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

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

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

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

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

Corresponding author:

Jullien M. Flynn

jmf422@cornell.edu

**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 highly abundant satellite family arose from a related satellite in the branch leading to the virilis phylad 4.5 - 11 million years ago before exploding in abundance in some species of the clade. 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. By analyzing multiple strains of *D. virilis*, we saw that one centromere-proximal satellite is increasing in abundance along a geographical gradient while the other is contracting in an anti-correlated manner, suggesting ongoing conflicts at the centromere. In conclusion, we illuminate several key attributes of satellite evolutionary dynamics that we hypothesize to be driven by processes like selection, meiotic drive, and constraints on satellite sequence and abundance.

**Introduction**

33 Repetitive DNA is abundant in most eukaryotic genomes, and is now understood to be correlated
34 with the manifold variation in genome size across the tree of life (Elliott and Gregory 2015). For
35 most species, transposable elements (TEs) dominate the repeat landscape, including in humans,
36 plants, and *Drosophila melanogaster*. Satellite DNA, which is characterized by tandem repeats
37 spanning long arrays, very rarely has dominated a genome to a similar extent as TEs. An
38 unprecedented case is that of *Drosophila virilis*, the Drosophila species with the largest estimated
39 genome size (up to 389 Mb) (Bosco *et al.* 2007), where some 40% of the genome is comprised of
40 just three simple 7-mer satellites: AAACTAC, AAACTAT, and AAATTAC (Gall *et al.* 1971; Gall and
41 Atherton 1974). Since the 1970s, there has been no follow-up to validate the amount of 7-mers
42 with modern techniques, or evolutionary studies to understand how and why these satellite repeats
43 expanded so explosively. The genomic composition of simple satellites in *D. virilis* provides an
44 excellent model for an investigation of the evolutionary dynamics involved in their expansion in the
45 genome as well as the technical challenges facing simple satellite analysis.
46 Satellites are rapidly evolving in sequence and copy number, and there is a high level of
47 variation in satellite content among and within species (Wei *et al.* 2014, 2018). The reasons for such
48 dramatic variation is not well understood, and cannot be fully explained by current models.
49 Satellites have been long hypothesized to be slightly deleterious and therefore governed primarily
50 by the strength of negative selection (Ohno 1972). However, the amount of satellite in the genome
51 that causes negative effects that could be selected against depends on many factors and cannot be
52 easily predicted (Charlesworth *et al.* 1994; Gregory 2001). The fact that most organisms have
53 satellite repeats in or near centromeres suggests that they are important for centromere function.
54 Satellite repeats can also be important for maintenance of the chromocenter and packaging of
55 chromosomes in the nucleus (Jagannathan *et al.* 2018, 2019), and the transcripts of some satellites

may be essential for fertility (Mills *et al*. 2019). In heterozygotes with alleles that differ in pericentromeric satellite sequence or abundance, one allele may assemble a stronger kinetochore during female meiosis I, increasing its probability of transmission into the egg (rather than polar bodies). This transmission advantage, known as centromere drive, allows satellites to rapidly change in composition in the population, regardless of their whole-organism fitness effects (Henikoff et al. 2001). If satellite DNA is an essential component of genomes or is only a burden (i.e. is selfish), it is still not clear why some species have almost no pericentromeric satellite DNA while others, like *D. virilis*, possess pericentromeric satellites that make up almost half of the genome.

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

78 Genome-wide characterization of satellites has taken off since high-throughput sequencing
79 has become widely available. We have learned from several informative studies about the
80 sequences and relative abundances of satellites in various species (Pavlek *et al.* 2015; Flynn *et al.*
81 2017; de Lima *et al.* 2017; Wei *et al.* 2018), but technical challenges may prevent accurate
82 quantitative estimates. Satellites may be more prone to errors or biases in the sequencing process
83 that do not affect the better studied regions of the genome. Satellites are difficult to assemble even
84 with long-read sequencing (Chang and Larracuente 2019). The genome assembly of *D. virilis* is
85 approximately half its estimated genome size by flow cytometry (~200 Mb vs 389 Mb) (Bosco *et al.*
86 2007), and it is likely that much of what is missing is simple satellite DNA. However, even using
87 alignment-free raw read methods have not produced satellite DNA estimates that approach the
88 amount that is missing from the genome assembly and was estimated from early work (Gall *et al.*
89 1971; Gall and Atherton 1974; Wei *et al.* 2018). Now, as long read sequencing is also being
90 exploited to study satellites, we must evaluate satellite DNA abundance estimates to assess if there
91 are platform-specific biases that may affect evolutionary analysis of satellite DNA.
92 The purpose of this paper is two-fold; first to explore the technical biases preventing
93 accurate characterization and quantification of simple satellites, and second to use a comparative
94 approach to understand the evolutionary dynamics of the extremely abundant 7mers in the *D. virilis*
95 group. First, we characterize satellites in *D. virilis* sequencing data from different platforms and
96 assess biases that affect accurate satellite characterization. We then use comparative genomics and
97 cytogenetics in *D. virilis* and its sister species to understand the composition and changes in the
98 highly abundant simple satellites. Finally we sequence multiple strains of *D. virilis* and sister species
99 to estimate polymorphism in satellite abundance and infer processes that may be influencing their
100 evolution. From this we infer that there are likely a variety of understudied processes affecting

101 satellite DNA in this organism, including positive selection, meiotic drive, and constraints and trade-
102 offs between satellites.
103
104 **RESULTS**
105
106 **Technical biases in characterizing simple satellites from sequencing**
107 *Long-read genome assemblies have an under-representation of simple satellites*
108 Long-read sequencing technologies have an advantage because of their long reads, but a
109 disadvantage due to their high error rate, prompting a need for extensive alignments for error-
110 correction and assembly. First we asked whether assemblies from long read technologies can better
111 assemble simple satellite reads than the previous Sanger assembly. We compared the amount of
112 simple 7-mer satellites (AAACTAC, AAACTAT, AAATTAC, AAACAAC) in three *D. virilis* genome
113 assemblies: the CAF1 assembly produced from Sanger sequencing (Drosophila 12 Genomes
114 Consortium *et al.* 2007), a PacBio assembly produced by our group by ~100x coverage (available at
115 https://www.ncbi.nlm.nih.gov/bioproject/?term=txid7214[Organism:noexp]), and a Nanopore
116 assembly produced from ~20x sequencing coverage (Miller *et al.* 2018). All assemblies were
117 approximately the same size at ~200 Mb. The PacBio and Nanopore assemblies contained a
118 similarly low amount of simple 7-mer satellites, 29 and 28 kb, respectively. The CAF1 assembly,
119 however contained 7.36 Mb of these satellites. This discrepancy is likely largely due to the
120 difference in assembly algorithms used for short read and long read data. Long reads must be
121 aligned and corrected to be incorporated into the assembly because of their high error rate,
122 whereas this is not necessary for Sanger-based assemblies. Use of modified methods can improve
123 assemblies of repetitive regions (Chang and Larracuente 2019), but for highly homogeneous simple

