each unique
797 insertion with the age-of-allele test script from Blumenstiel et al. [70] in Mathematica v12.0.0.0.
798 The probability distribution is also conditioned on population demography of North American *D.*
799 *melanogaster* populations outlined in the original paper [70]. North American *D. melanogaster*
800 populations experienced two bottlenecks in their migration from Africa [167, 168] that led to
801 non-equilibrium TE evolution [70]. We compared the probability distribution to the frequency of
802 insertions detected in the 28 DGRP samples with >25X average coverage of the centromere
803 islands. We excluded insertions in the Y chromosome given the high rate of duplication. The
804 expected frequency of each insertion in the population is the probability of observing an insertion
805 as frequent or more frequent within the population multiplied by the observed number of
806 insertions in our population sample. We calculated significant deviations from neutrality as
807 probability of observing as many or fewer, or as many or greater, copies < 0.1.

COMPETING INTERESTS STATEMENT

The authors have no competing interests.

DATA AVAILABILITY STATEMENT

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

ACKNOWLEDGEMENTS

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

REFERENCES

1. Allshire RC, Karpen GH: **Epigenetic regulation of centromeric chromatin: old dogs, new tricks?** *Nat Rev Genet* 2008, **9:**923-937.
2. Garrido-Ramos MA: **Satellite DNA: An Evolving Topic.** *Genes* 2017, **8**.
3. Hartley G, O'Neill RJ: **Centromere Repeats: Hidden Gems of the Genome.** *Genes* 2019, **10**.
4. Pidoux AL, Allshire RC: **Kinetochore and heterochromatin domains of the fission yeast centromere.** *Chromosome Res* 2004, **12:**521-534.
5. Ohzeki J-I, Bergmann JH, Kouprina N, Noskov VN, Nakano M, Kimura H, Earnshaw WC, Larionov V, Masumoto H: **Breaking the HAC Barrier: histone H3K9 acetyl/methyl balance regulates CENP-A assembly.** *EMBO J* 2012, **31:**2391-2402.

836 6. Choo KH: **Centromerization.** *Trends Cell Biol* 2000, **10:**182-188.
837 7. Henikoff S, Dalal Y: **Centromeric chromatin: what makes it unique?** *Curr Opin*
838 *Genet Dev* 2005, **15:**177-184.
839 8. Henikoff S, Furuyama T: **Epigenetic inheritance of centromeres.** *Cold Spring Harb*
840 *Symp Quant Biol* 2010, **75:**51-60.
841 9. Black BE, Cleveland DW: **Epigenetic centromere propagation and the nature of**
842 **CENP-a nucleosomes.** *Cell* 2011, **144:**471-479.
843 10. Mendiburo MJ, Padeken J, Fülöp S, Schepers A, Heun P: **Drosophila CENH3 is**
844 **sufficient for centromere formation.** *Science* 2011, **334:**686-690.
845 11. McKinley KL, Cheeseman IM: **The molecular basis for centromere identity and**
846 **function.** *Nat Rev Mol Cell Biol* 2016, **17:**16-29.
847 12. Blower MD, Sullivan BA, Karpen GH: **Conserved Organization of Centromeric**
848 **Chromatin in Flies and Humans.** *Developmental Cell* 2002, **2:**319-330.
849 13. Ribeiro SA, Vagnarelli P, Dong Y, Hori T, McEwen BF, Fukagawa T, Flors C, Earnshaw
850 WC: **A super-resolution map of the vertebrate kinetochore.** *Proc Natl Acad Sci U S A*
851 2010, **107:**10484-10489.
852 14. Bergmann JH, Rodríguez MG, Martins NMC, Kimura H, Kelly DA, Masumoto H,
853 Larionov V, Jansen LET, Earnshaw WC: **Epigenetic engineering shows H3K4me2 is**
854 **required for HJURP targeting and CENP-A assembly on a synthetic human**
855 **kinetochore.** *EMBO J* 2011, **30:**328-340.
856 15. Martins NMC, Bergmann JH, Shono N, Kimura H, Larionov V, Masumoto H, Earnshaw
857 WC: **Epigenetic engineering shows that a human centromere resists silencing**
858 **mediated by H3K27me3/K9me3.** *Mol Biol Cell* 2016, **27:**177-196.
859 16. Lo AW, Magliano DJ, Sibson MC, Kalitsis P, Craig JM, Choo KH: **A novel chromatin**
860 **immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding**
861 **neocentromere DNA.** *Genome Res* 2001, **11:**448-457.
862 17. Alonso A, Fritz B, Hasson D, Abrusan G, Cheung F, Yoda K, Radlwimmer B, Ladurner
863 AG, Warburton PE: **Co-localization of CENP-C and CENP-H to discontinuous**
864 **domains of CENP-A chromatin at human neocentromeres.** *Genome Biol* 2007,
865 **8:**R148.
866 18. Hasson D, Panchenko T, Salimian KJ, Salman MU, Sekulic N, Alonso A, Warburton PE,
867 Black BE: **The octamer is the major form of CENP-A nucleosomes at human**
868 **centromeres.** *Nat Struct Mol Biol* 2013, **20:**687-695.
869 19. Nergadze SG, Piras FM, Gamba R, Corbo M, Cerutti F, McCarter JGW, Cappelletti E,
870 Gozzo F, Harman RM, Antczak DF, et al: **Birth, evolution, and transmission of**
871 **satellite-free mammalian centromeric domains.** *Genome Res* 2018, **28:**789-799.
872 20. Kasai F, Garcia C, Arruga MV, Ferguson-Smith MA: **Chromosome homology between**
873 **chicken (*Gallus gallus domesticus*) and the red-legged partridge (*Alectoris rufa*);**
874 **evidence of the occurrence of a neocentromere during evolution.** *Cytogenet Genome*
875 *Res* 2003, **102:**326-330.
876 21. Marshall OJ, Chueh AC, Wong LH, Choo KHA: **Neocentromeres: new insights into**
877 **centromere structure, disease development, and karyotype evolution.** *Am J Hum*
878 *Genet* 2008, **82:**261-282.
879 22. Rocchi M, Archidiacono N, Schempp W, Capozzi O, Stanyon R: **Centromere**
880 **repositioning in mammals.** *Heredity* 2012, **108:**59-67.

