1 **Evolutionary dynamics of abundant 7 bp satellites in the genome of *Drosophila virilis***

2

3 Jullien M. Flynn[1], Manyuan Long[2], Rod A. Wing[3], Andrew G. Clark[1]

4

5 [1]Department of Molecular Biology and Genetics, Cornell University, Ithaca, USA

6 [2]Department of Ecology and Evolution, University of Chicago, Chicago, USA

7 [3]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, Arizona,

8 USA

Abstract

The factors that drive the rapid changes in satellite DNA genomic composition we see in eukaryotes are not well understood. *Drosophila virilis* has one of the highest relative amounts of simple satellites of any organism that has been studied, with an estimated 40+% of its genome composed of a few related 7 bp satellites. Here we use *D. virilis* as a model to understand technical biases affecting satellite sequencing and the evolutionary processes that drive satellite composition. By analyzing sequencing data from Illumina, PacBio, and Nanopore platforms, we identify platform-specific biases and suggest best practices for accurate characterization of satellites by sequencing. We use comparative genomics and cytogenetics to demonstrate that the satellite family arose in the branch leading to the virilis phylad 4.5 - 11 million years ago. The most abundant satellite is conserved in sequence and location in the pericentromeric region but has diverged widely in abundance among species, whereas the satellites nearest the centromere are rapidly turning over in sequence composition. The amount of satellite DNA is lowest in the species with the smallest effective population size in the *D. virilis* phylad, contrary to what would be expected if satellites were only evolving under weak negative selection. We found evidence for transposable elements flanking the satellite arrays as well as recent insertions into the arrays. By analyzing multiple strains of *D. virilis*, we saw that one centromere-proximal satellite is undergoing an expansion, while the other is contracting in a correlated manner, suggesting possible trade-offs or competition between satellites. In conclusion, we illuminate several key attributes of satellite evolutionary dynamics

that we hypothesize to be driven by processes like positive selection, meiotic drive, and constraints on satellite sequence and abundance.

Author summary

Tandem arrays of sequence repeats, or satellite DNA, makes up large and variable proportions of eukaryote genomes. Satellites may be an important component of genomes, but their evolution is not well understood because they are subject to a variety of different evolutionary forces due to their repetitive nature and proximity to centromeres, and they are difficult to sequence. Here we use a fruit fly *Drosophila virilis* – whose genome experienced a recent explosion in satellite DNA to make up 40+% of its genome – to understand the evolutionary processes that drive satellite composition. We found that the satellite emerged and expanded rapidly, and is variable among related species. The satellites have different patterns of evolution depending on their proximity to the centromere. We show evidence that the satellites may act selfishly and result in trade-offs on other satellites due to the push-back of selection.

Introduction

Repetitive DNA is abundant in most eukaryotic genomes, and is now understood to be correlated with the manifold variation in genome size across the tree of life (Elliott and Ryan Gregory 2015). For most species, transposable elements (TEs) dominate the repeat landscape, including in humans, plants, and *Drosophila melanogaster*. Satellite DNA, which is characterized by tandem repeats spanning long arrays, very rarely has dominated a genome to a similar

51 extent as TEs. An unprecedented case is that of *Drosophila virilis*, the Drosophila species with
52 the largest estimated genome size (up to 389 Mb) (Bosco *et al.* 2007), where some 40% of the
53 genome is comprised of just three simple 7-mer satellites: AAACTAC, AAACTAT, and AAATTAC
54 (Gall *et al.* 1971; Gall and Atherton 1974). Since the 1970s, there has been no follow-up to
55 validate the amount of 7-mers with modern techniques, or evolutionary studies to understand
56 how and why these satellite repeats expanded so explosively. The genomic composition of
57 simple satellites in *D. virilis* provides an excellent model for an investigation of the evolutionary
58 dynamics involved in their expansion in the genome as well as the technical challenges facing
59 simple satellite analysis.

60 Satellites are rapidly evolving in sequence and copy number, and there is a high level of
61 variation in satellite content among and within species (Wei *et al.* 2014, 2018). The reasons for
62 such dramatic variation is not well understood, and cannot be fully explained by current
63 models. Satellites have been long hypothesized to be slightly deleterious and therefore
64 governed primarily by the strength of negative selection (Ohno 1972). However, the amount of
65 satellite in the genome that causes negative effects that could be selected against depends on
66 many factors and cannot be easily predicted (Charlesworth *et al.* 1994; Gregory 2001). The fact
67 that most organisms have satellite repeats in or near centromeres suggests that they are
68 important for centromere function. Satellite repeats can also be important for maintenance of
69 the chromocenter and packaging of chromosomes in the nucleus (Jagannathan *et al.* 2018,
70 2019), and the transcripts of some satellites may be essential for fertility (Mills *et al*. 2019). In
71 heterozygotes with alleles that differ in pericentromeric satellite sequence or abundance, one
72 allele may assemble a stronger kinetochore during female meiosis I, increasing its probability of

73 transmission into the egg (rather than polar bodies). This transmission advantage, known as
74 centromere drive, allows satellites to rapidly change in composition in the population,
75 regardless of their whole-organism fitness effects (Henikoff 2001). If satellite DNA is an
76 essential component of genomes or is only a burden (i.e. is selfish), it is still not clear why some
77 species have almost no pericentromeric satellite DNA while others, like *D. virilis*, possess
78 pericentromeric satellites that make up almost half of the genome.

79 Comparing the satellites of *D. virilis* to those of its sister species can elucidate when the
80 abundant satellites arose, and how rapidly their copy numbers and sequences evolved. *D. virilis*
81 is 4.5 MY diverged from its sister species *D. novamexicana* and *D. americana*, which are both
82 restricted to North America, unlike globally-distributed *D. virilis* (Caletka and McAllister 2004).
83 *D. novamexicana* and *D. americana* have a smaller estimated genome size than *D. virilis* (~250
84 Mb vs. 389 Mb), suggesting these species may have less satellite content (Bosco *et al.* 2007).
85 Additionally, using intra-species comparisons across global populations can give indications
86 about factors that may be influencing satellite dynamics. For example, in *D. melanogaster*,
87 patterns of abundance of the *Prodsat* satellite closely mirror the migration patterns of species,
88 suggesting an ongoing expansion of this satellite (Wei *et al.* 2014). Genetic drift or meiotic drive
89 may contribute to patterns of geographical gradients of satellite abundance. We can also use
90 intra-species data to pose hypotheses about non-neutral processes that may be driving satellite
91 content. Previous work has shown evidence for conflicts or trade-offs between satellites within
92 the genome, and these constraints can be illuminated by analyzing satellites in several strains
93 (Flynn *et al.* 2017, 2018).

94 Genome-wide characterization of satellites has taken off since high-throughput
95 sequencing has become widely available. We have learned from several informative studies
96 about the sequences and relative abundances of satellites in various species (Pavlek *et al.* 2015;
97 Flynn *et al.* 2017; de Lima *et al.* 2017; Wei *et al.* 2018), but technical challenges may be
98 preventing accurate quantitative estimates in some cases. Satellites may be more prone to
99 errors or biases in the sequencing process that do not affect the better studied regions of the
100 genome. Satellites are difficult to assemble even with long-read sequencing (Chang and
101 Larracuente 2019). The genome assembly of *D. virilis* is approximately half its estimated
102 genome size by flow cytometry (~200 Mb vs 389 Mb) (Bosco *et al.* 2007), and it is likely that
103 much of what is missing is simple satellite DNA. However, even using alignment-free raw read
104 methods have not produced satellite DNA estimates that approach the amount that is missing
105 from the genome assembly and was estimated from early work (Gall *et al.* 1971; Gall and
106 Atherton 1974; Wei *et al.* 2018). Now, as long read sequencing is also being exploited to study
107 satellites, we must evaluate satellite DNA abundance estimates to assess if there are platform-
108 specific biases that may affect evolutionary analysis of satellite DNA.
109 The purpose of this paper is two-fold; first to explore the technical biases preventing
110 accurate characterization and quantification of simple satellites, and second to use a
111 comparative approach to understand the evolutionary dynamics of the extremely abundant
112 7mers in the *D. virilis* group. First, we characterize satellites in *D. virilis* sequencing data from
113 different platforms and assess biases that affect accurate satellite characterization. We then
114 use comparative genomics and cytogenetics in *D. virilis* and its sister species to understand the
115 composition and changes in the highly abundant simple satellites. Finally we sequence multiple

116 strains of *D. virilis* and sister species to estimate polymorphism in satellite abundance and infer
117 processes that may be influencing their evolution. From this we infer that there are likely a
118 variety of understudied processes affecting satellite DNA in this organism, including positive
119 selection, meiotic drive, and constraints and trade-offs between satellites.
120
121 **RESULTS**
122
123 **Technical biases in characterizing simple satellites from sequencing**
124 *Long-read genome assemblies have an under-representation of simple satellites*
125 Long-read sequencing technologies have an advantage because of their long reads, but a
126 disadvantage due to their high error rate, prompting a need for extensive alignments for error-
127 correction and assembly. First we asked whether assemblies from long read technologies can
128 better assemble simple satellite reads than the previous Sanger assembly. We compared the
129 amount of simple 7-mer satellites (AAACTAC, AAACTAT, AAATTAC, AAACAAC) in three *D. virilis*
130 genome assemblies: the CAF1 assembly produced from Sanger sequencing (Drosophila 12
131 Genomes Consortium *et al.* 2007), a PacBio assembly produced by our group by ~100x coverage
132 (available at https://www.ncbi.nlm.nih.gov/bioproject/?term=txid7214[Organism:noexp]), and a
133 Nanopore assembly produced from ~20x sequencing coverage (Miller *et al.* 2018). All
134 assemblies were approximately the same size at ~200 Mb. The PacBio and Nanopore
135 assemblies contained a similarly low amount of simple 7-mer satellites, 29 and 28 kb,
136 respectively. The CAF1 assembly, however contained 7.36 Mb of these satellites. This
137 discrepancy is likely largely due to the difference in assembly algorithms used for short read

138 and long read data. Long reads must be aligned and corrected to be incorporated into the
139 assembly because of their high error rate, whereas this is not necessary for Sanger-based
140 assemblies. Use of modified methods can improve assemblies of repetitive regions (Chang and
141 Larracuente 2019), but for highly homogeneous simple satellites, whose arrays span 10-100x
142 longer than the current maximum read length, it is practically impossible to produce a
143 continuous assembly.
144
145 *Simulations to assess simple repeat quantification from long read sequencing data*
146 Due to assembly issues of simple satellites, they must be quantified from raw unassembled
147 reads. Long read sequencing data poses a significant challenge because of the high error rate
148 including a high indel rate in the raw reads. We therefore used two different approaches along
149 with simulations to assess their accuracy. The first approach used k-Seek to select repeat-rich
150 reads and then Phobos to quantify satellites. This approach allows for *de novo* discovery of
151 satellite sequences. We used Noise-Cancelling Repeat Finder (NCRF, Harris *et al.* 2018) for our
152 second approach, providing our target satellites. Both methods are relatively sensitive to
153 imperfect repeats, which we expect with the high error rate of long-read sequencing.
154 To evaluate our approaches, we created a mock *D. virilis*-like genome containing
155 pericentromeric and centromeric repeats on each of 5 chromosomes (See Materials and
156 Methods). We then simulated 10x PacBio reads from this genome, and then quantified
157 satellites using both approaches. NCRF works by doing alignments of target satellites to the
158 reads and allowing up to a user-specified maximum divergence. To determine the most
159 appropriate maximum divergence, we simulated a range of values for this parameter from 18-