124 satellites, whose arrays span 10-100x longer than the current maximum read length, it is practically
125 impossible to produce a continuous assembly.
126
127 *Simulations to assess simple repeat quantification from long read sequencing data*
128 Due to assembly issues of simple satellites, they must be quantified from raw unassembled reads.
129 Long read sequencing data poses a significant challenge because of the high error rate including a
130 high indel rate in the raw reads. We therefore used two different approaches along with
131 simulations to assess their accuracy. The first approach used k-Seek (Wei et al. 2014) to select
132 repeat-rich reads and then Phobos (https://www.ruhr-uni-bochum.de/ecoevo/cm/cm_phobos.htm)
133 to quantify satellites. This approach allows for *de novo* discovery of satellite sequences. We used
134 Noise-Cancelling Repeat Finder (NCRF, Harris *et al.* 2019) for our second approach, providing our
135 target satellites. Both methods are relatively sensitive to imperfect repeats, which we expect with
136 the high error rate of long-read sequencing.
137 To evaluate our approaches, we created a mock *D. virilis*-like genome containing
138 pericentromeric and centromeric repeats on each of five chromosomes (See Materials and
139 Methods). We then simulated 10x PacBio reads from this genome, and then quantified satellites
140 using both approaches. NCRF works by doing alignments of target satellites to the reads and
141 allowing up to a user-specified maximum divergence. To determine the most appropriate maximum
142 divergence, we simulated a range of values for this parameter from 18-30% and chose the lowest
143 asymptotic value - which was 25% in this case (Figure S1). NCRF found almost the same amount of
144 satellites that truly existed in the mock genome whereas the k-Seek + Phobos method only found
145 about 20% (Figure 1A).
146

147 *The amounts and biases in simple 7mer repeats differ between Nanopore and PacBio sequencing*
148 *reads*
149 Next, we quantified simple satellites in the long-read data generated from our 100x PacBio
150 sequencing and 20x Nanopore sequencing using the two approaches mentioned above. Unlike in
151 the simulations, both approaches produced very similar (but lower than expected) estimates at 8.8-
152 10.9 Mb for the PacBio data (Figure 1B). The Nanopore data contained almost 3 times the 7mer
153 satellites compared to PacBio, with 23.4 - 28.2 Mb (Figure 1C). This may represent a platform-
154 specific difference in the ability to sequence long arrays of simple tandem repeats. Both the PacBio
155 reads and the Nanopore reads contained a greater amount of simple satellites than data produced
156 in our lab previously with Illumina HiSeq sequencing (Wei et al. 2018), however did not approach
157 the estimated >100 Mb in the genome.
158 Both the PacBio and Nanopore reads contained large amounts of what we expect to be
159 artefactual repeats, which were found with the k-Seek + Phobos approach, and validated with
160 NCRF. NCRF found 4.4 Mb (normalized to 1x genome coverage) of AAACGAC in the PacBio reads.
161 This satellite was not found in the Nanopore data or Illumina data (this and previous studies) or in
162 previous studies that characterized the most abundant satellites in *D. virilis*. Manual inspection
163 proved that the AAACGAC satellite was the true consensus found in long arrays in the reads and did
164 not represent an error in our approaches' characterization of satellites. Similarly, AAATCAC,
165 AGCCTAT, ACAGGCT, and AATGG were found in megabase quantities (after normalization) in the
166 Nanopore data - whereas these satellites were not found in Illumina or PacBio data. We suggest
167 these satellites are also technical artifacts introduced at the base-calling level.
168

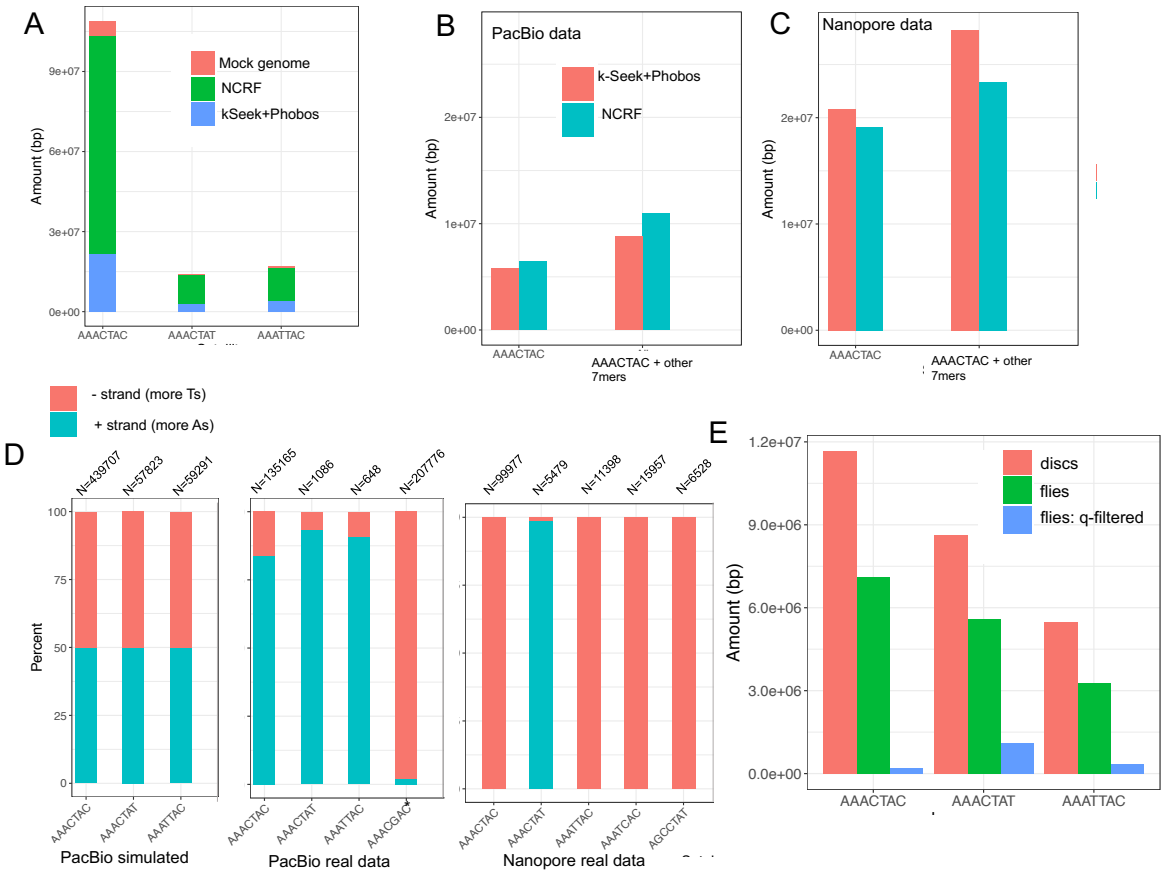


**Fig1**: Issues in quantifying simple satellites in sequencing data (all data shown is *D. virilis*). (**A**) Cumulative stacked barplot comparing the performance of the two tested approaches on PacBio data simulated with PBSim from a mock genome. (**B)** Comparing the results of the two approaches 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.