881 23. Miller JT, Dong F, Jackson SA, Song J, Jiang J: **Retrotransposon-related DNA**
882 **sequences in the centromeres of grass chromosomes.** *Genetics* 1998, **150:**1615-1623.
883 24. Presting GG, Malysheva L, Fuchs J, Schubert I: **A Ty3/gypsy retrotransposon-like**
884 **sequence localizes to the centromeric regions of cereal chromosomes.** *Plant J* 1998,
885 **16:**721-728.
886 25. Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J:
887 **Functional rice centromeres are marked by a satellite repeat and a centromere-**
888 **specific retrotransposon.** *Plant Cell* 2002, **14:**1691-1704.
889 26. Nagaki K, Cheng Z, Ouyang S, Talbert PB, Kim M, Jones KM, Henikoff S, Buell CR,
890 Jiang J: **Sequencing of a rice centromere uncovers active genes.** *Nat Genet* 2004,
891 **36:**138-145.
892 27. Nagaki K, Song J, Stupar RM, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones
893 KM, Dawe RK, et al: **Molecular and cytological analyses of large tracks of**
894 **centromeric DNA reveal the structure and evolutionary dynamics of maize**
895 **centromeres.** *Genetics* 2003, **163:**759-770.
896 28. Ferreri GC, Brown JD, Obergfell C, Jue N, Finn CE, O'Neill MJ, O'Neill RJ: **Recent**
897 **amplification of the kangaroo endogenous retrovirus, KERV, limited to the**
898 **centromere.** *J Virol* 2011, **85:**4761-4771.
899 29. Ferreri GC, Marzelli M, Rens W, O'Neill RJ: **A centromere-specific retroviral element**
900 **associated with breaks of synteny in macropodine marsupials.** *Cytogenet Genome Res*
901 2004, **107:**115-118.
902 30. de Sotero-Caio CG, Cabral-de-Mello DC, Calixto MdS, Valente GT, Martins C, Loreto
903 V, de Souza MJ, Santos N: **Centromeric enrichment of LINE-1 retrotransposons and**
904 **its significance for the chromosome evolution of Phyllostomid bats.** *Chromosome Res*
905 2017, **25:**313-325.
906 31. Carbone L, Harris RA, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, Meyer
907 TJ, Herrero J, Roos C, Aken B, et al: **Gibbon genome and the fast karyotype evolution**
908 **of small apes.** *Nature* 2014, **513:**195-201.
909 32. Carbone L, Harris RA, Mootnick AR, Milosavljevic A, Martin DIK, Rocchi M, Capozzi
910 O, Archidiacono N, Konkel MK, Walker JA, et al: **Centromere remodeling in *Hoolock***
911 ***leuconedys* (Hylobatidae) by a new transposable element unique to the gibbons.**
912 *Genome Biol Evol* 2012, **4:**648-658.
913 33. Contreras-Galindo R, Kaplan MH, He S, Contreras-Galindo AC, Gonzalez-Hernandez
914 MJ, Kappes F, Dube D, Chan SM, Robinson D, Meng F, et al: **HIV infection reveals**
915 **widespread expansion of novel centromeric human endogenous retroviruses.**
916 *Genome Res* 2013, **23:**1505-1513.
917 34. Zahn J, Kaplan MH, Fischer S, Dai M, Meng F, Saha AK, Cervantes P, Chan SM, Dube
918 D, Omenn GS, et al: **Expansion of a novel endogenous retrovirus throughout the**
919 **pericentromeres of modern humans.** *Genome Biol* 2015, **16:**74.
920 35. Klein SJ, O'Neill RJ: **Transposable elements: genome innovation, chromosome**
921 **diversity, and centromere conflict.** *Chromosome Res* 2018, **26:**5-23.
922 36. Birchler JA, Presting GG: **Retrotransposon insertion targeting: a mechanism for**
923 **homogenization of centromere sequences on nonhomologous chromosomes.** *Genes*
924 *Dev* 2012, **26:**638-640.