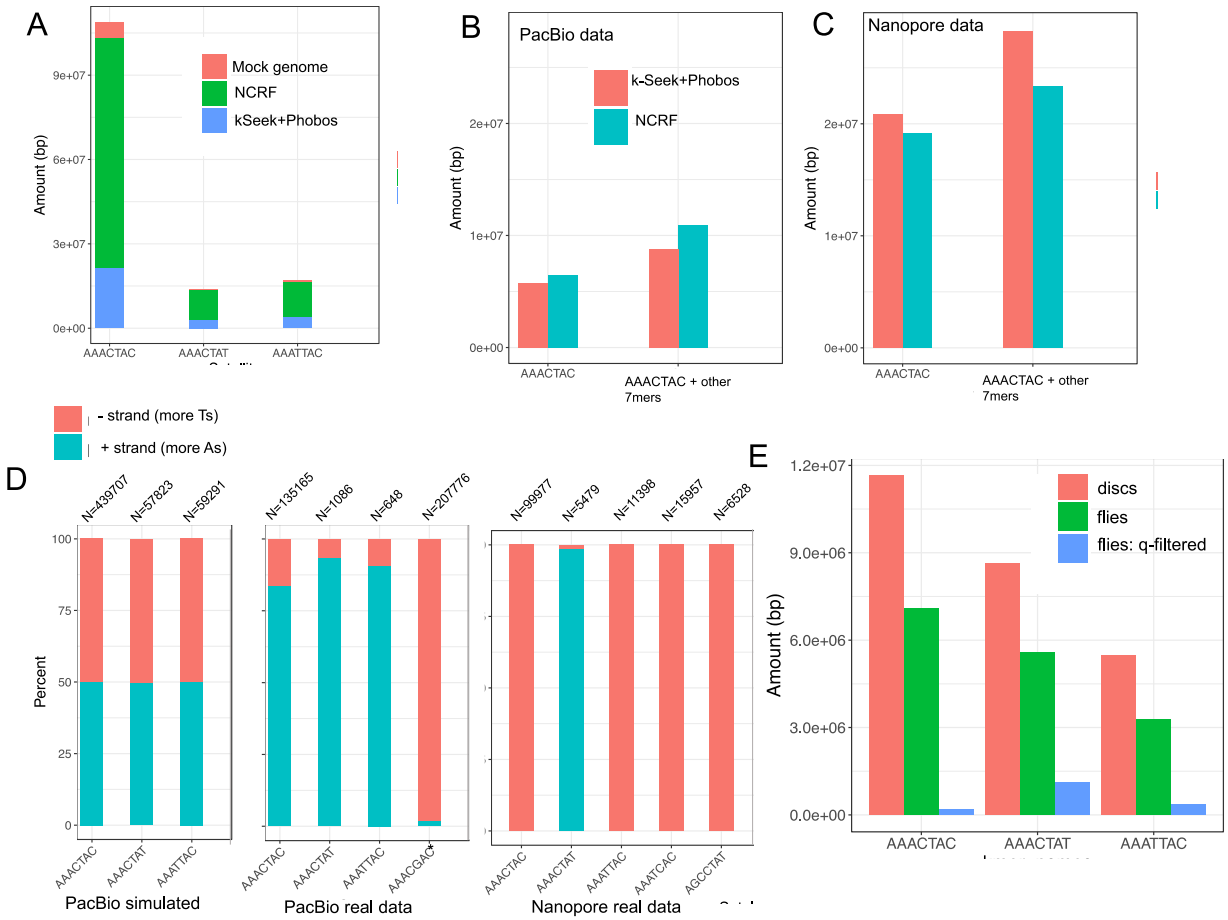
160 30% and chose the lowest asymptotic value - which was 25% in this case (Figure S1). NCRF
161 found almost the same amount of satellites that truly existed in the mock genome whereas the
162 k-Seek + Phobos method only found about 20% (Figure 1A).
163
164 *The amounts and biases in simple 7mer repeats differ between Nanopore and PacBio*
165 *sequencing reads*
166 Next, we quantified simple satellites in the long-read data generated from our 100x PacBio
167 sequencing and 20x Nanopore sequencing using the two approaches mentioned above. Unlike
168 in the simulations, both approaches produced very similar (but lower than expected) estimates
169 at 8.8-10.9 Mb for the PacBio data (Figure 1B). The Nanopore data contained almost 3 times
170 the 7mer satellites compared to PacBio, with 23.4 - 28.2 Mb (Figure 1C). This may represent a
171 platform-specific difference in the ability to sequence long arrays of simple tandem repeats.
172 Both the PacBio reads and the Nanopore reads contained a greater amount of simple satellites
173 than data produced in our lab previously with Illumina HiSeq sequencing (Wei et al. 2018).
174 Both the PacBio and Nanopore reads contained large amounts of what we expect to be
175 artifactual repeats, which were found with the k-Seek + Phobos approach, and validated with
176 NCRF. NCRF found 4.4 Mb (normalized to 1x genome coverage) of AAACGAC in the PacBio
177 reads. This satellite was not found in the Nanopore data or Illumina data (this and previous
178 studies) or in previous studies that characterized the most abundant satellites in *D. virilis*.
179 Manual inspection proved that the AAACGAC satellite was the true consensus found in long
180 arrays in the reads and did not represent an error in our approaches' characterization of
181 satellites. Similarly, AAATCAC, AGCCTAT, ACAGGCT, and AATGG were found in megabase

182 quantities (after normalization) in the Nanopore data - whereas these satellites were not found

183 in Illumina or PacBio data. We suggest these satellites are also technical artifacts introduced at

184 the base-calling level.

185



186

Fig1: Issues in quantifying simple satellites in sequencing data (all data shown is *D. virilis*). (**A**) Cumulative stacked barplot comparing the performance of the two tested approaches on PacBio data simulated with PBSim from a mock genome. (**B)** Comparing the results of the two approaches on the real PacBio data; "other" refers to additional satellites in the family, including suspected artefactual ones (AAAGCAC for PacBio and AAATCAC + AGCCTAT for Nanopore). (**C**) Same as B for Nanopore data. (**D**) Strand biases in the sequenced satellites in long read sequencing data. Satellites with asterisks are suspected artefactual ones. N refers to the number of satellite regions of reads used for the calculation. (E) Amount of satellites quantified in different datasets: imaginal discs (pure diploid), compared to flies (some polyteny), and fly data that has been quality filtered.

187 In the PacBio data, the relative amounts of 7mer satellites (AAACTAC, AAACTAT, and
188 AAATTAC) were lower than expected. This additional evidence led us to hypothesize that there
189 were context-specific errors in our PacBio data affecting our particular satellites.
190 If the sequencing were unbiased, we would expect to have an equal amount of satellites being
191 detected on reads coming from both DNA strands. We evaluated the strand bias in the
192 simulated and real long-read data for the three most abundant true satellites, as well as some
193 artifactual satellites. We arbitrarily label the positive strand as AAACTAC and the negative
194 strand as GTAGTTT, etc. In the simulated data, the positive and negative strands of satellites
195 were detected in equal amounts (Figure 1D). However, there was a strong strand bias for all
196 satellites in both the PacBio and Nanopore data (Figure 1D). For PacBio, the real satellites
197 AAACTAC, AAACTAT, AAATTAC had a positive strand bias, whereas the artifactual satellite had a
198 negative strand bias: 98% of the reads with this satellite were from the negative strand. Based
199 on communication with PacBio representatives, this issue seemed to be caused by context-
200 specific issues with base calling algorithms used for this sequencing run. As base calling
201 algorithms improve, these issues will likely begin to be remedied. In the Nanopore data, strand
202 biases were even more extreme: the negative strand was sequenced almost exclusively for real
203 satellites AAACTAC and AAATTAC and suspect satellite AAATCAC. However, the AAACTAT real
204 satellite was sequenced almost exclusively on the positive strand. In this case, strand biases
205 may be caused by unsequenceable secondary structures developing more frequently on one
206 strand of the satellite DNA than the other. We analyzed Illumina NextSeq reads for *D. virilis*, and
207 no such strand bias was found.
208
209 *D. virilis whole-flies have 40% less pericentromeric satellites than non-polytene tissue*

210
211 Polyteny occurs in all differentiated tissues of Dipterans, and is characterized by
212 multiple rounds of local DNA replication within the same nucleus and without cell division, a
213 process known as endoreduplication (Smith and Orr-Weaver 1991; Kim *et al.* 2011). However,
214 the pericentromeric heterochromatin, where most satellite DNA is located, is under-replicated
215 (Belyaeva *et al.* 1998). It has never been tested if the level of polyteny in an adult fly makes a
216 difference in the estimate of satellites per genome. Thus, we sequenced adult male flies (which
217 have multiple polytene tissues) and imaginal discs (which are diploid) from male larvae and
218 compared the amount of simple satellites in these datasets. We used Illumina sequencing and
219 PCR-free library preparations to reduce known PCR bias (Wei *et al.* 2018). We found that for
220 each of the four most abundant 7mer satellites in the *D. virilis* genome, there was
221 approximately 40% less in the flies compared to the imaginal discs (Figure 1E). This pattern is
222 not observed for microsatellites which are known to localize outside of pericentromeric
223 heterochromatin (Figure S2A). We also analyzed publicly available *D. melanogaster* data,
224 including flies, imaginal discs, and salivary glands (which are the most extreme in polyteny), and
225 observed this same pattern of under-replication of satellite repeats in polytene tissues (Figure
226 S2B and S2C).
227
228 *Reads with tandem repeats had lower quality scores in Illumina data*
229
230 Upon inspection with FastQC of our data from the polyteny analysis, we found a bimodal
231 distribution of quality scores, with one peak at 22 and another at 37 (Figure S3A). After filtering
232 low quality reads, the majority of the reads with simple satellites were removed (Figure S3). The

233 amount of satellites were reduced by ~15 x after quality filtering (Figure 1E). It is apparent that
234 in our dataset, simple satellite-containing reads were highly enriched for low quality scores. We
235 examined other published *D. virilis* Illumina datasets to evaluate if this issue existed in other
236 sequencing runs. Two other datasets were available and the one that was produced on the
237 Illumina NextSeq platform like our data (Miller *et al.* 2018) showed the same pattern of biased
238 quality scores in repetitive reads (Figure S4). The dataset produced on the HiSeq platform (Wei
239 *et al.* 2018) did not show this pattern. It should be noted however that the amounts of 7mer
240 satellites sequenced in the NextSeq datasets were higher than the HiSeq dataset. Our libraries
241 were multiplexed with other non-*D. virilis* group samples from unrelated projects and only
242 represented ~20% of the total sequenced lane so that we would not have issues related to low
243 complexity. We also noticed this pattern (but less dramatically) in our Illumina sequencing of
244 multiple strains.
245
246 **Related species have similar but fewer simple repeats**
247
248 *D. novamexicana* and *americana* which are 0.38 MY diverged from each other, are sister
249 species of *D. virilis*, which is approximately 4.5 MY diverged (Caletka and McAllister 2004)
250 (Figure 2A). We sequenced these species with high coverage PacBio runs and characterized and
251 quantified satellites. We emphasize the comparison of relative satellite amounts since all are
252 likely under-represented. Interestingly, we found the most abundant satellite in *D. virilis*,
253 AAACTAC, is also the most abundant in *D. novamexicana*, albeit with about half the total
254 amount. The second and third most abundant repeats, AAACTAT and AAATTAC, however were

255 barely present in *D. novamexicana*. The second most abundant satellite in *D. novamexicana*
256 was AAACAAC, whereas in *D. virilis* there is only a small amount. *D. americana* has a similar
257 satellite composition as *D. novamexicana*, but because of issues with a different PacBio
258 sequencing chemistry used, we do not report the data here.