In the PacBio data, the relative amounts of 7mer satellites (AAACTAC, AAACTAT, and AAATTAC) were lower than expected. This additional evidence led us to hypothesize that there were context-specific errors in our PacBio data affecting our particular satellites. If the sequencing were unbiased, we would expect to have an equal amount of satellites being detected on reads coming from both DNA strands. We evaluated the strand bias in the simulated and real long-read data for the three most abundant true satellites, as well as some artefactual satellites. We

176 arbitrarily label the positive strand as AAACTAC and the negative strand as GTAGTTT, etc. In the
177 simulated data, the positive and negative strands of satellites were detected in equal amounts
178 (Figure 1D). However, there was a strong strand bias for all satellites in both the PacBio and
179 Nanopore data (Figure 1D). For PacBio, the real satellites AAACTAC, AAACTAT, AAATTAC had a
180 positive strand bias, whereas the artefactual satellite had a negative strand bias: 98% of the reads
181 with this satellite were from the negative strand. Based on communication with PacBio
182 representatives, this issue seemed to be caused by context-specific issues with base calling
183 algorithms used for this sequencing run. As base calling algorithms improve, these issues will likely
184 begin to be remedied. In fact, we received PacBio Circular Consensus Sequencing or "HiFi" data for
185 a closely related species, *D. americana*, and the base-calling issue was remedied. In the Nanopore
186 data, strand biases were even more extreme: the negative strand was sequenced almost exclusively
187 for real satellites AAACTAC and AAATTAC and suspect satellite AAATCAC. However, the AAACTAT
188 real satellite was sequenced almost exclusively on the positive strand. In this case, strand biases
189 may be caused by unsequenceable secondary structures developing more frequently on one strand
190 of the satellite DNA than the other. We analyzed Illumina NextSeq reads for *D. virilis*, and no such
191 strand bias was found.
192
193 *D. virilis whole-flies have 40% less pericentromeric satellites than non-polytene tissue*
194
195 Polyteny occurs in all differentiated tissues of Dipterans, and is characterized by multiple
196 rounds of local DNA replication within the same nucleus and without cell division, a process known
197 as endoreduplication (Smith and Orr-Weaver 1991; Kim *et al.* 2011). However, the pericentromeric
198 heterochromatin, where most satellite DNA is located, is under-replicated (Belyaeva *et al.* 1998). It

199 has never been tested if the level of polyteny in an adult fly makes a difference in the estimate of
200 satellites per genome. Thus, we sequenced adult male flies (which have multiple polytene tissues)
201 and imaginal discs (which are diploid) from male larvae and compared the amount of simple
202 satellites in these datasets. We used Illumina sequencing and PCR-free library preparations to
203 reduce known PCR bias (Wei *et al.* 2018). We found that for each of the four most abundant 7mer
204 satellites in the *D. virilis* genome, there was approximately 40% less in the flies compared to the
205 imaginal discs (Figure 1E). This pattern is not observed for microsatellites which are known to
206 localize outside of pericentromeric heterochromatin (Figure S2A). We also analyzed publicly
207 available *D. melanogaster* data, including flies, imaginal discs, and salivary glands (which are the
208 most extreme in polyteny), and observed this same pattern of under-replication of satellite repeats
209 in polytene tissues (Figure S2B and S2C).
210
211 *Reads with tandem repeats had lower quality scores in Illumina data*
212
213 Upon inspection with FastQC of our data from the polyteny analysis, we found a bimodal
214 distribution of quality scores, with one peak at 22 and another at 37 (Figure S3A). After filtering low
215 quality reads, the majority of the reads with simple satellites were removed (Figure S3). The
216 quantity of satellites was reduced by ~15 x after quality filtering (Figure 1E). It is apparent that in
217 our dataset, simple satellite-containing reads were highly enriched for low quality scores. We
218 examined other published *D. virilis* Illumina datasets to evaluate if this issue existed in other
219 sequencing runs. Two other datasets were available and the one that was produced on the Illumina
220 NextSeq platform like our data (Miller *et al.* 2018) showed the same pattern of biased quality scores
221 in repetitive reads (Figure S4). The dataset produced on the HiSeq platform (Wei *et al.* 2018) did not

222 show this pattern. It should be noted however that the amounts of 7mer satellites sequenced in the
223 NextSeq datasets were higher than the HiSeq dataset. Our libraries were multiplexed with other
224 non-*D. virilis* group samples from unrelated projects and only represented ~20% of the total
225 sequenced lane so that we would not have issues related to low complexity. We also noticed this
226 pattern (but less dramatically) in our Illumina sequencing of multiple strains.
227
228 **Related species have similar but fewer simple repeats**
229
230 *D. novamexicana* and *americana* which are 0.38 MY diverged from each other, are sister species of
231 *D. virilis*, which is approximately 4.5 MY diverged (Caletka and McAllister 2004) (Figure 2A). We
232 sequenced these species with high coverage PacBio runs and characterized and quantified satellites.
233 We emphasize the comparison of relative satellite amounts since all are likely under-represented.
234 *D. americana* was sequenced with PacBio HiFi reads, which eliminated artefactual satellites, but
235 make quantitative comparisons difficult since different chemistries have different efficiencies of
236 sequencing satellites. Nevertheless, we also found a high enrichment of 7bp satellites in *D.*
237 *novamexicana* and *D. americana* (Fig 2B). Interestingly, we found the most abundant satellite in *D.*
238 *virilis*, AAACTAC, is also the most abundant in *D. novamexicana* and *D. americana*, albeit with about
239 half the total amount. The second and third most abundant repeats, AAACTAT and AAATTAC,
240 however were not present in long tandem arrays in *D. novamexicana*. The second most abundant
241 satellite in *D. novamexicana* and *americana* was AAACAAC, whereas in *D. virilis* there is only a few
242 kilobases.
243 By analyzing sequencing data in more diverged species, we can infer when the AAACTAC
244 satellite family arose. *D. hydei* is approximately 26 MY diverged from *D. virilis* (Izumitani *et al.* 2016),

245 and we had PacBio long read data for this species. Here 7 bp satellites are again the most enriched
246 (Fig 2B), but the sequences are unrelated to those in *D. virilis* (ACCCATG, AAAGGTC from PacBio
247 data). We analyzed Illumina data for *D. montana*, another member of the virilis group that is 7-11
248 MY diverged from *D. virilis* (Ostrega and Thompson 1986; Spicer and Bell 2002) (Figure 2A). This
249 species does not have any AAACTAC family satellites, and in fact no enrichment of 7 bp satellites.
250 The most abundant satellite in *D. montana* is AAAC. From these data, we infer that the AAACTAC
251 family of satellites arose in the clade leading to the *D. virilis* phylad 4.5-11 MYA. We also analyzed
252 Illumina sequencing data for *D. lummei*, which is 3 MY diverged from *D. novamexicana/americana*
253 (Fig 2A). AAACTAC is conserved in *D. lummei,* but it is the only enriched 7 bp satellite in this species
254 and its relative estimated abundance is lower than the other three *D. virilis* phylad species.
255
256 **Complex satellites are also abundant in *D. virilis* group genomes**
257 We searched the high-quality genome assemblies for complex satellites (defined here as unit
258 lengths greater than 20 bp). In *D. virilis,* we found a 36-bp satellite
259 AAAACGACATAACTCCGCGCGGAGATATGACGTTCC making up ~800 kb of the assembly. This satellite
260 was found in previous studies and is thought to be associated with the possibly mobile element pDv
261 (Zelentsova et al 1986, Heikkinen et al. 1995). In *D. novamexicana*, we found a 32 bp satellite
262 AAAAGCTGATTGCTATATGTGCAATAGCTGAC along with a related 29 bp satellite. The 32 bp satellite
263 spanned over 1.1 Mb on a single 3 Mb contig in the *D. novamexicana* assembly. The non-satellite
264 portion of the contig had similarity to chromosome 6 (dot chromosome/Muller element F) (Figure
265 S5). In *D. americana*, we found this identical 32 bp satellite, but in total its span was only ~150 kb. In
266 all *D. virilis* group species, we also found a series of similar satellites varying in size (150-500 bp)

267 related to the previously described helitron central repeat that has expanded to tandem repeats in

268 the virilis group (Dias *et al.* 2015).

269

270 **Fluorescence *in situ* hybridization reveals evolutionary dynamics of 7 bp repeats**

271

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

273 *virilis* group. From our sequencing data, we know that the AAACTAC satellite is conserved between

274 *D. virilis, D. novamexicana,* and *D. americana*, but the abundance varies by approximately two-fold.