925 37. Sultana T, Zamborlini A, Cristofari G, Lesage P: **Integration site selection by**
926 **retroviruses and transposable elements in eukaryotes.** *Nature Reviews Genetics* 2017,
927 **18:**292-308.
928 38. Wong LH, Choo KHA: **Evolutionary dynamics of transposable elements at the**
929 **centromere.** *Trends Genet* 2004, **20:**611-616.
930 39. Sharma A, Wolfgruber TK, Presting GG: **Tandem repeats derived from centromeric**
931 **retrotransposons.** *BMC Genomics* 2013, **14**.
932 40. Bourc'his D, Bestor TH: **Meiotic catastrophe and retrotransposon reactivation in**
933 **male germ cells lacking Dnmt3L.** *Nature* 2004, **431:**96-99.
934 41. Kato M, Takashima K, Kakutani T: **Epigenetic control of CACTA transposon mobility**
935 **in Arabidopsis thaliana.** *Genetics* 2004, **168:**961-969.
936 42. Ariumi Y: **Guardian of the Human Genome: Host Defense Mechanisms against**
937 **LINE-1 Retrotransposition.** *Front Chem* 2016, **4:**28.
938 43. Chan FL, Wong LH: **Transcription in the maintenance of centromere chromatin**
939 **identity.** *Nucleic Acids Res* 2012, **40:**11178-11188.
940 44. Wong LH, Brettingham-Moore KH, Chan L, Quach JM, Anderson MA, Northrop EL,
941 Hannan R, Saffery R, Shaw ML, Williams E, Choo KHA: **Centromere RNA is a key**
942 **component for the assembly of nucleoproteins at the nucleolus and centromere.**
943 *Genome Res* 2007, **17:**1146-1160.
944 45. Quénet D, Dalal Y: **A long non-coding RNA is required for targeting centromeric**
945 **protein A to the human centromere.** *Elife* 2014, **3:**e03254.
946 46. Catania S, Pidoux AL, Allshire RC: **Sequence features and transcriptional stalling**
947 **within centromere DNA promote establishment of CENP-A chromatin.** *PLoS Genet*
948 2015, **11:**e1004986.
949 47. Chen C-C, Bowers S, Lipinszki Z, Palladino J, Trusiak S, Bettini E, Rosin L, Przewloka
950 MR, Glover DM, O'Neill RJ, Mellone BG: **Establishment of Centromeric Chromatin**
951 **by the CENP-A Assembly Factor CAL1 Requires FACT-Mediated Transcription.**
952 *Dev Cell* 2015, **34:**73-84.
953 48. Hall LE, Mitchell SE, O'Neill RJ: **Pericentric and centromeric transcription: a**
954 **perfect balance required.** *Chromosome Research* 2012, **20:**535-546.
955 49. Talbert PB, Henikoff S: **Transcribing Centromeres: Noncoding RNAs and**
956 **Kinetochore Assembly.** *Trends in Genetics* 2018, **34:**587-599.
957 50. Chuong EB, Elde NC, Feschotte C: **Regulatory activities of transposable elements:**
958 **from conflicts to benefits.** *Nat Rev Genet* 2017, **18:**71-86.
959 51. Jain M, Olsen HE, Turner DJ, Stoddart D, Bulazel KV, Paten B, Haussler D, Willard HF,
960 Akeson M, Miga KH: **Linear assembly of a human centromere on the Y**
961 **chromosome.** *Nature Biotechnology* 2018, **36:**321-323.
962 52. Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri
963 L, Dishuck PC, Rhie A, et al: **The structure, function and evolution of a complete**
964 **human chromosome 8.** *Nature* 2021, **593:**101-107.
965 53. Chang C-H, Chavan A, Palladino J, Wei X, Martins NMC, Santinello B, Chen C-C,
966 Erceg J, Beliveau BJ, Wu C-T, et al: **Islands of retroelements are major components**
967 **of Drosophila centromeres.** *PLoS Biol* 2019, **17:**e3000241.
968 54. Belser C, Baurens F-C, Noel B, Martin G, Cruaud C, Istace B, Yahiaoui N, Labadie K,
969 Hřibová E, Doležel J, et al: **Telomere-to-telomere gapless chromosomes of banana**
970 **using nanopore sequencing.** *Commun Biol* 2021, **4:**1047.