259 By looking at sequencing data in more diverged species, we can infer when the
260 AAACTAC satellite family arose. *D. montana*, another member of the virilis group that is 7-11
261 MY diverged from *D. virilis* (Ostrega and Thompson 1986; Spicer and Bell 2002) (Figure 2A),
262 does not have any AAACTAC family satellites, and in fact no enrichment of 7 bp satellites. The
263 most abundant satellite is AAAC, which may be related to the AAACTAC family. *D. hydei* is
264 approximately 26 MY diverged from *D. virilis* (Izumitani *et al.* 2016), and here 7 bp satellites are
265 again the most enriched, but the sequences are unrelated to those in *D. virilis* (ACCCATG,
266 AAAGGTC from PacBio data). From these data, we infer that the AAACTAC family of satellites
267 arose in the clade leading to the *D. virilis* phylad 4.5-11 MYA. We also analyzed Illumina
268 sequencing data for *D. lummei*, which is 3 MY diverged from *D. novamexicana/americana*.
269 AAACTAC is conserved in *D. lummei,* but it is the only enriched 7mer satellite in this species and
270 its estimated abundance is much lower than the other three *D. virilis* phylad species.
271

272 **Complex satellites are also abundant in *D. virilis* group genomes**

273 We searched the high-quality genome assemblies for complex satellites (defined here as unit
274 lengths greater than 20 bp), and we applied NCRF to search in the raw reads. In *D. virilis,* we
275 found a previously uncharacterized 36-bp satellite
276 AAAACGACATAACTCCGCGCGGAGATATGACGTTCC making up ~800 kb of our assembly. In *D.*

277 *novamexicana*, we found a previously uncharacterized 32 bp satellite

278 AAAAGCTGATTGCTATATGTGCAATAGCTGAC along with a related 29 bp satellite that spanned

279 over 1.1 Mb on a single contig. The 3 Mb contig was approximately half satellite and half non-

280 satellite. The non-satellite half had similarity to chromosome 6 (dot chromosome/Muller

281 element F) (Figure S5). In *D. americana*, we found this identical 32 bp satellite, but in total its

282 span was only ~150 kb. In all *D. virilis* group species, we also found a series of similar satellites

283 varying in size (150-500 bp) related to the previously described helitron central repeat that has

284 expanded to tandem repeats in the virilis group (Dias *et al.* 2015).

285

286 **Fluorescence in situ hybridization reveals evolutionary dynamics of 7mer repeats**

287

288 The location of the 7mer satellites on metaphase chromosomes has never been shown in the *D.*

289 *virilis* group. The AAACTAC satellite is conserved between all three species included in this

290 study: *D. virilis, D. novamexicana,* and *D. americana*, but the abundance varies by

291 approximately two-fold. The second most abundant satellites have turned over between *D.*

292 *virilis* and *novamexicana*/*americana*. We used FISH of the most abundant 7mers (AAACTAC,

293 AAACTAT, AAATTAC, AAACAAC) in these three sister species. *D. virilis* and *D. novamexicana*

294 have the same karyotype with five acrocentric chromosomes plus the very small F element or

295 "dot chromosome". The strain of *D. americana* we used has centromere-centromere fusions

296 between the X and 4th chromosomes and the 2nd and 3rd chromosomes.

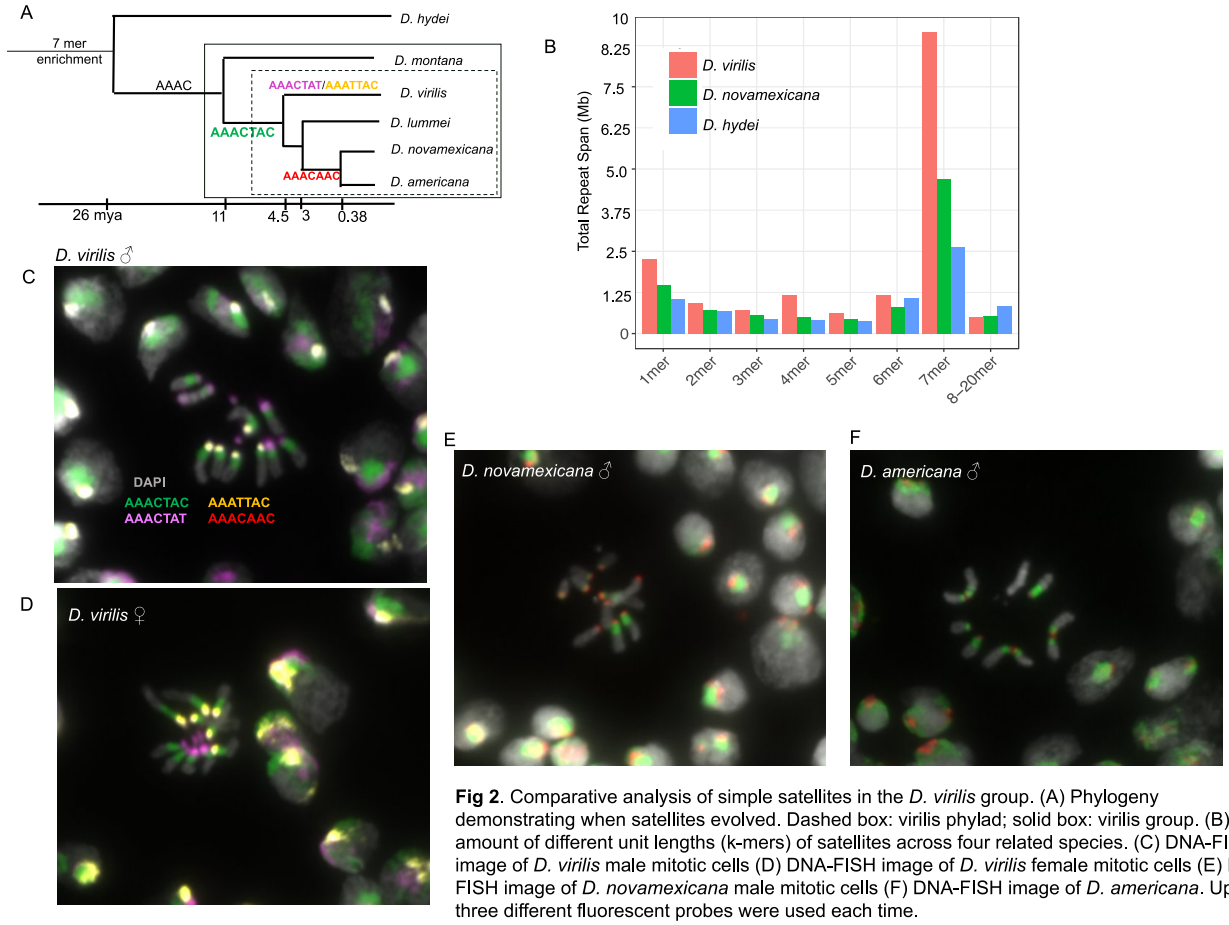


Fig 2. Comparative analysis of simple satellites in the *D. virilis* group. (A) Phylogeny demonstrating when satellites evolved. Dashed box: virilis phylad; solid box: virilis group. (B) Total amount of different unit lengths (k-mers) of satellites across four related species. (C) DNA-FISH image of *D. virilis* male mitotic cells (D) DNA-FISH image of *D. virilis* female mitotic cells (E) DNA-FISH image of *D. novamexicana* male mitotic cells (F) DNA-FISH image of *D. americana*. Up to three different fluorescent probes were used each time.

FISH results in *D. virilis* show that the most abundant satellite determined by sequencing, AAACTAC, is clearly the most abundant and occurs in approximately equal amounts in the pericentromeric region on the five pairs of large chromosomes. The Y chromosome appears to have slightly less AAACTAC satellite. The second and third most abundant satellites, AAATTAC and AAACTAT, are localized beyond the pericentromeric repeat at or near the centromere. There are five single chromosomes having each of these repeats, indicating that one chromosome pair has different satellite content - which we hypothesized to be the X and Y. Based on differences between male and female FISH results (Figure 2C and 2D), we suggest the Y chromosome has AAACTAT at both distal ends of the chromosome and AAACTAC only flanking one end, whereas the X chromosome has the other centromeric repeat AAATTAC. We were also

309 able to visualize the dot chromosomes in *D. virilis*, which we find is mostly composed of
310 AAACTAT. The AAACAAC satellite is present in small amounts in *D. virilis*, very likely on a single
311 chromosome (Figure S6).
312 We estimated that *D. novamexicana* has approximately half the AAACTAC as *D. virilis*,
313 and visualizing it with FISH reveals a pattern that suggests aspects of its evolution. Its
314 pericentromeric localization is conserved. One chromosome pair has the same amount of
315 AAACTAC as *D. virilis*, whereas all other chromosomes have a very small amount (Figure 2E).
316 Based on the FISH images, it appears that it is the 5th chromosome in *D. novamexicana* that has
317 the greatest amount of pericentromeric AAACTAC conserved. The centromeric repeat on all
318 major chromosomes is AAACAAC in *D. novamexicana* and *D. americana.* Our images illustrate
319 clearly the centromere-centromere fusion between chromosome X-4 and 2-3 in *D. americana*
320 with the satellites being maintained on both sides of the fusion (Figure 2F). None of the four
321 simple satellite probes bound to what we hypothesize is the Y chromosome of *D. novamexicana*
322 or *D. americana*. Based on the images we suggest that *D. americana* has an intermediate
323 amount of pericentromeric AAACTAC satellite compared to *D. virilis* and *novamexicana*.
324
325 **Some satellite-containing reads are linked to TEs**
326
327 We used RepeatMasker to detect if any of the reads containing satellites also contain
328 transposable elements. TEs might be located in islands within the simple repeats at the
329 centromere as in *D. melanogaster* (Chang *et al.* 2019) (Figure 3A), in more distal regions
330 flanking the pericentromeric heterochromatin (Figure 3B), or some TEs may have inserted into