275 The second most abundant satellites have turned over between *D. virilis* and

276 *novamexicana*/*americana*. We used FISH of the most abundant 7mers (AAACTAC, AAACTAT,

277 AAATTAC, AAACAAC) in these three sister species. *D. virilis* and *D. novamexicana* have the same

278 karyotype with five acrocentric chromosomes plus the very small F element or "dot chromosome".

279 The strain of *D. americana* we used has centromere-centromere fusions between the X and 4$^{th}$

280 chromosomes and the 2$^{nd}$ and 3$^{rd}$ chromosomes.

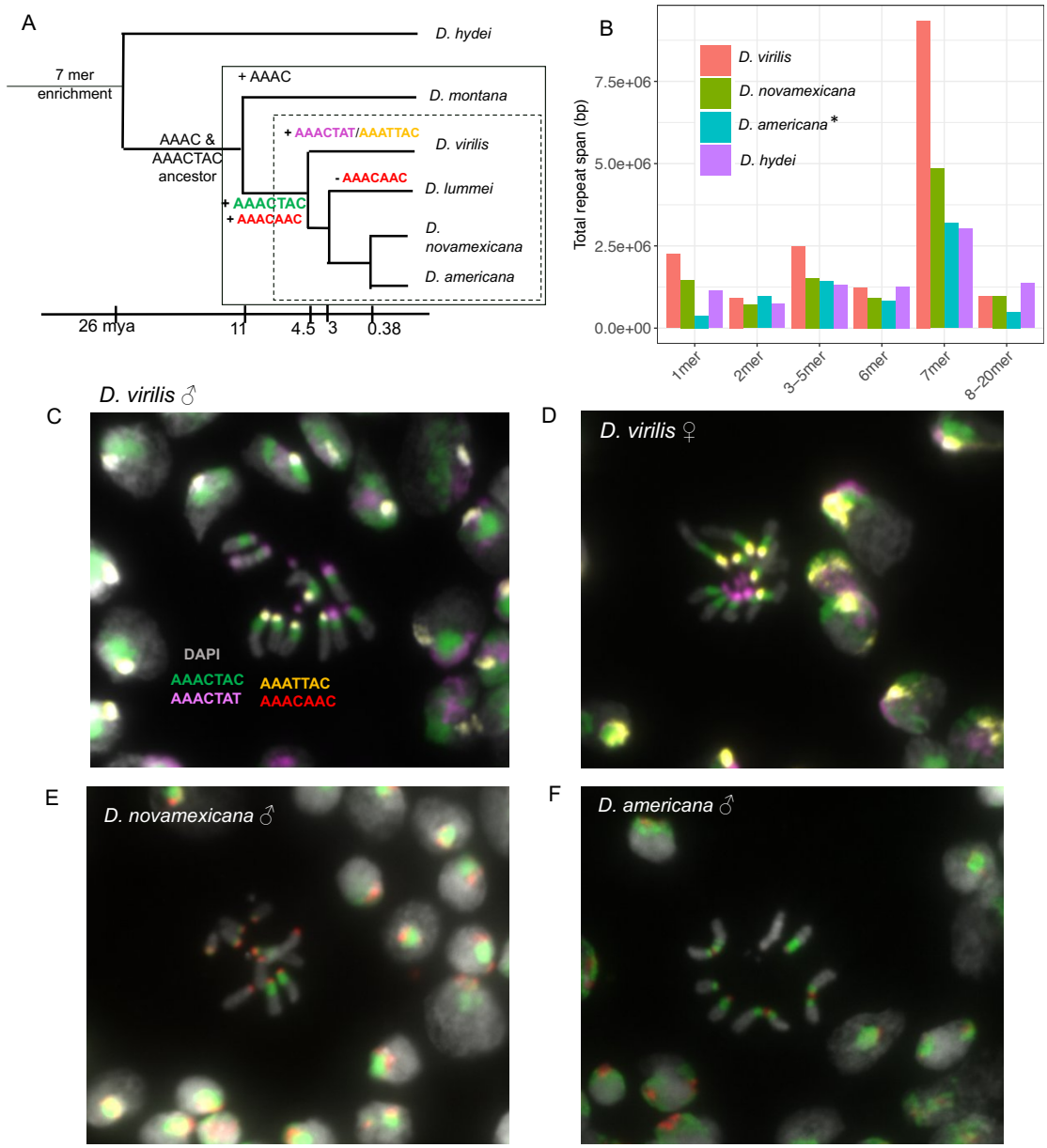


**Fig 2**. Comparative analysis of simple satellites in the *D. virilis* group. (A) Phylogeny demonstrating when satellites arose (+) and were lost (-). Dashed box: virilis phylad; solid box: virilis group. (B) Total amount of different unit lengths (k-mers) of satellites across four related species. The (*) for *D. americana* indicates that it was sequenced with PacBio HiFi reads, whereas the other species were sequenced with chemistry version 2.0. (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

283 pericentromeric region on the five pairs of large chromosomes. The Y chromosome appears to have
284 slightly less AAACTAC satellite. The second and third most abundant satellites, AAATTAC and
285 AAACTAT, are localized more proximally at or near the centromere. There are five single
286 chromosomes having each of these satellite, indicating that one chromosome pair has different
287 satellite content - which we hypothesized to be the X and Y. Based on differences between male
288 and female FISH results (Figure 2C and 2D), we suggest the Y chromosome has AAACTAT at both
289 distal ends of the chromosome and AAACTAC only flanking one end, whereas the X chromosome
290 has the other centromeric repeat AAATTAC. We were also able to visualize the dot chromosomes in
291 *D. virilis*, which we find is mostly composed of AAACTAT. The AAACAAC satellite is present in small
292 amounts in *D. virilis*, very likely on a single chromosome (Figure S6).
293 We estimated that *D. novamexicana* has approximately half the AAACTAC as *D. virilis*, and
294 visualizing it with FISH reveals a pattern that suggests aspects of its evolution. Its pericentromeric
295 localization is conserved. One chromosome pair has the same amount of AAACTAC as *D. virilis*,
296 whereas all other chromosomes have a very small amount (Figure 2E). Based on the FISH images, it
297 appears that it is the 5th chromosome in *D. novamexicana* that has the greatest amount of
298 pericentromeric AAACTAC conserved. The centromeric repeat on all major chromosomes is
299 AAACAAC in *D. novamexicana* and *D. americana.* Our images illustrate clearly the centromere-
300 centromere fusion between chromosome X-4 and 2-3 in *D. americana* with the satellites being
301 maintained on both sides of the fusion (Figure 2F). None of the four simple satellite probes bound
302 to the Y chromosome of *D. novamexicana* or *D. americana*. Based on the images we suggest that *D.*
303 *americana* has an intermediate amount of pericentromeric AAACTAC satellite compared to *D. virilis*
304 and *novamexicana*.
305

306 **Some satellite-containing reads are linked to TEs**

307

308 We used RepeatMasker to detect if any of the reads containing satellites also contain transposable
309 elements. TEs might be located in islands within the simple repeats at the centromere as in *D.*
310 *melanogaster* (Chang *et al.* 2019) (Figure 3A), in more distal regions flanking the pericentromeric
311 heterochromatin (Figure 3B), or some TEs may have inserted into long pericentromeric satellite
312 arrays (Figure 3C). For AAACTAC (and its artefactual counterpart AAACGAC), ~3.5% of reads
313 (2473/75,364) also contained at least 500 bp of a TE insertion. In the satellite reads that also
314 contain TE sequences, TEs were enriched at the beginning and ends of reads, concordant with the
315 hypothesis that a high proportion of the TEs we found are flanking the long arrays of AAACTAC
316 distally (Figure 3B). In order to understand how many reads would be expected to contain both
317 satellites and TEs if TEs flanked this satellite and were not interspersed, we simulated a situation
318 where a large satellite block flanked a large TE block. This simulation revealed a much smaller
319 amount of reads containing both satellites and TEs (0.06 %). This result suggests that not only are
320 TEs flanking the pericentromeric satellite AAACTAC, but there have likely been TE insertions into the
321 satellite arrays. Likely flanking the proximal end of the pericentromeric satellite are the centromeric
322 satellites AAACTAT or AAATTAC. 144 reads contained both AAACTAC and AAACTAT repeats (0.19%
323 of AAACTAC reads) and 94 reads contained both AAACTAC and AAATTAC repeats (0.12% of
324 AAACTAC reads). Based on our simulations, these proportions of overlapping reads are consistent
325 with our expectation based on our FISH results that the pericentromeric and centromeric satellites
326 have relatively clean boundaries and they directly flank each other.