971 55. Wang B, Yang X, Jia Y, Xu Y, Jia P, Dang N, Wang S, Xu T, Zhao X, Gao S, et al:
972 **High-quality Arabidopsis thaliana Genome Assembly with Nanopore and HiFi Long**
973 **Reads.** *Genomics Proteomics Bioinformatics* 2021.
974 56. Naish M, Alonge M, Wlodzimierz P, Tock AJ, Abramson BW, Schmucker A,
975 Mandakova T, Jamge B, Lambing C, Kuo P, et al: **The genetic and epigenetic**
976 **landscape of the Arabidopsis centromeres.** *Science* 2021, **374:**eabi7489.
977 57. Hales KG, Korey CA, Larracuente AM, Roberts DM: **Genetics on the Fly: A Primer on**
978 **the Drosophila Model System.** *Genetics* 2015, **201:**815-842.
979 58. Rahman R, Chirn G-W, Kanodia A, Sytnikova YA, Brembs B, Bergman CM, Lau NC:
980 **Unique transposon landscapes are pervasive across *Drosophila melanogaster***
981 **genomes.** *Nucleic Acids Res* 2015, **43:**10655-10672.
982 59. Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L,
983 Sisu C, Li JJ, et al: **Comparative analysis of the transcriptome across distant species.**
984 *Nature* 2014, **512:**445-448.
985 60. Andersen PR, Tirian L, Vunjak M, Brennecke J: **A heterochromatin-dependent**
986 **transcription machinery drives piRNA expression.** *Nature* 2017, **549:**54-59.
987 61. Laktionov PP, Maksimov DA, Romanov SE, Antoshina PA, Posukh OV, White-Cooper
988 H, Koryakov DE, Belyakin SN: **Genome-wide analysis of gene regulation mechanisms**
989 **during Drosophila spermatogenesis.** *Epigenetics Chromatin* 2018, **11:**14.
990 62. Tettey TT, Gao X, Shao W, Li H, Story BA, Chitsazan AD, Glaser RL, Goode ZH,
991 Seidel CW, Conaway RC, et al: **A Role for FACT in RNA Polymerase II Promoter-**
992 **Proximal Pausing.** *Cell Rep* 2019, **27:**3770-3779.e3777.
993 63. Metcalfe CJ, Casane D: **Modular organization and reticulate evolution of the ORF1**
994 **of Jockey superfamily transposable elements.** *Mobile DNA* 2014, **5:**19.
995 64. Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, Tarone AM, Turlapati L, Zichner
996 T, Zhu D, Lyman RF, et al: **Natural variation in genome architecture among 205**
997 ***Drosophila melanogaster* Genetic Reference Panel lines.** *Genome Res* 2014, **24:**1193-
998 1208.
999 65. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y,
1000 Magwire MM, Cridland JM, et al: **The *Drosophila melanogaster* Genetic Reference**
1001 **Panel.** *Nature* 2012, **482:**173-178.
1002 66. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H,
1003 Jamilloux V, Quesneville H, González J: **Population-scale long-read sequencing**
1004 **uncovers transposable elements associated with gene expression variation and**
1005 **adaptive signatures in Drosophila.** *Nat Commun* 2022, **13:**1948.
1006 67. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD: **Structural variants exhibit**
1007 **widespread allelic heterogeneity and shape variation in complex traits.** *Nat Commun*
1008 2019, **10:**4872.
1009 68. Long E, Evans C, Chaston J, Udall JA: **Genomic Structural Variations Within Five**
1010 **Continental Populations of *Drosophila melanogaster*.** *G3: Genes|Genomes|Genetics*
1011 2018, **8:**3247-3253.
1012 69. Nelson MG, Linheiro RS, Bergman CM: **McClintock: An Integrated Pipeline for**
1013 **Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing**
1014 **Data.** *G3* 2017, **7:**2763-2778.
1015 70. Blumenstiel JP, Chen X, He M, Bergman CM: **An age-of-allele test of neutrality for**
1016 **transposable element insertions.** *Genetics* 2014, **196:**523-538.