331 long pericentromeric satellite arrays (Figure 3C). For AAACTAC (and its artifactual counterpart
332 AAACGAC), ~3.5% of reads (2473/75,364) also contained at least 500 bp of a TE insertion. In the
333 reads that contain TE sequences, satellite repeats were enriched at the beginning and ends of
334 reads, concordant with the hypothesis that a high proportion of the TEs we found are actually
335 flanking the long arrays of AAACTAC distally (Figure 3B). In order to understand how many
336 reads would be expected to contain both satellites and TEs if TEs flanked this satellite and were
337 not interspersed, we simulated a situation where a large satellite block flanked a large TE block.
338 This simulation revealed a much smaller amount of reads containing both satellites and TEs
339 (0.06 %). This result suggests that not only are TEs flanking the pericentromeric satellite
340 AAACTAC, but there have likely been TE insertions into the satellite arrays. Likely flanking the
341 proximal end of the pericentromeric satellite are the centromeric satellites AAACTAT or
342 AAATTAC. 144 reads contained both AAACTAC and AAACTAT repeats (0.19% of AAACTAC reads)
343 and 94 reads contained both AAACTAC and AAATTAC repeats (0.12% of AAACTAC reads). Based
344 on our simulations, these proportions of overlapping reads are consistent with our expectation
345 based on our FISH results that the pericentromeric and centromeric satellites have relatively
346 clean boundaries and they directly flank each other.
347

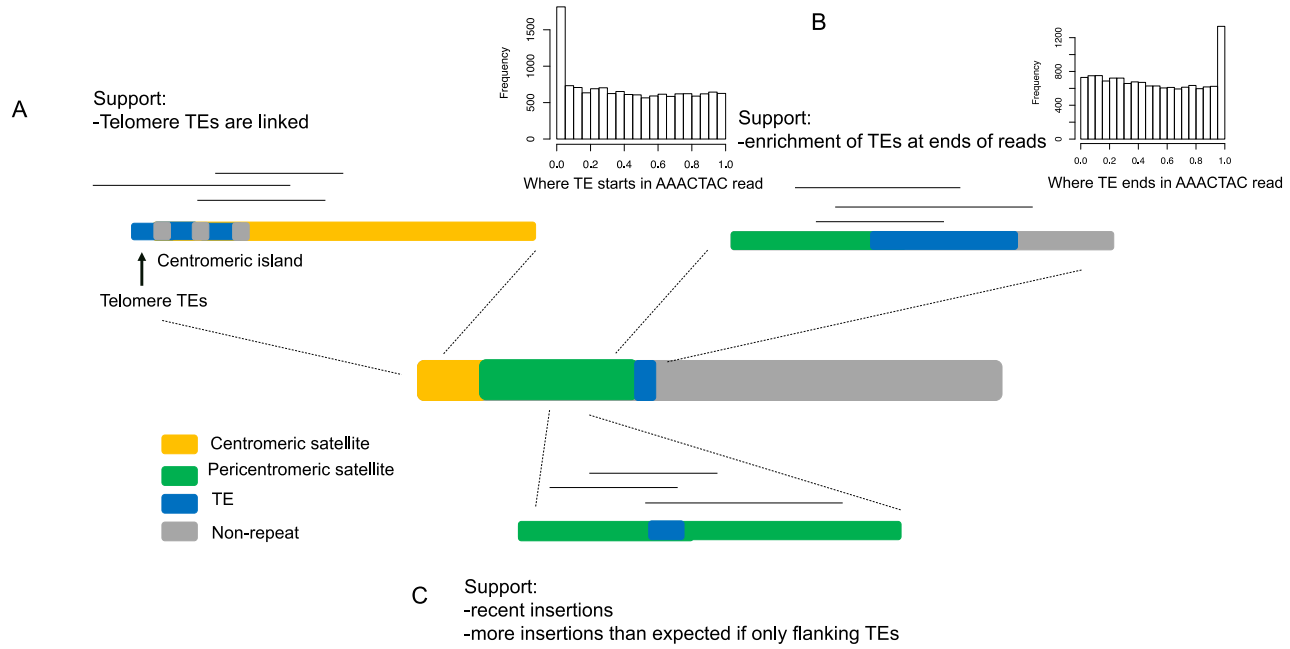


Fig 3. Transposable elements are in close proximity to satellite arrays. (**A**) Satellites near the centromere may be linked to potential islands of retroelements and telomeric TEs. (**B**) Heterochromatic TEs flank satellite arrays. Histograms show the start and end position of transposable element in the satellite-rich read. (**C**) Transposable elements may have inserted into the satellite arrays.

Many of the TE insertions into satellites seemed to be very young. 2080/2473 TE-containing AAACTAC reads had a TE insertion with less than 15% divergence from the Repbase consensus, which is the expected error rate of PacBio reads. These insertions included the superfamilies: DNA elements, LINE/CR1, LINE/I-Jockey, LINE/Penelope, LINE/R1, LTR/Copia, LTR/Gypsy, LTR/Pao, and Helitrons. We acknowledge however, that there may be a detection bias for insertions that are less divergent from the consensus. We also remind readers that there were likely fewer satellite reads sequenced than expected, and that this may have biased these results if satellite-only regions were sequenced less efficiently than satellite-TE regions. For centromeric satellites AAATTAC and AAACTAT, the results are more difficult to interpret since these were more strongly under-represented. However, 300/715 reads of AAACTAT and 49/385 reads AAATTAC contained TEs. Like the AAACTAC pericentromeric satellite, most TE

361 insertions were low divergence from the Repbase consensus in the centromeric satellite reads.
362 288/300 and 46/49 TE containing reads had a TE insertion > 500 bp with < 15% divergence for
363 AAACTAT and AAATTAC, respectively.
364 There were differences in the TE composition of reads with different satellites. For the
365 pericentromeric satellite AAACTAC, Gypsy-10_Dvi was the most enriched, followed by Helitrons
366 (Helitron-1N1_DVir, Helitron-1_DVir, Helitron-2N1_DVir, Helitron-2_DVir). For the AAACTAT
367 centromeric satellite, Gypsy-10_Dvi was again the most enriched, followed by Penelope. For
368 the AAATTAC centromeric satellite, Gypsy-2_DVir was the most enriched followed by Penelope.
369 In both AAACTAC and AAACTAT reads, CR1-1_DVi was the second or third most abundant TE.
370 Interestingly, R1 was present in relatively high amounts in AAACTAC. In 110/132 of these R1-
371 AAACTAC reads, rDNA sequences were not also linked. This suggests that some R1 elements,
372 which are generally localized to rDNA loci, have jumped into or near satellite arrays. This is
373 concordant with findings that some R1 elements are located outside rDNA loci in Drosophila
374 (Stage and Eikbush 2009). All centromeres in *D. virilis* are acrocentric, meaning that the
375 telomeric TEs Het-A and TART (Casacuberta and Pardue 2003) are likely near the centromere
376 satellites. We found 12 reads linked to AAATTAC that contained matches to TART. Only two
377 reads linked to any satellite contained a sequence matching HeT-A. We also used BLAST to
378 detect matches between the genome assembly (masked from the 7mer satellites) and the 7mer
379 satellite reads. We could not detect any unique regions of the genome that matched non-
380 satellite sequence on the reads. Non-satellite parts of satellite reads matched hundreds of
381 places in the genome each.
382

383 **Variation in *D. virilis* group global strains**

384

385 *D. virilis* is globally distributed while its sister species are localized to North America, with *D.*
386 *novamexicana* more restricted than *D. americana.* Patterns of variation in satellites may reveal
387 potential mechanisms that can be hypothesized to be driving satellite evolution. Additionally, *D.*
388 *americana* has a polymorphic fusion between the X and 4th chromosomes, so we can identify
389 differences in satellite composition associated with the fusion. This fusion has been shown to
390 be currently undergoing meiotic drive, potentially mediated by a larger total centromere or
391 pericentromere size in the fused strains compared to the non-fused strains (Stewart et al.). On
392 the other hand, chromosome fusions are often caused by Robertsonian translocations with loss
393 of some non-essential DNA, which might include pericentromeric satellites (Schubert and Lysak
394 2011).

395 We used Illumina sequencing with PCR-free library preparation and k-Seek to estimate
396 the abundance of 7mer satellites across 12 worldwide strains of *D. virilis*, eight strains of *D.*
397 *americana* (including four strains that have the X-4 fusion and four that do not), and five strains
398 of *D. novamexicana* (Table S1). All sequenced strains were male except a female of the *D. virilis*
399 inbred strain 87 as a comparison. A PCA using only the four most abundant 7mers shows
400 clustering of the three species, but the separation is much more dramatic in the PCA using the
401 20 most abundant simple satellites (Figure S7). Overall, *D. virilis* had the highest AAACTAC
402 satellite content as well as the highest variation, with *D. americana* intermediate between *D.*
403 *virilis* and *D. novamexicana* (Figure 4A). Using different normalization procedures including
404 mapping and GC correction (see Materials and Methods), produced the same relative ranking

405 of satellite abundances between species. In all cases, the inbred strain from which the genome
406 sequence was produced had the lowest abundance of AAACTAC. In the case of *D. virilis*, this
407 difference was very high. This was not due to a normalization bias as we did mapping-free
408 normalization.