327

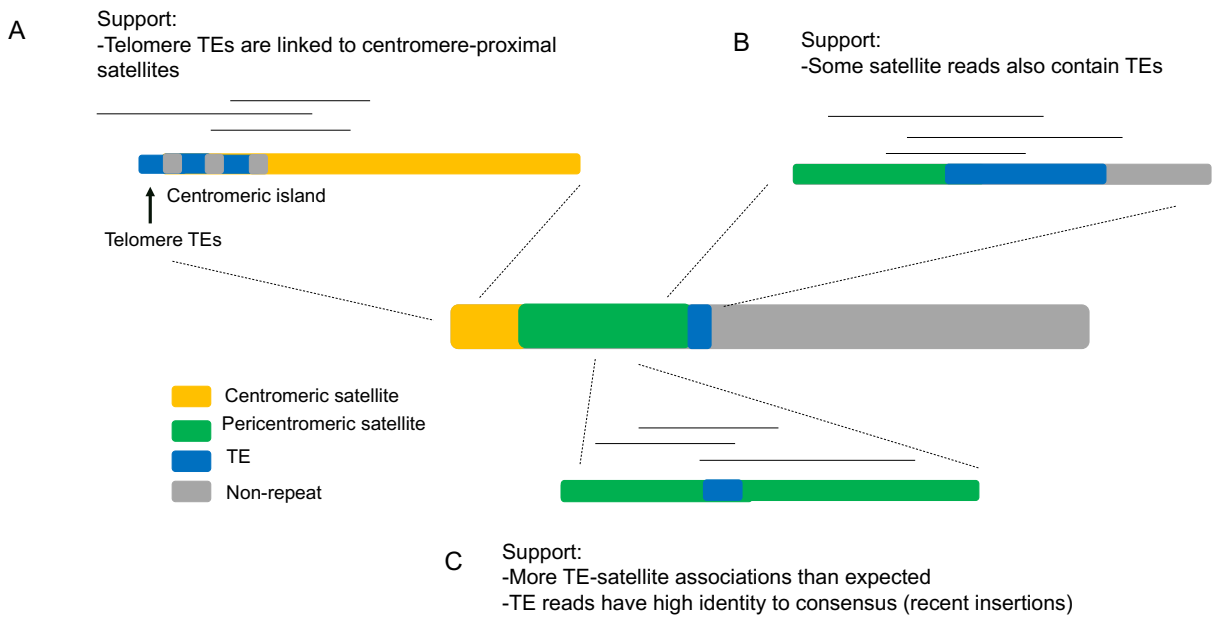


**Fig 3**. Transposable elements are in close proximity to satellite arrays. (**A**) Satellites near the centromere may be linked to potential islands of retroelements (Chang et al. 2019) and telomeric TEs. (**B**) Heterochromatic TEs flank satellite arrays. Histograms show the start and end position of transposable element in the satellite-rich read. (**C**) Transposable elements may have inserted into the satellite arrays.

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

339 consensus in the centromere-proximal satellite reads. 288/300 and 46/49 TE containing reads had a
340 TE insertion > 500 bp with < 15% divergence for AAACTAT and AAATTAC, respectively.
341 There were differences in the TE composition of reads with different satellites. For the
342 pericentromeric satellite AAACTAC, Gypsy-10_Dvi was the most enriched, followed by Helitrons
343 (Helitron-1N1_DVir, Helitron-1_DVir, Helitron-2N1_DVir, Helitron-2_DVir). For the AAACTAT
344 centromeric satellite, Gypsy-10_Dvi was again the most enriched, followed by Penelope. For the
345 AAATTAC centromeric satellite, Gypsy-2_DVir was the most enriched followed by Penelope. In both
346 AAACTAC and AAACTAT reads, CR1-1_DVi was the second or third most abundant TE. Interestingly,
347 R1 was present in relatively high amounts in AAACTAC. In 110/132 of these R1-AAACTAC reads,
348 rDNA sequences were not also linked. This suggests that some R1 elements, which are generally
349 localized to rDNA loci, have jumped into or near satellite arrays. This is concordant with findings
350 that some R1 elements are located outside rDNA loci in Drosophila (Stage and Eikbush 2009). All
351 centromeres in *D. virilis* are acrocentric, meaning that the telomeric TEs Het-A and TART
352 (Casacuberta and Pardue 2003) are likely near the centromere satellites. We found 12 reads linked
353 to AAATTAC that contained matches to TART. Only two reads linked to any satellite contained a
354 sequence matching HeT-A. We also used BLAST to detect matches between the genome assembly
355 (masked from the 7mer satellites) and the 7mer satellite reads. We could not detect any unique
356 regions of the genome that matched non-satellite sequence on the reads because they had low
357 quality matches to hundreds of places in the genome each.
358
359 **Variation in *D. virilis* group global strains**
360

361 *D. virilis* is globally distributed while its sister species are localized to North America, with *D.*
362 *novamexicana* more restricted than *D. americana.* Patterns of variation in satellites may reveal
363 potential mechanisms that can be hypothesized to be driving satellite evolution. Additionally, *D.*
364 *americana* has a polymorphic fusion between the X and 4th chromosomes, so we may be able to
365 identify differences in satellite composition associated with the fusion. This fusion has been shown
366 to be currently undergoing meiotic drive, potentially mediated by a larger total centromere or
367 pericentromere size in the fused strains compared to the non-fused strains (Stewart et al. 2019). On
368 the other hand, chromosome fusions are often caused by Robertsonian translocations with loss of
369 some non-essential DNA, which might include pericentromeric satellites (Schubert and Lysak 2011).
370 We used Illumina sequencing with PCR-free library preparation and k-Seek to estimate the
371 abundance of 7mer satellites across 12 worldwide strains of *D. virilis*, eight strains of *D. americana*
372 (including four strains that have the X-4 fusion and four that do not), and five strains of *D.*
373 *novamexicana* (Table S1). All sequenced strains were male except a female of the *D. virilis* inbred
374 strain 87 as a comparison. A PCA using only the four most abundant 7mers shows clustering of the
375 three species, but the separation is much more dramatic in the PCA using the 20 most abundant
376 simple satellites (Figure S7). Overall, *D. virilis* had the highest AAACTAC satellite content as well as
377 the highest variation, with *D. americana* intermediate between *D. virilis* and *D. novamexicana*
378 (Figure 4A). Using different normalization procedures including mapping and GC correction (see
379 Materials and Methods), produced the same relative ranking of satellite abundances between
380 species. In all cases, the inbred strain from which the genome sequence was produced had the
381 lowest abundance of AAACTAC. In the case of *D. virilis*, this difference was very high. This was not
382 due to a normalization bias as we did mapping-free normalization.