1017 71. Kofler R, Betancourt AJ, Schlötterer C: **Sequencing of pooled DNA samples (Pool-Seq)**
1018 **uncovers complex dynamics of transposable element insertions in *Drosophila***
1019 ***melanogaster*.** *PLoS Genet* 2012, **8:**e1002487.
1020 72. Kofler R, Nolte V, Schlötterer C: **Tempo and Mode of Transposable Element Activity**
1021 **in Drosophila.** *PLoS Genet* 2015, **11:**e1005406.
1022 73. Chang C-H, Larracuente AM: **Heterochromatin-Enriched Assemblies Reveal the**
1023 **Sequence and Organization of the *Drosophila melanogaster* Y Chromosome.**
1024 *Genetics* 2019, **211:**333-348.
1025 74. Ghildiyal M, Zamore PD: **Small silencing RNAs: an expanding universe.** *Nat Rev*
1026 *Genet* 2009, **10:**94-108.
1027 75. Ernst C, Odom DT, Kutter C: **The emergence of piRNAs against transposon invasion**
1028 **to preserve mammalian genome integrity.** *Nature Communications* 2017, **8**.
1029 76. Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M: **Recurrent insertion and**
1030 **duplication generate networks of transposable element sequences in the *Drosophila***
1031 ***melanogaster* genome.** *Genome Biol* 2006, **7:**R112.
1032 77. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ:
1033 **Discrete small RNA-generating loci as master regulators of transposon activity in**
1034 **Drosophila.** *Cell* 2007, **128:**1089-1103.
1035 78. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H,
1036 Siomi MC: **A slicer-mediated mechanism for repeat-associated siRNA 5' end**
1037 **formation in Drosophila.** *Science* 2007, **315:**1587-1590.
1038 79. Mohn F, Sienski G, Handler D, Brennecke J: **The rhino-deadlock-cutoff complex**
1039 **licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila.**
1040 *Cell* 2014, **157:**1364-1379.
1041 80. Parhad SS, Tu S, Weng Z, Theurkauf WE: **Adaptive Evolution Leads to Cross-Species**
1042 **Incompatibility in the piRNA Transposon Silencing Machinery.** *Dev Cell* 2017,
1043 **43:**60-70.e65.
1044 81. Rozhkov NV, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie
1045 WR, Hannon GJ, Evgen'ev MB: **Small RNA-based silencing strategies for transposons**
1046 **in the process of invading Drosophila species.** *RNA* 2010, **16:**1634-1645.
1047 82. Quénerch'du E, Anand A, Kai T: **The piRNA pathway is developmentally regulated**
1048 **during spermatogenesis in Drosophila.** *RNA* 2016, **22:**1044-1054.
1049 83. Wei X, Eickbush DG, Speece I, Larracuente AM: **Heterochromatin-dependent**
1050 **transcription of satellite DNAs in the *Drosophila melanogaster* female germline.** *Elife*
1051 2021, **10**.
1052 84. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J: **Population genomics**
1053 **of transposable elements in *Drosophila melanogaster*.** *Mol Biol Evol* 2011, **28:**1633-
1054 1644.
1055 85. Montgomery E, Charlesworth B, Langley CH: **A test for the role of natural selection in**
1056 **the stabilization of transposable element copy number in a population of *Drosophila***
1057 ***melanogaster*.** *Genetical Research* 1987, **49:**31-41.
1058 86. Charlesworth B, Campos JL: **The relations between recombination rate and patterns**
1059 **of molecular variation and evolution in Drosophila.** *Annu Rev Genet* 2014, **48:**383-
1060 403.
1061 87. Kent TV, Uzunović J, Wright SI: **Coevolution between transposable elements and**
1062 **recombination.** *Philos Trans R Soc Lond B Biol Sci* 2017, **372**.