409 Satellite abundances in *D. virilis* displayed a pattern that appeared to be correlated to
410 the geographic location strains were collected from. For the centromeric satellite AAATTAC,
411 there was a linear decrease in abundance from West to East then South following migration
412 from Beringia (Throckmorton 1982) beginning in China (Figure 4C). For the centromeric satellite
413 AAACTAT, the pattern was the opposite; a linear increase in abundance from West to East then
414 South (Figure 4D). We analyzed variation in the satellite repeats to determine if the apparent
415 linear expansion in AAATTAC also correlated with newer and hence more homogeneous
416 sequences. On average, the centromeric satellite arrays were very homogeneous (average
417 above 99% sequence identity in Illumina reads). However, AAACTAT had a slightly higher
418 sequence identity than AAATTAC (Figure 4 C and D). There was no pattern in average sequence
419 identity with respect to geography for any of the satellites (Figure 4 C and D). The
420 pericentromeric satellite AAACTAC has almost identical sequence divergence across the three
421 species (~98.5%). When comparing between a male and female of the same strain, the male
422 had lower sequence identity. In *D. americana* and *D. novamexicana*, the AAACAAC satellite had
423 lower average sequence identity at 97.5%.
424

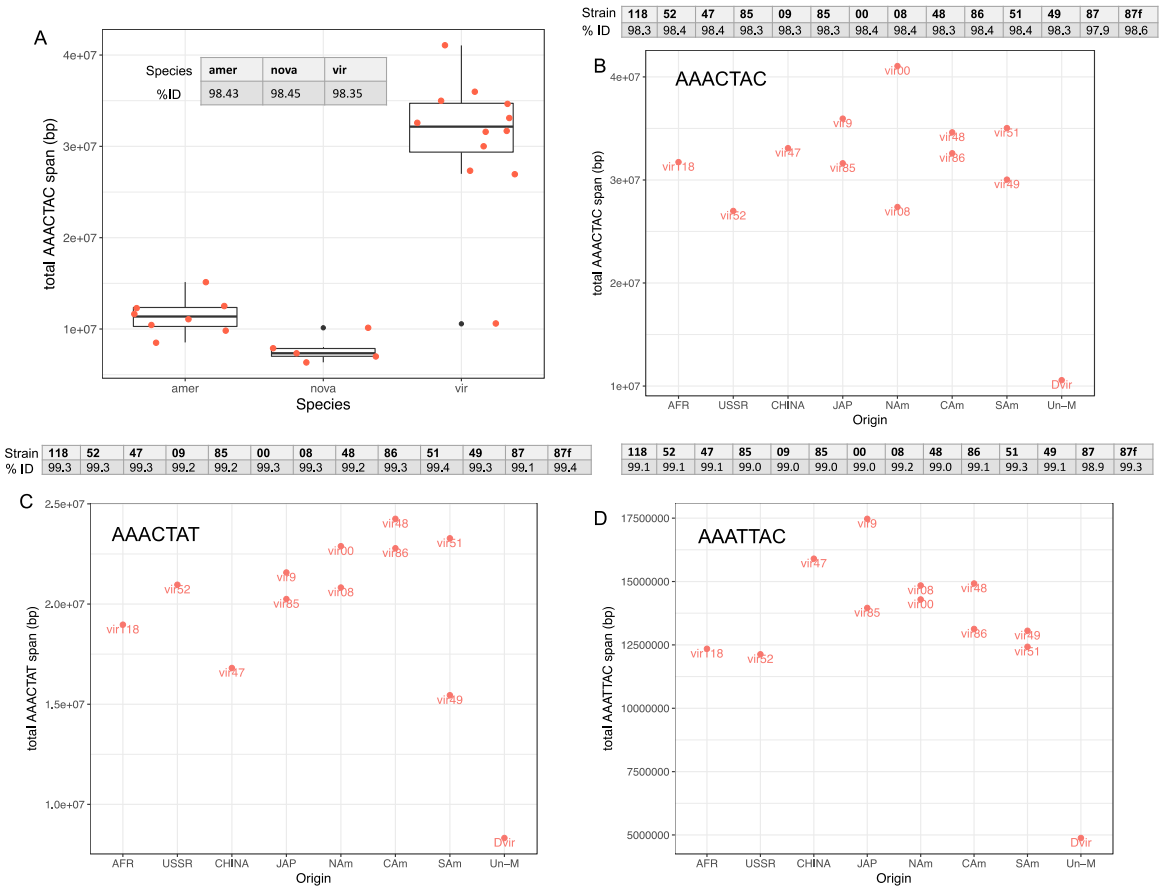


Fig 4. Variation in satellites across species and strains. (**A**) AAACTAC abundance across the three species. (**B**) AAACTAC abundance across strains originating from different localities. (**C**) AAACTAC (centromere satellite) abundance across strains originating from different localities. (**D**) AAATTAC abundance across strains originating from different localities.

There were also several other satellites besides the most abundant four that varied between the *D. virilis* group species. AAACAAT was either absent or in low abundance in *D. virilis* and *novamexicana*, but present between 100,000 - 250,000 copies in *D. americana,* indicating a very recent expansion of this satellite. AAAAAC was present in ~100,000 copies in *D. americana* and *D. novamexicana*, whereas it is almost absent in *D. virilis*. To differentiate Y-specific satellites, we included a female of strain 87, the reference genome strain, in our Illumina sequencing run in addition to a male. There were several Y-specific satellites in *D. virilis* found by Wei et al. 2018 and validated by our data, which varied between species in the group

435 (AATAATAG, AATAGATT, and ACATAT). All had different relative patterns of abundance
436 between species
437 (https://github.com/jmf422/D_virilis_satellites/blob/master/Intra_inter_species_sequencing/vi
438 rilis_group_intra-inter_species.html). There was no detectable difference in centromeric or
439 pericentromeric satellite abundance in *D. americana* strains with and without the polymorphic
440 centromere-centromere fusion that might be explained by a change in satellite abundance
441 resulting from molecular events surrounding the fusion (Figure S8).
442
443 Amplified repeats of the DINE-1 helitron have been found on the *D. virilis* 5th and Y
444 chromosomes (Dias *et al.* 2015). We also examined variation in this satellite abundance by
445 counting reads that mapped to this family of satellites. There was no striking pattern with
446 respect to geography, however the strain with the highest AAACTAC content had the lowest
447 DINE-1 content, besides the inbred strain 87, which had lower satellite content all-around
448 (Figure S9). We expect AAACTAC and DINE-1 to be both in the pericentromeric region of Chr5 in
449 *D. virilis* based on our FISH results and the previous work. The male contained 2.6x more DINEs
450 than the female of the same strain, indicating that ~70% of the DINE-1 repeats in the genome
451 are located on the Y chromosome in *D. virilis*.
452
453 **Discussion**
454
455 Here, we used the satellite DNA-rich genome of *D. virilis* to highlight three previously
456 uncharacterized mechanisms for biases that occur in sequencing and analyzing satellite DNA.

457 We emphasize that comparing satellite DNA amounts between different platforms (e.g.
458 Illumina, PacBio, Nanopore, and even different versions of each) should be done with caution
459 as each technology has its own biases. We have found that issues arise when multiple kilobase
460 molecules of simple satellite DNA are attempted to be sequenced by long-read platforms. In
461 the case of PacBio, systematic errors in base calling may be introduced when sequencing
462 through long arrays of satellites. This issue is not specific to our satellites, as a recent study has
463 also found systematic errors and strand biases in shorter arrays of human satellites in both
464 PacBio and Nanopore reads (Mitsuhashi *et al.*). Circular consensus sequencing (CCS) or "HiFi", a
465 type of sequencing offered by PacBio which allows an accurate consensus to be produced after
466 multiple rounds of sequencing the same molecule, may be more appropriate for sequencing
467 analysis of satellite DNA. Preliminary results based on collaborative efforts with PacBio
468 representatives show no systematic errors in satellite sequences with the new CCS platform. In
469 the case of Nanopore, it is possible that a similar satellite-specific base calling errors exists, or
470 that there is a strand-specific difference in secondary or tertiary structures that occur in long
471 strands of simple satellite DNA. We caution readers in interpreting simple satellite DNA from
472 long-read sequencing data and suggest validation with satellites of known sequence and
473 abundance if available or Illumina reads (without quality filtering). Long read platforms are
474 already improving their chemistry and software for better satellite characterization. Because
475 long reads are likely to cross boundaries of different repetitive regions, long read sequencing
476 proved useful in understanding the length of the satellite arrays and TE insertions into them.
477 Moreover, we demonstrate that the abundance of satellites in pericentromeric
478 heterochromatin are underestimated when sequencing flies compared to pure diploid tissue

479 because of polyteny. We caution readers in performing quality filtering before simple satellite
480 analysis, as satellite containing reads may be enriched for lower quality scores.
481 From comparative analysis of satellites, we found the abundant AAACTAC family of
482 satellites arose in the branch leading to the *virilis* phylad 4.5-11 MYA (Figure 2A). Interestingly,
483 the most abundant satellite in *D. montana*, 7-11 MY diverged, is AAAC. The AAACTAC and AAAC
484 satellites may have been derived from one another. However, an enrichment of different 7 bp
485 satellites does exist in the even more distantly related *D. hydei.* From both FISH and sequencing
486 analysis, we found that *D. virilis* has the highest total amount of AAACTAC family satellites, *D.*
487 *novamexicana* has the lowest, and *D. americana* intermediate. Unlike the pericentromeric
488 satellite, the centromere-proximal satellite sequence has turned over between *D. virilis* and *D.*
489 *americana/novamexicana*. The AAACAAC satellite is present in *D. virilis* in relatively low
490 amounts, likely on a single chromosome. This satellite then became the centromeric satellite in
491 *D. americana* and *novamexicana*, replacing the AAACTAT and AAATTAC satellites. Interestingly,
492 *D. lummei* has lower satellite content than the other species and its only high-abundance
493 simple satellite is AAACTAC. Although sequencing quantified only up to 30 Mb of the AAACTAC
494 family of satellites, FISH confirmed that these satellites are extremely abundant in *D. virilis* and
495 the 40% of the genome estimate seems realistic.
496
497 We can make hypotheses about how and why the satellites expanded in *D. virilis*. We
498 know that mutation rates for changes in copy number of satellite DNA are high, and potentially
499 have a tendency to expand rather than contract in the absence of selection (Flynn *et al.* 2017,
500 2018). High rates of mutation must be accompanied by a regime that would allow a satellite

501 copy number increase to sweep the population - which could be mediated by positive selection
502 if there is a benefit of the satellite increase, or centromere drive if the phenomenon is at play in
503 this group. Alternatively, in a situation where satellites are slightly deleterious, small effective
504 population sizes in isolated populations or continued bottlenecks could allow satellites to
505 expand in the genome without being removed by selection. However, *D. novamexicana* has the
506 lowest effective population size of the virilis pylad and yet it has the lowest amount of satellite
507 DNA. We already know that the centromere-to-centromere fusions in *D. americana* have
508 undergone meiotic drive hypothesized to be mediated by the increase in centromere total size
509 with the fusion (Stewart et al.). The mechanism allowing drive in *D. americana* may have been
510 at play in the branch leading to *D. virilis* or may be currently occurring. Why have satellites not
511 expanded to this extent in the other species? *D. virilis* might have some attributes about its
512 biology that made the satellite expansion favorable or allowable. For example, genome size is
513 positively correlated with development time in Drosophilidae (Gregory and Johnston 2008). *D.*
514 *virilis* has a slow development time, and this may have evolved in concert with the expansion in
515 satellite abundance in its genome.
516 We can also use data from multiple strains to make hypotheses about factors driving
517 satellite DNA evolution in *D. virilis*. Ancestrally, *D. virilis* had a relatively small effective
518 population size in an isolated range in Asia, and has undergone a recent population and range
519 expansion (Mirol et al. 2008). The amount of the most abundant pericentromeric satellite,
520 AAACTAC, does not vary greatly across strains from different localities. Assuming we sequenced
521 a strain from the ancestral range, this suggests that population bottlenecks did not allow
522 AAACTAC to expand, and the satellite expansion likely occurred before the population

523 expansion. However, the abundance of the centromere-proximal satellites does show a pattern
524 with respect to where samples were collected. There is increasing AAACTAT abundance on a
525 gradient from west to east, correlated with decreasing AAATTAC on the same gradient.
526 AAACTAT may be experiencing an expansion or starting to occupy additional centromeres,
527 while the AAATTAC satellite may be decreasing in parallel because of a trade-off or maximum
528 limit on satellite amount. AAACTAT is located on the Y, two other chromosomes and the dot,
529 and AAATTAC is located on the X and two other chromosomes. AAACTAT has a slightly higher
530 average sequence identity than AAATTAC, concordant with the hypothesis that it has recently
531 increased its copy number.