383 Satellite abundances in *D. virilis* displayed a pattern that appeared to be correlated to the
384 geographic location from which strains were collected. For the centromeric satellite AAATTAC,
385 there was a linear decrease in abundance from West to East then South following probable
386 migration from Beringia (Throckmorton 1982) beginning in China (Figure 4C). For the centromeric
387 satellite AAACTAT, the pattern was the opposite; a linear increase in abundance from West to East
388 then South (Figure 4D). We also analyzed sequence variation in the satellite repeats across strains
389 and species to determine if there were any interesting patterns. On average, the centromeric
390 satellite arrays were very homogeneous (average above 99% sequence identity in Illumina reads).
391 However, AAACTAT had a slightly higher sequence identity than AAATTAC (Figure 4 C and D). There
392 was no pattern in average sequence identity with respect to geography for any of the satellites
393 (Figure 4 C and D). The pericentromeric satellite AAACTAC has almost identical sequence divergence
394 across the three species (~98.5%). When comparing between a male and female of the same strain,
395 the male had lower sequence identity. In *D. americana* and *D. novamexicana*, the AAACAAC satellite
396 had lower average sequence identity at 97.5%.

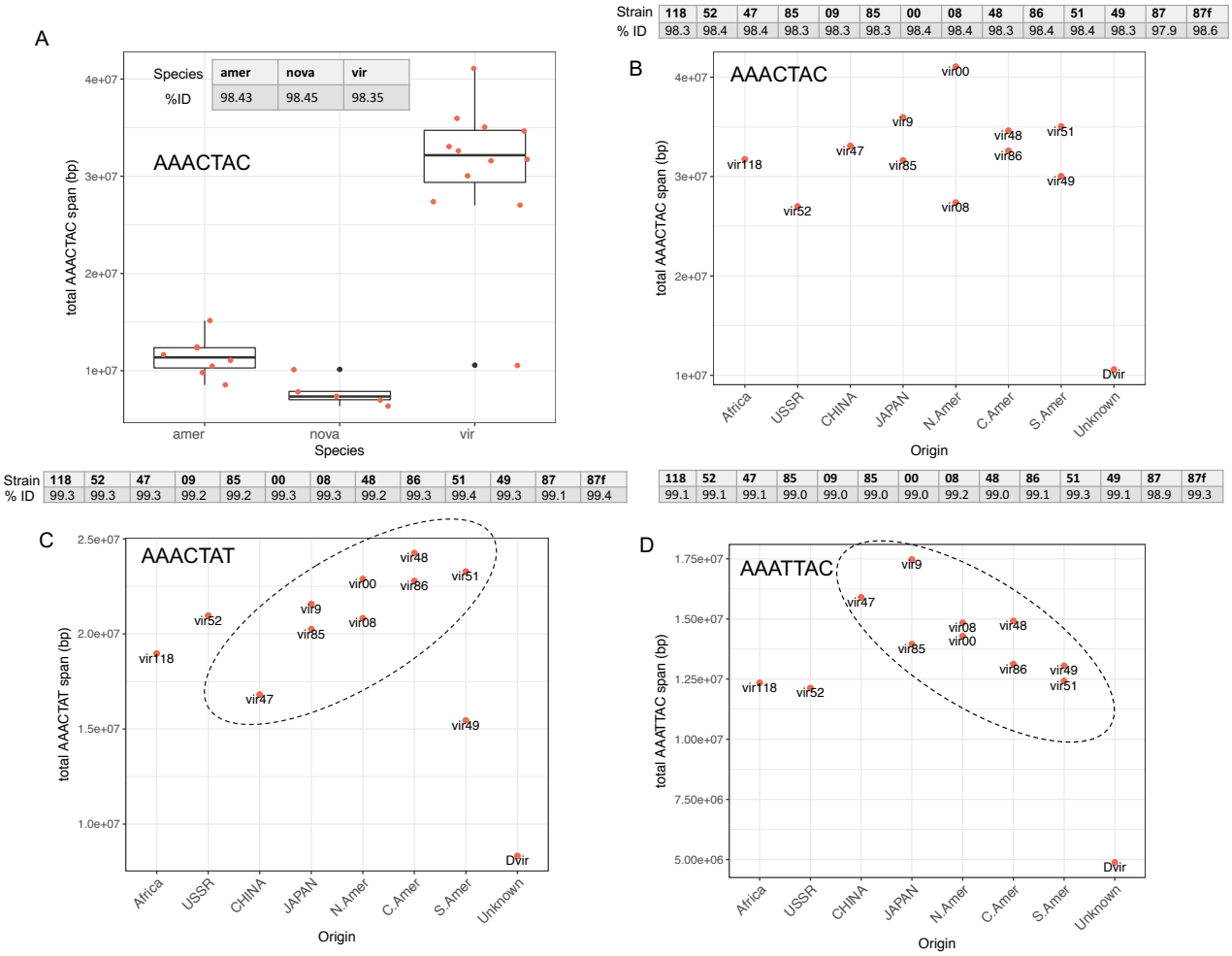


**Fig 4**. Variation in satellites across species and strains. (**A**) AAACTAC total abundance across the three species. (**B**) AAACTAC, (**C**) AAACTAC, and (**D**) AAATTAC, abundance across *D. virilis* strains originating from different localities (x axis). The strain Dvir is strain 87, the inbred strain used for genome assemblies.

397 There were also several other satellites besides the most abundant four that varied

398 between the *D. virilis* group species. AAACAAT was either absent or in low abundance in *D. virilis*

399 and *novamexicana*, but present between 100,000 - 250,000 copies in *D. americana,* indicating a

400 very recent expansion of this satellite. AAAAAC was present in ~100,000 copies in *D. americana* and

401 *D. novamexicana*, whereas it is almost absent in *D. virilis*. To differentiate Y-specific satellites, we

402 included a female of strain 87, the reference genome strain, in our Illumina sequencing run in

403 addition to a male. There were several Y-specific satellites in *D. virilis* found by Wei et al. 2018 and

404 validated by our data, which varied between species in the group (AATAATAG, AATAGATT, and
405 ACATAT). All had different patterns of relative abundance between species
406 (https://github.com/jmf422/D_virilis_satellites/blob/master/Intra_inter_species_sequencing/virilis
407 _group_intra-inter_species.html). There was no detectable difference in centromeric or
408 pericentromeric satellite abundance in *D. americana* strains with vs. without the polymorphic
409 centromere-centromere fusion. We conclude that molecular events surrounding the fusion did not
410 produce any changes in satellite abundance (Figure S8).
411
412 Amplified repeats of the DINE-1 helitron have been found on the *D. virilis* 5th and Y
413 chromosomes (Dias *et al.* 2015). We also examined variation in this satellite abundance by counting
414 reads that mapped to this family of satellites. There was no striking pattern with respect to
415 geography, however the strain with the highest AAACTAC content had the lowest DINE-1 content,
416 besides the inbred strain 87, which had lower satellite content all-around (Figure S9). We expect
417 AAACTAC and DINE-1 to be both in the pericentromeric region of Chr5 in *D. virilis* based on our FISH
418 results and the previous work. The male contained 2.6x more DINEs than the female of the same
419 strain, indicating that ~70% of the DINE-1 repeats in the genome are located on the Y chromosome
420 in *D. virilis*.
421
422 **Discussion**
423
424 Here, we used the satellite DNA-rich genome of *D. virilis* to highlight three previously
425 uncharacterized mechanisms for biases that occur in sequencing and analyzing satellite DNA. We
426 emphasize that comparing satellite DNA amounts between different platforms (e.g. Illumina,