1063 88. Bergman CM, Bensasson D: **Recent LTR retrotransposon insertion contrasts with**
1064 **waves of non-LTR insertion since speciation in *Drosophila melanogaster*.** *Proc Natl*
1065 *Acad Sci U S A* 2007, **104:**11340-11345.
1066 89. Blumenstiel JP, Hartl DL, Lozovsky ER: **Patterns of insertion and deletion in**
1067 **contrasting chromatin domains.** *Mol Biol Evol* 2002, **19:**2211-2225.
1068 90. Chang CH, Gregory LE, Gordon KE, Meiklejohn CD, Larracuente AM: **Unique**
1069 **structure and positive selection promote the rapid divergence of Drosophila Y**
1070 **chromosomes.** *Elife* 2022, **11**.
1071 91. Comeron JM, Ratnappan R, Bailin S: **The many landscapes of recombination in**
1072 ***Drosophila melanogaster*.** *PLoS Genet* 2012, **8:**e1002905.
1073 92. Miller DE, Smith CB, Kazemi NY, Cockrell AJ, Arvanitakis AV, Blumenstiel JP,
1074 Jaspersen SL, Hawley RS: **Whole-Genome Analysis of Individual Meiotic Events in**
1075 ***Drosophila melanogaster* Reveals That Noncrossover Gene Conversions Are**
1076 **Insensitive to Interference and the Centromere Effect.** *Genetics* 2016, **203:**159-171.
1077 93. Miller DE, Takeo S, Nandanan K, Paulson A, Gogol MM, Noll AC, Perera AG, Walton
1078 KN, Gilliland WD, Li H, et al: **A Whole-Chromosome Analysis of Meiotic**
1079 **Recombination in *Drosophila melanogaster*.** *G3* 2012, **2:**249-260.
1080 94. Khurana JS, Wang J, Xu J, Koppetsch BS, Thomson TC, Nowosielska A, Li C, Zamore
1081 PD, Weng Z, Theurkauf WE: **Adaptation to P Element Transposon Invasion in**
1082 ***Drosophila melanogaster*.** *Cell* 2011, **147:**1551-1563.
1083 95. Kofler R: **Dynamics of Transposable Element Invasions with piRNA Clusters.** *Mol*
1084 *Biol Evol* 2019, **36:**1457-1472.
1085 96. Said I, McGurk MP, Clark AG, Barbash DA: **Patterns of piRNA Regulation in**
1086 **Drosophila Revealed through Transposable Element Clade Inference.** *Mol Biol Evol*
1087 2022, **39**.
1088 97. Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C: **Molecular dissection of a**
1089 **natural transposable element invasion.** *Genome Res* 2018, **28:**824-835.
1090 98. Anxolabéhère D, Kidwell MG, Periquet G: **Molecular characteristics of diverse**
1091 **populations are consistent with the hypothesis of a recent invasion of *Drosophila***
1092 ***melanogaster* by mobile P elements.** *Mol Biol Evol* 1988, **5:**252-269.
1093 99. Hill T, Schlotterer C, Betancourt AJ: **Hybrid Dysgenesis in *Drosophila simulans***
1094 **Associated with a Rapid Invasion of the P-Element.** *PLoS Genet* 2016, **12:**e1005920.
1095 100. Wolfgruber TK, Sharma A, Schneider KL, Albert PS, Koo D-H, Shi J, Gao Z, Han F, Lee
1096 H, Xu R, et al: **Maize centromere structure and evolution: sequence analysis of**
1097 **centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons.**
1098 *PLoS Genet* 2009, **5:**e1000743.
1099 101. Schneider KL, Xie Z, Wolfgruber TK, Presting GG: **Inbreeding drives maize**
1100 **centromere evolution.** *Proc Natl Acad Sci U S A* 2016, **113:**E987-996.
1101 102. Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, Toyoda A, Fujiyama A,
1102 Tarutani Y, Kakutani T: **Centromere-targeted de novo integrations of an LTR**
1103 **retrotransposon of *Arabidopsis lyrata*.** *Genes Dev* 2012, **26:**705-713.
1104 103. Tsukahara S, Kobayashi A, Kawabe A, Mathieu O, Miura A, Kakutani T: **Bursts of**
1105 **retrotransposition reproduced in Arabidopsis.** *Nature* 2009, **461:**423-426.
1106 104. Langley CH, Montgomery E, Hudson R, Kaplan N, Charlesworth B: **On the role of**
1107 **unequal exchange in the containment of transposable element copy number.** *Genet*
1108 *Res* 1988, **52:**223-235.

1109 105. Hollister JD, Gaut BS: **Epigenetic silencing of transposable elements: a trade-off**
1110 **between reduced transposition and deleterious effects on neighboring gene**
1111 **expression.** *Genome Res* 2009, **19:**1419-1428.
1112 106. Lee YCG: **The Role of piRNA-Mediated Epigenetic Silencing in the Population**
1113 **Dynamics of Transposable Elements in *Drosophila melanogaster*.** *PLoS Genet* 2015,
1114 **11:**e1005269.
1115 107. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H: **Genome-wide evidence for**
1116 **local DNA methylation spreading from small RNA-targeted sequences in**
1117 **Arabidopsis.** *Nucleic Acids Res* 2011, **39:**6919-6931.
1118 108. Cridland JM, Thornton KR, Long AD: **Gene expression variation in *Drosophila***
1119 ***melanogaster* due to rare transposable element insertion alleles of large effect.**
1120 *Genetics* 2015, **199:**85-93.
1121 109. Srivastava S, Foltz DR: **Posttranslational modifications of CENP-A: marks of**
1122 **distinction.** *Chromosoma* 2018, **127:**279-290.
1123 110. Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ:
1124 **piRNA-Guided Genome Defense: From Biogenesis to Silencing.** *Annu Rev Genet*
1125 2018, **52:**131-157.
1126 111. Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD: **PIWI-interacting RNAs:**
1127 **small RNAs with big functions.** *Nat Rev Genet* 2019, **20:**89-108.
1128 112. Onishi R, Yamanaka S, Siomi MC: **piRNA- and siRNA-mediated transcriptional**
1129 **repression in Drosophila, mice, and yeast: new insights and biodiversity.** *EMBO Rep*
1130 2021, **22:**e53062.
1131 113. Scelfo A, Fachinetti D: **Keeping the Centromere under Control: A Promising Role**
1132 **for DNA Methylation.** *Cells* 2019, **8**.
1133 114. Altemose N, Logsdon GA, Bzikadze AV, Sidhwani P, Langley SA, Caldas GV, Hoyt SJ,
1134 Uralsky L, Ryabov FD, Shew CJ, et al: **Complete genomic and epigenetic maps of**
1135 **human centromeres.** *Science* 2022, **376:**eabl4178.
1136 115. Doolittle WF, Ford Doolittle W, Sapienza C: **Selfish genes, the phenotype paradigm**
1137 **and genome evolution.** *Nature* 1980, **284:**601-603.
1138 116. Orgel LE, Crick FH: **Selfish DNA: the ultimate parasite.** *Nature* 1980, **284:**604-607.
1139 117. Dimitri P, Junakovic N: **Revising the selfish DNA hypothesis: new evidence on**
1140 **accumulation of transposable elements in heterochromatin.** *Trends Genet* 1999,
1141 **15:**123-124.
1142 118. Cosby RL, Chang N-C, Feschotte C: **Host-transposon interactions: conflict,**
1143 **cooperation, and cooption.** *Genes Dev* 2019, **33:**1098-1116.
1144 119. Danilevskaya ON, Slot F, Traverse KL, Hogan NC, Pardue ML: **Drosophila telomere**
1145 **transposon HeT-A produces a transcript with tightly bound protein.** *Proc Natl Acad*
1146 *Sci U S A* 1994, **91:**6679-6682.
1147 120. Mason JM, Biessmann H: **The unusual telomeres of Drosophila.** *Trends Genet* 1995,
1148 **11:**58-62.
1149 121. Pardue M-L, DeBaryshe PG: **Drosophila telomeres: A variation on the telomerase**
1150 **theme.** *Fly* 2008, **2:**101-110.
1151 122. Pardue ML, Debaryshe PG: **Drosophila telomere transposons: genetically active**
1152 **elements in heterochromatin.** *Genetica* 2000, **109:**45-52.
1153 123. Saint-Leandre B, Nguyen SC, Levine MT: **Diversification and collapse of a telomere**
1154 **elongation mechanism.** *Genome Res* 2019, **29:**920-931.