532 Our observation of rapid evolution and enrichment of AAACTAC in *D. virilis* in a short
533 evolutionary time period (a few million years) is consistent with the centromere-drive model to
534 account for the evolution of centromere complexity in genetic conflict (Malik and Bayes, 2006).
535 In this model, the asymmetric female meiosis can cause competition between the centromeres
536 with or without newly formed satellites (or with more or less satellites) to be included into the
537 oocyte to pass to next generation. A consequence of the competition would be runway
538 expansions of centromeric satellites, and rapid replacements by novel satellites. According to
539 this model, our observation of increasing AAACTAT abundance in a west-east gradient may
540 reflect an ongoing centromere drive combined with selection “pushing back”.

541 Interestingly, in *D. novamexicana*, AAACTAC was greatly reduced in the pericentromeric
542 regions on all chromosome pairs except one. Based on the FISH images in *D. novamexicana* and
543 *americana*, we hypothesize that it is the 5th chromosome that has the high amount of
544 AAACTAC satellite. This is interesting because previous work has shown that the 5th

545 chromosome contains a high amount of DINE-1 helitron satellite in *D. virilis* but not in *D.*
546 *americana* (Dias et al. 2015). We found that all chromosomes including Chr5 contain a large
547 amount of AAACTAC in *D. virilis*. The DINE-1 repeat and the pericentromeric satellite may both
548 be present (and possibly in conflict or competing) on the *D. virilis* 5th chromosome, but in the
549 branch leading to *D. novamexicana*, AAACTAC took over. DINE-1 had a relatively consistent
550 amount across different *D. virilis* strains, however the strain with the highest AAACTAC amount
551 is an outlier with a lower DINE-1 amount (Figure S9). This may indicate a maximum threshold of
552 satellites was reached on this chromosome, and one satellite had to reduce its abundance. We
553 have seen evidence for this trade-off, or appearance of competitive exclusion, being invoked
554 under selection in our previous studies (Flynn *et al.* 2017, 2018). There may have been a similar
555 conflict on Chr5 of *D. novamexicana*, where AAACTAC retained a high copy number to prevent
556 DINE-1 from expanding. Interestingly, the opposite has occurred on the *D. novamexicana* and
557 *D. americana* Y chromosome, where AAACTAC family satellites are absent but DINE repeats are
558 abundant. A potential mechanism mediating apparent stabilizing selection on total satellite
559 abundance is that satellites can act as a sink for heterochromatin factors, with their abundance
560 affecting chromatin state (Lemos *et al*. 2010).

561 The AAACTAC satellite has remained conserved in sequence and location in the virilis
562 phylad. It has also maintained high levels of sequence identity that is equal in the three species
563 we sequenced (98.5% based on Illumina reads). The conservation may reflect a constraint due
564 to selection or a pervasive mechanism of concerted evolution. The periodicity of the sequence
565 may stabilize the DNA helix wrapping around nucleosomes, or it may be constrained by
566 coevolution of an important satellite DNA binding protein (Maio *et al.* 1977; Jagannathan *et al.*

567 2018). Additionally, within the AAACTAC family, the position and identity of the four A-
568 nucleotides are conserved in all four satellites (AAACTAC, AAATTAC, AAACTAT, AAACAAC) -
569 which may indicate constraint based on the above mechanisms. Conservation of particular
570 satellite unit lengths and "AA" periodicities have been found in other divergent species
571 (Lowman and Bina 1990). Concerted evolution of satellites could be achieved by repeated
572 recycling of units by copy number changes associated with replication slippage or unequal
573 recombination or gene conversion (Walsh 1987; Elder and Turner 1995). However,
574 recombination in the pericentromeric heterochromatin has never been detected in wild-type
575 flies (Mehrotra and McKim 2006; Hughes *et al.* 2018). On the other hand, if recombination were
576 occurring, satellite arrays will eventually be lost unless they are conserved by selection
577 (Charlesworth *et al.* 1986). Clearly, we are still lacking in understanding how and why long
578 simple satellite arrays maintain their homogeneity, and whether recombination plays a role in
579 their dynamics. Concordant with the hypothesis that recombination is playing a role, males
580 have lower average sequence identity in the 7mer satellites than females, which could indicate
581 increased decay on the Y chromosome where there no homologous recombination (Figure 4A).
582 Moreover, our results suggest that transposable elements flank the AAACTAC satellite
583 array (likely more distally to the centromere) and some TEs have inserted within the array. Our
584 analysis suggests that most TE insertions are recent, but because of the under-representation
585 of satellite containing reads, we cannot estimate the number of TE insertions that have
586 occurred into the satellite arrays. Our analysis is also not precise enough to determine if there
587 are islands of TEs at the centromere in *D. virilis* as has been demonstrated in *D. melanogaster*
588 (Chang et al. 2019).

589 We found no difference in centromeric and pericentromeric satellites abundances
590 between *D. americana* strains that differ in their X-4 fusion status. This suggests that the fusion
591 event did not result in a large loss of satellites, making the total centromere and
592 pericentromere size is indeed larger on the X-4 fused chromosome than the single unfused
593 chromosome, concordant with the hypothesis that a larger centromeric region results in
594 centromere drive (Stewart et al.). Another interesting observation from the sequencing of
595 multiple strains of the three species was that in all cases, the inbred strain that the reference
596 genome was made from had the lowest amount of AAACTAC. For *D. virilis*, this difference was
597 extreme. It is tempting to speculate that the process of inbreeding and/or long periods in the
598 lab may have driven the reduction in pericentromeric satellite abundance.
599 In conclusion, our results show very rapid dynamics in the abundant satellites of the *D.*
600 *virilis* group that are likely explained by various cellular and population-level forces that are not
601 yet understood. Further studies can test if there is a species-specific upper limit to satellite
602 amount per genome or per chromosome upon which negative fitness effects occur, which may
603 result in trade-offs or competition between satellites. Centromere drive may be an important
604 process affecting satellite evolution in this species group, and might partially explain why the
605 satellites expanded 4.5-11 MYA, why satellite sequences at the centromere turned over more
606 rapidly, and why there is a gradient of increasing satellite content related to geographical
607 distribution of strains. A more extensive study to determine if inbreeding or extended periods
608 in the lab drives a reduction in satellite abundance will help illuminate the processes that are
609 important for maintaining satellite content. Determining the frequency of recombination in the
610 large pericentromeric heterochromatin blocks in species like *D. virilis* will be challenging but

important for understanding how the satellites maintain homogeneity in their sequence. To understand the role of satellites and the importance of their sequence, unit length, and abundance, researchers can strive to develop methods to engineer satellites by modifying specific bases and their abundances.

MATERIALS AND METHODS

All scripts for analyzing the data and to produce the results we show are here: https://github.com/jmf422/D_virilis_satellites. Illumina sequencing reads generated for this study are deposited in NCBI SRA under accession PRJNA548201. Raw PacBio reads will be deposited under the same accession. Both will be released upon publication.

Characterizing satellite DNA from genome assemblies

All scripts and R markdown files used for this analysis are provided in https://github.com/jmf422/D_virilis_satellites/tree/master/Genome_assembly_analysis. We used genome assemblies produced by the PacBio sequencing project (https://www.ncbi.nlm.nih.gov/bioproject/?term=txid7214[Organism:noexp]) of *D. virilis*, *D. novamexicana*, and *D. americana*. We also downloaded the *D. virilis* genome produced by Nanopore sequencing from (Miller *et al.* 2018), and the CAF1 assembly from (Drosophila 12 Genomes Consortium *et al.* 2007). We used Phobos (https://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm) and Tandem repeats finder (Benson 1999) to

633 characterize simple and complex satellites in these genome assemblies. To identify the
634 chromosomal linkage of complex satellites in the genome assembly, we produced a dotplot
635 with D-GENIES (Cabanettes and Klopp 2018).
636
637 **Characterizing satellite DNA from raw long reads**
638
639 Characterizing and quantifying satellites from long reads is a challenge because of the
640 sequencing high error rate. We used two approaches to characterize satellites from raw long
641 reads. The first approach, we call k-Seek + Phobos, in which we first broke the reads into 100 bp
642 subreads and ran k-Seek on them. k-Seek is very efficient for analyzing many reads, however is
643 not very sensitive for reads with a high error rate since it was designed for Illumina reads (Wei
644 *et al.* 2014). If k-Seek found satellites on at least one subread, we would run the complete
645 parent read through Phobos. Phobos is more sensitive to imperfect repeats and error rates, but
646 cannot handle huge quantities of data; thus why we only ran the portion of reads identified by
647 k-Seek to have tandem repeats. This approach allowed us to characterize satellites *de novo* and
648 quantify them. All scripts for the analysis of long reads with the k-Seek + Phobos approach are
649 located here:
650 https://github.com/jmf422/D_virilis_satellites/tree/master/LongRead_kseek_Phobos. The
651 second approach we used is Noise-Cancelling Repeat Finder (NCRF, (Harris *et al.*)). This
652 program was designed to quantify satellites from long reads with high error rates. However, it
653 cannot identify satellites *de novo* and requires specific satellite sequences to search for. NCRF
654 also requires a “max divergence allowed” parameter, which we tuned with simulations (see

655 below). Scripts used for the NCRF approach are located here:

656 https://github.com/jmf422/D_virilis_satellites/tree/master/LongRead_NCRF.

657

658 We did simulations to assess both approaches:

659 https://github.com/jmf422/D_virilis_satellites/tree/master/Simulations. First, we created a

660 simplified mock *D. virilis* genome with a satellite DNA composition estimated from our FISH

661 results. We could not use the genome assembly because it contained very little satellite DNA.

662 Specifically, each chromosome had a centromeric satellite either AAATTAC or AAACTAC

663 followed by the pericentromeric satellite AAACTAC, combined taking up 40% of the genome.