427 PacBio, Nanopore, and even different versions of each) should be done with caution as each
428 technology has its own biases. We have found that issues arise when long arrays of simple satellite
429 DNA are attempted to be sequenced by long-read platforms. In the case of PacBio, systematic
430 errors in base calling may be introduced when sequencing through long arrays of satellites. This
431 issue is not specific to our satellites, as a recent study has also found systematic errors and strand
432 biases in shorter arrays of human satellites in both PacBio and Nanopore reads (Mitsuhashi *et al.*
433 2019). Circular consensus sequencing (CCS) or "HiFi", a type of sequencing offered by PacBio which
434 allows an accurate consensus to be produced after multiple rounds of sequencing the same
435 molecule, may be more appropriate for sequencing analysis of satellite DNA. No systematic errors
436 in satellite sequences resulted with the new CCS platform after collaboration with PacBio
437 representatives. In the case of Nanopore, it is possible that a similar satellite-specific base calling
438 errors exists, or that there is a strand-specific difference in secondary or tertiary structures that
439 occur in long strands of simple satellite DNA. We caution readers in interpreting simple satellite
440 DNA from long-read sequencing data and suggest validation with satellites of known sequence and
441 abundance if available or Illumina reads (without quality filtering). Long read platforms are already
442 improving their chemistry and software for better satellite characterization. Because long reads are
443 likely to cross boundaries of different repetitive regions, long read sequencing proved useful in
444 understanding the length of the satellite arrays and TE insertions into them. Moreover, we
445 demonstrate that the abundance of satellites in pericentromeric heterochromatin are
446 underestimated when sequencing flies compared to pure diploid tissue because of polyteny. We
447 caution readers in performing quality filtering before simple satellite analysis, as satellite containing
448 reads may be enriched for lower quality scores.

449 From comparative analysis of satellites, we found the abundant AAACTAC family of satellites
450 arose in the branch leading to the *virilis* phylad 4.5-11 MYA (Figure 2A). Interestingly, the most
451 abundant satellite in *D. montana*, 7-11 MY diverged, is AAAC. The AAACTAC and AAAC satellites
452 were likely derived from a common ancestor satellite (Fig 2A). From both FISH and sequencing
453 analysis, we found that *D. virilis* has the highest total amount of AAACTAC family satellites, *D.*
454 *novamexicana* has about half of *D. virilis*, and *D. americana* intermediate between the two species.
455 *D. lummei* has the lowest relative satellite content, and its only high-abundance simple satellite is
456 AAACTAC. Unlike the pericentromeric satellite, the centromere-proximal satellite sequence has
457 turned over between *D. virilis* and *D. americana/novamexicana*. The AAACAAC satellite likely
458 evolved in the branch leading to the virilis phylad since it is present in three of the four species
459 studied (Fig 2A). AAACAAC is present in *D. virilis* in relatively low amounts, whereas it became the
460 centromeric satellite on almost all chromosomes in *D. americana* and *novamexicana*. The AAACTAT
461 and AAATTAC satellites are unique to *D. virilis* and occupy the centromeric region. The emerging
462 pattern is that the centromere-proximal satellites have turned over more rapidly than the
463 pericentromeric satellite. This is likely due to satellites participating in conflicts at centromeres
464 (Bayes and Malik 2006, and discussed below). Although sequencing quantified only up to 30 Mb of
465 the AAACTAC family of satellites, FISH confirmed that these satellites are extremely abundant in *D.*
466 *virilis* and the 40% of the genome estimate seems realistic.
467 We can make hypotheses about how and why the satellites expanded in *D. virilis*. We know
468 that mutation rates for changes in copy number of satellite DNA are high, and potentially have a
469 tendency to expand rather than contract in the absence of selection (Flynn *et al.* 2017, 2018). High
470 rates of mutation must be accompanied by a regime that would allow a satellite copy number
471 increase to sweep the population - which could be mediated by positive selection if there is a

472 benefit of the satellite increase, or centromere drive if the phenomenon is at play. Alternatively, in
473 a situation where satellites are slightly deleterious, small effective population sizes in isolated
474 populations or continued bottlenecks could allow satellites to expand in the genome without being
475 removed by selection. However, *D. novamexicana* has the lowest effective population size of the
476 virilis phylad and yet it has the lowest amount of satellite DNA. We already know that the
477 centromere-to-centromere fusions in *D. americana* have undergone meiotic drive hypothesized to
478 be mediated by the increase in centromere total size with the fusion (Stewart et al. 2019). The
479 mechanism allowing drive in *D. americana* may have been at play in the branch leading to *D. virilis*
480 or may be currently occurring. Why have satellites not expanded to this extent in the other species?
481 *D. virilis* might have some attributes about its biology that made the satellite expansion favorable or
482 allowable. For example, genome size is positively correlated with development time in
483 Drosophilidae (Gregory and Johnston 2008). *D. virilis* has a slow development time, and this may
484 have evolved in concert with the expansion in satellite abundance in its genome.
485 We can use data from multiple strains to make hypotheses about factors driving satellite
486 DNA evolution in *D. virilis*. Ancestrally, *D. virilis* had a relatively small effective population size in an
487 isolated range in Asia, and has undergone a recent population and range expansion (Mirol et al.
488 2008). The amount of the most abundant pericentromeric satellite AAACTAC, does not show a
489 geographical pattern across the global strains. Assuming we sequenced a strain from the ancestral
490 range, this suggests that population bottlenecks were not what allowed AAACTAC to expand, and
491 the satellite expansion likely occurred before the population expansion.
492 Our observation of rapid evolution and enrichment of AAACTAC in *D. virilis* in a short
493 evolutionary time period (a few million years) is consistent with the centromere-drive model to
494 account for the evolution of centromere complexity in genetic conflict (Malik and Bayes, 2006). In

495 this model, the asymmetric female meiosis can cause competition between the centromeres with
496 or without newly formed satellites or with more or less satellites, to be included into the oocyte to
497 pass to next generation. A consequence of the competition would be runway expansions of
498 centromeric satellites, and rapid replacements by novel satellites. We hypothesize that the pattern
499 of the centromere-proximal satellite AAACTAT increasing on a geographical gradient while AAATTAC
500 decreases along the same gradient is driven by centromeric conflicts. AAACTAT may be starting to
501 occupy centromeres that AAATTAC occupied, benefitting from a transmission advantage
502 (centromere drive), while the AAATTAC satellite may be decreasing in parallel because of selection
503 "pushing back", for example because of a maximum limit on satellite amount in the centromeric
504 region. Another line of evidence that centromere related conflicts are playing a role is the rapid rate
505 of turnover of the centromere-proximal satellites compared to the pericentromeric satellite.
506 Interestingly, in *D. novamexicana*, AAACTAC was greatly reduced in the pericentromeric
507 regions on all chromosome pairs except one. Based on the FISH images in *D. novamexicana* and
508 *americana*, we hypothesize that it is the 5th chromosome that has the high amount of AAACTAC
509 satellite. This is interesting because previous work has shown that the 5th chromosome contains a
510 high amount of DINE-1 helitron satellite in *D. virilis* but not in *D. americana* (Dias et al. 2015). This
511 may be evidence of past and ongoing competition and trade-offs between the DINE-1 satellite and
512 AAACTAC. We found that all chromosomes including Chr5 contain a large amount of AAACTAC in *D.*
513 *virilis*. DINE-1 had a relatively consistent amount across different *D. virilis* strains, however the
514 strain with the highest AAACTAC amount is an outlier with a lower DINE-1 amount (Figure S9). This
515 may indicate a maximum threshold of satellites was reached on this chromosome, and one satellite
516 had to reduce its abundance. We have seen evidence for this trade-off, or appearance of
517 competitive exclusion, being invoked under selection in our previous studies (Flynn *et al.* 2017,