1155 124. McGurk MP, Dion-Côté A-M, Barbash DA: **Rapid evolution at the Drosophila**
1156 **telomere: transposable element dynamics at an intrinsically unstable locus.** *Genetics*
1157 2021, **217**.
1158 125. Moazed D, Bühler M, Buker SM, Colmenares SU, Gerace EL, Gerber SA, Hong EJE,
1159 Motamedi MR, Verdel A, Villén J, Gygi SP: **Studies on the mechanism of RNAi-**
1160 **dependent heterochromatin assembly.** *Cold Spring Harb Symp Quant Biol* 2006,
1161 **71:**461-471.
1162 126. Choi ES, Strålfors A, Castillo AG, Durand-Dubief M, Ekwall K, Allshire RC:
1163 **Identification of noncoding transcripts from within CENP-A chromatin at fission**
1164 **yeast centromeres.** *J Biol Chem* 2011, **286:**23600-23607.
1165 127. Chan FL, Marshall OJ, Saffery R, Kim BW, Earle E, Choo KHA, Wong LH: **Active**
1166 **transcription and essential role of RNA polymerase II at the centromere during**
1167 **mitosis.** *Proc Natl Acad Sci U S A* 2012, **109:**1979-1984.
1168 128. Rošić S, Köhler F, Erhardt S: **Repetitive centromeric satellite RNA is essential for**
1169 **kinetochore formation and cell division.** *J Cell Biol* 2014, **207:**335-349.
1170 129. Bobkov GOM, Gilbert N, Heun P: **Centromere transcription allows CENP-A to**
1171 **transit from chromatin association to stable incorporation.** *J Cell Biol* 2018,
1172 **217:**1957-1972.
1173 130. Ohkuni K, Kitagawa K: **Endogenous transcription at the centromere facilitates**
1174 **centromere activity in budding yeast.** *Curr Biol* 2011, **21:**1695-1703.
1175 131. Sandler L, Novitski E: **Meiotic Drive as an Evolutionary Force.** *The American*
1176 *Naturalist* 1957, **91:**105-110.
1177 132. Akera T, Chmátal L, Trimm E, Yang K, Aonbangkhen C, Chenoweth DM, Janke C,
1178 Schultz RM, Lampson MA: **Spindle asymmetry drives non-Mendelian chromosome**
1179 **segregation.** *Science* 2017, **358:**668-672.
1180 133. Akera T, Trimm E, Lampson MA: **Molecular Strategies of Meiotic Cheating by Selfish**
1181 **Centromeres.** *Cell* 2019, **178:**1132-1144.e1110.
1182 134. Henikoff S, Ahmad K, Malik HS: **The centromere paradox: stable inheritance with**
1183 **rapidly evolving DNA.** *Science* 2001, **293:**1098-1102.
1184 135. Malik HS, Henikoff S: **Adaptive Evolution of Cid, a Centromere-Specific Histone in**
1185 **Drosophila.** *Genetics* 2001, **157:**1293-1298.
1186 136. Kumon T, Ma J, Akins RB, Stefanik D, Nordgren CE, Kim J, Levine MT, Lampson MA:
1187 **Parallel pathways for recruiting effector proteins determine centromere drive and**
1188 **suppression.** *Cell* 2021, **184:**4904-4918.e4911.
1189 137. Kipling D, Warburton PE: **Centromeres, CENP-B and Tigger too.** *Trends Genet* 1997,
1190 **13:**141-145.
1191 138. Khost DE, Eickbush DG, Larracuente AM: **Single-molecule sequencing resolves the**
1192 **detailed structure of complex satellite DNA loci in *Drosophila melanogaster.*** *Genome*
1193 *Res* 2017, **27:**709-721.
1194 139. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
1195 **features.** *Bioinformatics* 2010, **26:**841-842.
1196 140. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper
1197 A, Markowitz S, Duran C, et al: **Geneious Basic: an integrated and extendable**
1198 **desktop software platform for the organization and analysis of sequence data.**
1199 *Bioinformatics* 2012, **28:**1647-1649.