664 The non-satellite DNA portion of the genome was generated randomly with a 40% GC content.

665 We then used PBSim (Ono *et al.* 2013) to simulate PacBio reads and we used these simulated

666 reads for multiple analyses. First, we used them to tune the max divergence parameter of NCRF

667 by running NCRF repeatedly with max divergence parameters ranging 18-30%. We found that

668 the amount of satellites found, particularly the most abundant one, levelled off at 25% max

669 divergence. This is the parameter value we used moving forward. We also used these simulated

670 reads to quantify satellites with both approaches and compare them. Finally, we used these

671 simulated reads to assess strand biases in long read sequencing data (see below).

672

673 **Identification of biases in simple satellites in long read data**

674

675 We suspected that there were biases in the satellite DNA found in the *D. virilis* group PacBio

676 (and Nanopore) data because we found high abundance satellites that had never been found

677 before with other types of data, and so we suspected they were artifactual. These artifactual
678 satellites were found with both kSeek + Phobos and NCRF approaches, but were not found in
679 the simulated data. We tried testing for a strand bias in reads that contained satellite DNA.
680 Using both the summarized output from NCRF and validated with a custom script
681 (LongRead_NCRF/which_strand_pacbio_script.sh), we counted the satellite DNA stretches that
682 originated from each the positive and negative strand. The positive strand is defined as the one
683 that contains the satellite AAACTAC and derivatives (more As than Ts), and the negative strand
684 is the one that contains the reverse complement (e.g. GTAGTTT, more Ts than As). We did this
685 for the three satellites used in the simulated data and real and artefactual satellites found in
686 the PacBio and Nanopore data. Detailed analysis and visualization of the biases is shown here:
687 LongRead_kseek_Phobos/longread_analysis.html
688
689 **Sequencing of polytene and non-polytene tissue**
690
691 To acquire *D. virilis* pure diploid tissue, we dissected male 3rd instar larvae and collected
692 imaginal discs including the eye-antennal disc and wing discs. Approximately 100 larvae were
693 required to get enough DNA (>1 ug). We also collected ~5 adult flies for fly libraries. We used
694 the inbred genome assembly strain 87 for these libraries. DNA was extracted with Qiagen
695 DNeasy blood and tissue kit and PCR-free libraries were prepared. Libraries were run on an
696 Illumina NextSeq with 1 x 150 bp reads, and each sample took up approximately 7% of the
697 flowcell. The other libraries run on this flowcell were from an unrelated project including
698 RNAseq from other species. Reads were analyzed with k-Seek both before and after filtering

699 with Trimmomatic (Bolger et al. 2014). FastQC was run to evaluate the quality of the reads.
700 Scripts are here: https://github.com/jmf422/D_virilis_satellites/tree/master/Polyteny. We also
701 analyzed publicly available *D. melanogaster* data from the same strain and same sequencing
702 platform of embryos (non-polytene), salivary glands (extreme polyteny) from (Yarosh and
703 Spradling 2014), and flies (varied levels of polyteny) from (Gutzwiller *et al.* 2015).
704
705 **Fluorescence in situ hybridization of satellite DNAs**
706
707 We followed the protocol of (Larracuente and Ferree 2015) for satellite DNA FISH. We ordered
708 the following probes from IDT with 5' modifications: $(AAACTAC)_6$ with alexa-488 fluorophore,
709 $(AAACTAT)_6$ with Cyanine5 fluorophore, $(AAATTAC)_6$ with Cyanine3 fluorophore, $(AAACAAC)_6$
710 with Cyanine3 fluorophore, and $(AAACGAC)_6$ with Cyanine5 fluorophore. We hybridized three
711 probes at a time, to allow for similar probes to compete to result in specific hybridization with
712 the rationale shown in (Beliveau *et al.* 2015). Hybridization temperature was 32°C. We imaged
713 on an Olympus fluorescent microscope and Metamorph capture system at the Cornell Imaging
714 Facility. Composite images were produced with ImageJ.
715
716 **Characterizing TEs linked to satellites and satellites anchored to the genome assembly**
717
718 We extracted the reads identified to have the 7 bp satellites on them from NCRF results, and
719 then we ran RepeatMasker (http://www.repeatmasker.org) on these reads using parameters:
720 "-nolow" and "-species Drosophila". All reads had at least 500 bp of tandem satellite on them

721 according to NCRF default parameters, and to avoid spurious identification of TEs from semi-
722 repetitive fragments, we described only TEs in reads that also had at least 500 bp of a TE
723 identified from RepeatMasker. We also BLASTed the same satellite reads to the genome
724 assembly to evaluate if satellite reads could be anchored to the genome assembly. Analysis
725 scripts are here: https://github.com/jmf422/D_virilis_satellites/tree/master/TEs_satellites.
726
727 **Sequencing of multiple *D. virilis* group strains**
728
729 We obtained as many strains of *D. virilis* that have information about where they were
730 collected as possible. This included 12 strains as live stocks we obtained either from stocks in
731 our lab or from the Drosophila species stock center (Table S1). We also prepared a female
732 library for strain 87 for when we wanted to differentiate Y-specific satellite patterns. We also
733 obtained five strains of *D. novamexicana* and eight strains of *D. americana*. All were obtained
734 from live stocks and the inbred genome strains were included for both species as well (strain 14
735 and G96, respectfully). For *D. americana*, we included four strains that have the chromosome X-
736 4 fusion and four strains that do not have it, based on communication with the Bryant
737 McAllister lab. DNA was extracted as above from five flies each and samples were prepared
738 identically as above and sequenced on 50% of 3 flowcells of Illumina NextSeq 1 x 150 bp reads.
739 We dispersed the samples from each species between multiple flowcells. Our samples took up
740 only half the flowcell with the other half being occupied by a RNAseq libraries from an
741 unrelated project.

742 All scripts used to analyze these data are located here:

743 https://github.com/jmf422/D_virilis_satellites/tree/master/Intra_inter_species_sequencing.

744 Reads were evaluated with FastQC and not filtered for quality based on the potential bias of

745 Illumina quality scores on satellites. We used k-Seek to quantify satellites. We tried several

746 normalization strategies but decided the most appropriate was a mapping-free normalization.

747 We estimated average depth by dividing the total number of bases sequenced by the estimated

748 genome size by flow cytometry (Bosco *et al.* 2007). We believe this was the best option in this

749 case because: 1) we were concerned about a mapping bias because for each species the strain

750 that the genome assembly was made from may have more reads map to it; 2) after masking the

751 genome from the 7mer satellites and also discluding the X and Y contigs (because we had male

752 and female strains, and the Y chromosome contained more low GC regions) - there was little

753 difference in coverage based on GC content. We include results when we used a mapping based

754 GC normalization in the sub-directory "AlternativeNormalization".

755 We used NCRF with modified parameters (minlength=100, maxdiv=10) to characterize

756 the average sequence identity of satellite arrays from the Illumina data. To quantify DINE-1

757 satellites across *D. virilis* strains, we produced a library of DINE-1 satellite variants based on our

758 PacBio genome analysis. We then mapped Illumina reads to this library and normalized the

759 number of reads that mapped to any sequence in the library by the estimated depth. We also

760 analyzed Illumina DNA sequencing reads of *D. montana* (Parker et al. 2018) and *D. lummei*

761 (Ahmed-Braimah *et al.* 2017) with k-Seek to identify the most abundant satellites and whether

762 or not the AAACTAC satellite family was present.

763

764 **Acknowledgments**

765

766 We thank Yasir Ahmed-Braimah for helpful discussions and advice for some analyses. We also
767 thank Bryant McAllister for providing *D. americana* strains along with their fusion status. We
768 are grateful for Elissa Cosgrove's help with some computational trouble-shooting and Asha
769 Jain's help in preparing DNA sequencing libraries. We also thank Danny Miller for useful
770 discussions and for providing the raw Nanopore reads from his study. Sarah Kingan and Jane
771 Landolin, from PacBio, were very helpful in exploring the causes for the artefactual repeats. We
772 thank Amanda Larracuente for advice on FISH protocols and the Cornell Imaging Facility for use
773 of their microscope. This project was funded by NIH grant number: GM116113 to RAW, ML, and
774 AGC. JMF was supported by an NSERC PGS D fellowship.

775
776
777 **References**

778 Ahmed-Braimah Y. H., R. L. Unckless, Andrew G. Clark, 2017 Evolutionary Dynamics of Male
779 Reproductive Genes in the *Drosophila virilis* Subgroup. G3. 7:3145-3155.

780 Beliveau B. J., A. N. Boettiger, M. S. Avendaño, R. Jungmann, R. B. McCole, *et al.*, 2015 Single-molecule
781 super-resolution imaging of chromosomes and in situ haplotype visualization using Oligopaint FISH
782 probes. Nat. Commun. 6: 7147.

783 Belyaeva E. S., I. F. Zhimulev, E. I. Volkova, A. A. Alekseyenko, Y. M. Moshkin, *et al.*, 1998 Su(UR)ES: a
784 gene suppressing DNA underreplication in intercalary and pericentric heterochromatin of
785 Drosophila melanogaster polytene chromosomes. Proc. Natl. Acad. Sci. U. S. A. 95: 7532–7537.

786 Benson G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research

27: 573–580.

Bolger A. M., M. Lohse, B. Usadel, 2014 Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

Bosco G., P. Campbell, J. T. Leiva-Neto, and T. A. Markow, 2007 Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177: 1277–1290.

Cabanettes F., and C. Klopp, 2018 D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ 6: e4958.

Caletka B. C., and B. F. McAllister, 2004 A genealogical view of chromosomal evolution and species delimitation in the Drosophila virilis species subgroup. Mol. Phylogenet. Evol. 33: 664–670.

Casacuberta E., and M.-L. Pardue, 2003 HeT-A elements in Drosophila virilis: retrotransposon telomeres are conserved across the Drosophila genus. Proc. Natl. Acad. Sci. U. S. A. 100: 14091–14096.

Chang C.-H., and A. M. Larracuente, 2019 Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the Drosophila melanogaster Y Chromosome. Genetics 211: 333–348.

Chang C.-H., A. Chavan, J. Palladino, X. Wei, N. M. C. Martins, *et al.*, 2019 Islands of retroelements are the major components of Drosophila centromeres. PLoS Biology 17(5):e3000241

Charlesworth B., C. H. Langley, and W. Stephan, 1986 The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics 112: 947–962.

Charlesworth B., P. Sniegowski, and W. Stephan, 1994 The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371: 215–220.

807 Dias G. B., P. Heringer, M. Svartman, and G. C. S. Kuhn, 2015 Helitrons shaping the genomic architecture
808 of Drosophila: enrichment of DINE-TR1 in α- and β-heterochromatin, satellite DNA emergence, and
809 piRNA expression. Chromosome Res. 23: 597–613.

810 Drosophila 12 Genomes Consortium, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman, *et al.*, 2007
811 Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.

812 Elder J. F. Jr, and B. J. Turner, 1995 Concerted evolution of repetitive DNA sequences in eukaryotes. Q.
813 Rev. Biol. 70: 297–320.

814 Elliott T. A., and T. Ryan Gregory, 2015 What's in a genome? The C-value enigma and the evolution of
815 eukaryotic genome content. Philos. Trans. R. Soc. Lond. B Biol. Sci. 370: 20140331.

816 Flynn J. M., I. Caldas, M. E. Cristescu, and A. G. Clark, 2017 Selection Constrains High Rates of Tandem
817 Repetitive DNA Mutation in Daphnia pulex. Genetics genetics.300146.2017.

818 Flynn J. M., S. E. Lower, D. A. Barbash, and A. G. Clark, 2018 Rates and Patterns of Mutation in Tandem
819 Repetitive DNA in Six Independent Lineages of Chlamydomonas reinhardtii. Genome Biol. Evol. 10:
820 1673–1686.

821 Gall J., E. Cohen, and M. Polan, 1971 Repetitive DNA sequences in Drosophila. Chromosoma 33.
822 https://doi.org/10.1007/bf00284948

823 Gall J. G., and D. D. Atherton, 1974 Satellite DNA sequences in Drosophila virilis. J. Mol. Biol. 85: 633–
824 664.

825 Gregory T. R., 2001 Coincidence, coevolution, or causation? DNA content, cell size, and the C-value
826 enigma. Biol. Rev. Camb. Philos. Soc. 76: 65–101.

827 Gregory T. R., and J. S. Johnston, 2008 Genome size diversity in the family Drosophilidae. Heredity 101:
828 228–238.

829 Gutzwiller F., C. R. Carmo, D. E. Miller, D. W. Rice, I. L. G. Newton, *et al.*, 2015 Dynamics of Wolbachia
830 pipientis Gene Expression Across the Drosophila melanogaster Life Cycle. G3 5: 2843–2856.

831 Harris R. S., M. Cechova, and K. Makova, Noise-Cancelling Repeat Finder: Uncovering tandem repeats in
832 error-prone long-read sequencing data:

833 Henikoff S., 2001 The Centromere Paradox: Stable Inheritance with Rapidly Evolving DNA. Science 293:
834 1098–1102.

835 Hughes S. E., D. E. Miller, A. L. Miller, and R. S. Hawley, 2018 Female Meiosis: Synapsis, Recombination,
836 and Segregation in. Genetics 208: 875–908.

837 Izumitani H. F., Y. Kusaka, S. Koshikawa, M. J. Toda, and T. Katoh, 2016 Phylogeography of the Subgenus
838 Drosophila (Diptera: Drosophilidae): Evolutionary History of Faunal Divergence between the Old
839 and the New Worlds. PLoS One 11: e0160051.

840 Jagannathan M., R. Cummings, and Y. M. Yamashita, 2018 A conserved function for pericentromeric
841 satellite DNA. Elife 7. https://doi.org/10.7554/eLife.34122

842 Jagannathan M., R. Cummings, and Y. M. Yamashita, 2019 The modular mechanism of chromocenter
843 formation in. Elife 8. https://doi.org/10.7554/eLife.43938

844 Kim J. C., J. Nordman, F. Xie, H. Kashevsky, T. Eng, *et al.*, 2011 Integrative analysis of gene amplification
845 in Drosophila follicle cells: parameters of origin activation and repression. Genes Dev. 25: 1384–
846 1398.

847 Larracuente A. M., and P. M. Ferree, 2015 Simple Method for Fluorescence DNA In Situ Hybridization to
848 Squashed Chromosomes. Journal of Visualized Experiments.

849 Lemos B., A. T. Branco, D. L. Hartl, 2010 Epigenetic effects of polymorphic Y chromosomes modulate
850 chromatin components, immune response, and sexual conflict. PNAS 107:15826-15831.

851 Lima L. G. de, M. Svartman, and G. C. S. Kuhn, 2017 Dissecting the Satellite DNA Landscape in Three
852 Cactophilic Sequenced Genomes. G3 7: 2831–2843.

853 Lowman H., and M. Bina, 1990 Correlation between dinucleotide periodicities and nucleosome
854 positioning on mouse satellite DNA. Biopolymers 30: 861–876.

855 Malik H.S. and J.J. Bayes, 2006. Genetic conflicts during meiosis and the evolutionary origins of
856 centromere complexity. Biochem Soc Trans. 34: 569-573.

857 Maio J. J., F. L. Brown, and P. R. Musich, 1977 Subunit structure of chromatin and the organization of
858 eukaryotic highly repetitive DNA: recurrent periodicities and models for the evolutionary origins of
859 repetitive DNA. J. Mol. Biol. 117: 637–655.

860 Mehrotra S., and K. S. McKim, 2006 Temporal Analysis of Meiotic DNA Double-Strand Break Formation
861 and Repair in Drosophila Females. PLoS Genetics 2: e200.

862 Miller D. E., C. Staber, J. Zeitlinger, and R. Scott Hawley, 2018 Highly Contiguous Genome Assemblies of
863 15 Drosophila Species Generated Using Nanopore Sequencing. G3: Genes|Genomes|Genetics 8:
864 3131–3141.

865 Mills W. K., Y. C. G. Lee, A. M. Kochendoerfer, E. M. Dunleavy, G. H. Karpen. 2019 RNA transcribed from
866 heterochromatic simple-tandem repeats are required for male fertility and histone-protamine
867 exchange in *D. melanogaster*. Biorxiv doi: https://doi.org/10.1101/617175.

868 Mitsuhashi S., M. C. Frith, T. Mizuguchi, S. Miyatake, T. Toyota, *et al.*, 2019 Tandem-genotypes: Robust
869 detection of tandem repeat expansions from long DNA reads. Genome Biology 20:58.

870 Ohno S., 1972 So much "junk" DNA in our genome. Brookhaven Symp. Biol. 23: 366–370.

871 Ono Y., K. Asai, and M. Hamada, 2013 PBSIM: PacBio reads simulator—toward accurate genome
872 assembly. Bioinformatics 29: 119–121.

873 Ostrega M. S., and V. Thompson, 1986 Mitochondrial DNA Restriction site polymorphism in Drosophila
874 montana and drosophila virilis. Biochemical Systematics and Ecology 14: 515–519.

875 Parker D. J., R. A. W. Wiberg, U. Trivedi, V. I. Tyukmaeva, K. Gharbi, *et al.*, 2018 Inter and Intraspecific
876 Genomic Divergence in Drosophila montana Shows Evidence for Cold Adaptation. Genome Biol.
877 Evol. 10: 2086–2101.

878 Pavlek M., Y. Gelfand, M. Plohl, and N. Meštrović, 2015 Genome-wide analysis of tandem repeats in
879 Tribolium castaneum genome reveals abundant and highly dynamic tandem repeat families with
880 satellite DNA features in euchromatic chromosomal arms. DNA Res. 22: 387–401.

881 Schubert I., and M. A. Lysak, 2011 Interpretation of karyotype evolution should consider chromosome
882 structural constraints. Trends Genet. 27: 207–216.

883 Smith A. V., and T. L. Orr-Weaver, 1991 The regulation of the cell cycle during Drosophila
884 embryogenesis: the transition to polyteny. Development 112: 997–1008.

885 Spicer G. S., and C. D. Bell, 2002 Molecular Phylogeny of the Drosophila virilis Species Group (Diptera:
886 Drosophilidae) Inferred from Mitochondrial 12S and 16S Ribosomal RNA Genes. Annals of the
887 Entomological Society of America 95: 156–161.

888 Stage D. E., T. H. Eickbush, 2009 Origin of nascent lineages and the mechanisms used to prime second-
889 strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. Genome Biology 10:R49.

890 Stewart B. S., Y. H. Ahmed-Braimah, D. G. Cerne, B. F. McAllister, Female meiotic drive preferentially
891 segregates derived metacentric chromosomes in Drosophila.
892
893 Throckmorton LH: *The Genetics and Biology of Drosophila Volume 3b*. Academic Press: New York; 1982.
894
895 Walsh J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence DNAs.
896 Genetics 115: 553–567.

897 Wei K. H.-C., J. K. Grenier, D. A. Barbash, and A. G. Clark, 2014 Correlated variation and population
898 differentiation in satellite DNA abundance among lines of Drosophila melanogaster. Proc. Natl.
899 Acad. Sci. U. S. A. 111: 18793–18798.

900 Wei K. H.-C., S. E. Lower, I. V. Caldas, T. J. S. Sless, D. A. Barbash, *et al.*, 2018 Variable Rates of Simple
901 Satellite Gains across the Drosophila Phylogeny. Mol. Biol. Evol. 35: 925–941.

902 Yarosh W., and A. C. Spradling, 2014 Incomplete replication generates somatic DNA alterations within
903 Drosophila polytene salivary gland cells. Genes & Development 28: 1840–1855.

904 **Supporting Information Captions**

905 Figure S1. NCRF Simulations results. Running NCRF with different max error allowed on reads from
906 PBSIM off a mock genome composed of perfect repeats.

907 Figure S2. Polyteny vs. diploid analysis. (A) Microsatellites expected to not be exclusive in
908 pericentromeric heterochromatin (control for Fig 2E). (B) Analysis of *D. melanogaster* sequencing
909 data from embryos, flies, and salivary glands. (C) Microsatellites in *D. melanogaster* data, control

910 for Fig S2B.

911 Fig S3. Satellite containing reads are enriched for low quality scores in Illumina data. A) Quality score
912 distribution in the raw reads. (B) Quality score distribution after quality filtering. (C) GC distribution
913 of raw reads. (D) GC distribution of reads after quality filtering.

914 Fig S4. Satellite containing reads are enriched for low quality reads in one Illumina dataset (A) produced
915 on a NextSeq, but not on a second dataset (B) produced on a HiSeq.

916 Figure S5. Dotplot of D. novamexicana contig 000018F, which contained a high amount of the 32 bp
917 satellite (from 0-1.5 M) and its similarity to a large region on *D. virilis* chromosome 6 (dot).

918 Fig S6. FISH image of D. virilis male. The AAACAAC satellite appears on a single chromosome pair. This is
919 clear from the metaphase chromosomes (middle) as well as the interphase cells where two distinct
920 foci are localized. The distinct foci in interphase cells is in contrast to the AAACTAC satellite, which
921 takes up a large portion of the nucleus.

922 Fig S7. PCA of satellite DNA copy number of D. virilis group strains. (A) Using only the abundances of the
923 satellites AAACTAC, AAATTAC, AAACTAT, AAACAAC. (B) Using the abundances of the 20 most
924 abundant simple satellites.

925 Fig S8. Comparison of D. americana strains with and without the X-4 chromosomal fusion. (A) PCA using
926 the top 20 simple satellites. (B) Boxplot of AAACTAC copy number. (C) Boxplot of AAACAAC copy
927 number.

928 Fig S9. DINE amount vs. AAACTAC copy number. DINE amount is the number of reads mapped to the
929 DINE repeat / estimated depth.