1200 141. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2.** *Nat Methods*
1201 2012, **9:**357-359.
1202 142. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,
1203 Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format**
1204 **and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.
1205 143. Buske FA, Bodén M, Bauer DC, Bailey TL: **Assigning roles to DNA regulatory motifs**
1206 **using comparative genomics.** *Bioinformatics* 2010, **26:**860-866.
1207 144. Langmead B: **Aligning short sequencing reads with Bowtie.** *Curr Protoc*
1208 *Bioinformatics* 2010, **Chapter 11:**Unit 11.17.
1209 145. Han BW, Wang W, Zamore PD, Weng Z: **piPipes: a set of pipelines for piRNA and**
1210 **transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq,**
1211 **ChIP-seq and genomic DNA sequencing.** *Bioinformatics* 2015, **31:**593-595.
1212 146. Long AD, Macdonald SJ, King EG: **Dissecting complex traits using the Drosophila**
1213 **Synthetic Population Resource.** *Trends Genet* 2014, **30:**488-495.
1214 147. Grenier JK, Roman Arguello J, Moreira MC, Gottipati S, Mohammed J, Hackett SR,
1215 Boughton R, Greenberg AJ, Clark AG: **Global Diversity Lines–A Five-Continent**
1216 **Reference Panel of Sequenced Drosophila melanogaster Strains.** *G3:*
1217 *Genes|Genomes|Genetics* 2015, **5:**593-603.
1218 148. Pedersen BS, Quinlan AR: **Mosdepth: quick coverage calculation for genomes and**
1219 **exomes.** *Bioinformatics* 2018, **34:**867-868.
1220 149. Navarro-Dominguez B, Chang C-H, Brand CL, Muirhead CA, Presgraves DC,
1221 Larracuente AM: **Epistatic selection on a selfish supergene - drive, recombination,**
1222 **and genetic load.** *Elife* 2022, **11**.
1223 150. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,
1224 Lunter G, Marth GT, Sherry ST, et al: **The variant call format and VCFtools.**
1225 *Bioinformatics* 2011, **27:**2156-2158.
1226 151. Knaus BJ, Grünwald NJ: **vcfr: a package to manipulate and visualize variant call**
1227 **format data in R.** *Mol Ecol Resour* 2017, **17:**44-53.
1228 152. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable**
1229 **and accurate long-read assembly via adaptive -mer weighting and repeat**
1230 **separation.** *Genome Res* 2017, **27:**722-736.
1231 153. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
1232 **34:**3094-3100.
1233 154. Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ: **Contiguous and accurate**
1234 **de novo assembly of metazoan genomes with modest long read coverage.** *Nucleic*
1235 *Acids Res* 2016, **44:**e147.
1236 155. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly**
1237 **from long uncorrected reads.** *Genome Res* 2017, **27:**737-746.
1238 156. King EG, Merkes CM, McNeil CL, Hoofer SR, Sen S, Broman KW, Long AD,
1239 Macdonald SJ: **Genetic dissection of a model complex trait using the Drosophila**
1240 **Synthetic Population Resource.** *Genome Res* 2012, **22:**1558-1566.
1241 157. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL:
1242 **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
1243 158. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search**
1244 **tool.** *J Mol Biol* 1990, **215:**403-410.

159. Bailey TL, Johnson J, Grant CE, Noble WS: **The MEME Suite.** *Nucleic Acids Research* 2015, **43:**W39-W49.
160. Gel B, Diez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R: **regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests.** *Bioinformatics* 2016, **32:**289-291.
161. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Mol Biol Evol* 2013, **30:**772-780.
162. Stamatakis A: **RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.** *Bioinformatics* 2014, **30:**1312-1313.
163. 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.
164. Ohta T: **Allelic and nonallelic homology of a supergene family.** *Proc Natl Acad Sci U S A* 1982, **79:**3251-3254.
165. 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.
166. **Biostrings: Efficient manipulation of biological strings.** [https://bioconductor.org/packages/Biostrings]
167. Duchen P, Zivkovic D, Hutter S, Stephan W, Laurent S: **Demographic inference reveals African and European admixture in the North American Drosophila melanogaster population.** *Genetics* 2013, **193:**291-301.
168. Li H, Stephan W: **Inferring the demographic history and rate of adaptive substitution in Drosophila.** *PLoS Genet* 2006, **2:**e166.