518 2018). There may have been a similar conflict on Chr5 of *D. novamexicana*, where AAACTAC
519 retained a high copy number to prevent DINE-1 from expanding. Interestingly, the opposite has
520 occurred on the *D. novamexicana* and *D. americana* Y chromosome, where AAACTAC family
521 satellites are absent but DINE repeats are abundant. A potential mechanism mediating apparent
522 stabilizing selection on total satellite abundance is that satellites can act as a sink for
523 heterochromatin factors, with their abundance affecting chromatin state (Lemos *et al*. 2010).
524 The AAACTAC satellite has remained conserved in sequence and location in the virilis
525 phylad. It has also maintained high levels of sequence identity that is equal in the three species we
526 sequenced (98.5% based on Illumina reads). The conservation may reflect a constraint due to
527 selection or a pervasive mechanism of concerted evolution. The periodicity of the sequence may
528 stabilize the DNA helix wrapping around nucleosomes, or it may be constrained by coevolution of
529 an important satellite DNA binding protein (Maio *et al.* 1977; Jagannathan *et al.* 2018). Additionally,
530 within the AAACTAC family, the position and identity of the four A-nucleotides are conserved in all
531 four satellites (AAACTAC, AAATTAC, AAACTAT, AAACAAC) - which may indicate constraint based on
532 the above mechanisms. Conservation of particular satellite unit lengths and "AA" periodicities have
533 been found in other divergent species (Lowman and Bina 1990). Concerted evolution of satellites
534 could be achieved by repeated recycling of units by copy number changes associated with
535 replication slippage or unequal recombination or gene conversion (Walsh 1987; Elder and Turner
536 1995). However, recombination in the pericentromeric heterochromatin has never been detected
537 in wild-type flies (Mehrotra and McKim 2006; Hughes *et al.* 2018). On the other hand, if
538 recombination were occurring, satellite arrays will eventually be lost unless they are conserved by
539 selection (Charlesworth *et al.* 1986). Clearly, we are still lacking in understanding how and why long
540 simple satellite arrays maintain their homogeneity, and whether recombination plays a role in their

541 dynamics. Concordant with the hypothesis that recombination is playing a role, males have lower
542 average sequence identity in the 7 bp satellites than females, which could indicate increased decay
543 on the Y chromosome where there no homologous recombination (Figure 4A).

544 Moreover, our results suggest that transposable elements flank the AAACTAC satellite array
545 (likely more distally to the centromere) and some TEs have inserted within the array. Our analysis
546 suggests that most TE insertions are recent, but because of the under-representation of satellite
547 containing reads, we cannot estimate the number of TE insertions that have occurred into the
548 satellite arrays. Our analysis is also not precise enough to determine if there are islands of TEs at
549 the centromere in *D. virilis* as has been demonstrated in *D. melanogaster* (Chang et al. 2019).

550 We found no difference in centromeric and pericentromeric satellites abundances between
551 *D. americana* strains that differ in their X-4 fusion status. This suggests that the fusion event did not
552 result in a large loss of satellites, making the total centromere and pericentromere size is indeed
553 larger on the X-4 fused chromosome than the single unfused chromosome, concordant with the
554 hypothesis that a larger centromeric region results in centromere drive (Stewart et al. 2019).
555 Another interesting observation from the sequencing of multiple strains of the three species was
556 that in all cases, the inbred strain that the reference genome was made from had the lowest
557 amount of AAACTAC. For *D. virilis*, this difference was extreme. It is tempting to speculate that the
558 process of inbreeding and/or long periods in the lab may have driven the reduction in
559 pericentromeric satellite abundance.

560 In conclusion, our results show very rapid dynamics in the abundant satellites of the *D. virilis*
561 group that are likely explained by various cellular and population-level forces that are not yet
562 understood. Further studies can test if there is a species-specific upper limit to satellite amount per
563 genome or per chromosome upon which negative fitness effects occur, which may result in trade-

564 offs or competition between satellites. Centromere drive may be an important process affecting
565 satellite evolution in this species group, and might partially explain why the satellites expanded 4.5-
566 11 MYA, why satellite sequences at the centromere turned over more rapidly, and why there is a
567 gradient of increasing satellite content related to geographical distribution of strains. A more
568 extensive study to determine if inbreeding or extended periods in the lab drives a reduction in
569 satellite abundance will help illuminate the processes that are important for maintaining satellite
570 content. Determining the frequency of recombination in the large pericentromeric heterochromatin
571 blocks in species like *D. virilis* will be challenging but important for understanding how the satellites
572 maintain homogeneity in their sequence. To understand the role of satellites and the importance of
573 their sequence, unit length, and abundance, researchers can strive to develop methods to engineer
574 satellites by modifying specific bases and their abundances.
575
576 **MATERIALS AND METHODS**
577
578 All scripts for analyzing the data and to produce the results we show are here:
579 https://github.com/jmf422/D_virilis_satellites. Illumina sequencing reads generated for this study
580 are deposited in NCBI SRA under accession PRJNA548201. Raw PacBio reads will be deposited under
581 the same accession. Both will be released upon publication.
582
583 **Characterizing satellite DNA from genome assemblies**
584
585 All scripts and R markdown files used for this analysis are provided in
586 https://github.com/jmf422/D_virilis_satellites/tree/master/Genome_assembly_analysis.

587 We used genome assemblies produced by the PacBio sequencing project

588 (https://www.ncbi.nlm.nih.gov/bioproject/?term=txid7214[Organism:noexp]) of *D. virilis*, *D.*

589 *novamexicana*, and *D. americana*. We also downloaded the *D. virilis* genome produced by

590 Nanopore sequencing from (Miller *et al.* 2018), and the CAF1 assembly from (Drosophila 12

591 Genomes Consortium *et al.* 2007). We used Phobos (https://www.ruhr-uni-

592 bochum.de/spezzoo/cm/cm_phobos.htm) and Tandem repeats finder (Benson 1999) to

593 characterize simple and complex satellites in these genome assemblies. To identify the

594 chromosomal linkage of complex satellites in the genome assembly, we produced a dotplot with D-

595 GENIES (Cabanettes and Klopp 2018).

596

597 **Characterizing satellite DNA from raw long reads**

598

599 Characterizing and quantifying satellites from long reads is a challenge because of the sequencing

600 high error rate. We used two approaches to characterize satellites from raw long reads. The first

601 approach, we call k-Seek + Phobos, in which we first broke the reads into 100 bp subreads and ran

602 k-Seek on them. k-Seek is very efficient for analyzing many reads, however is not very sensitive for

603 reads with a high error rate since it was designed for Illumina reads (Wei *et al.* 2014). If k-Seek

604 found satellites on at least one subread, we would run the complete parent read through Phobos.

605 Phobos is more sensitive to imperfect repeats and error rates, but cannot handle huge quantities of

606 data; thus why we only ran the portion of reads identified by k-Seek to have tandem repeats. This

607 approach allowed us to characterize satellites *de novo* and quantify them. All scripts for the analysis

608 of long reads with the k-Seek + Phobos approach are located here:

609 https://github.com/jmf422/D_virilis_satellites/tree/master/LongRead_kseek_Phobos. The second

610 approach we used is Noise-Cancelling Repeat Finder (NCRF, (Harris *et al.* 2019) ). This program was
611 designed to quantify satellites from long reads with high error rates. However, it cannot identify
612 satellites *de novo* and requires specific satellite sequences to search for. NCRF also requires a "max
613 divergence allowed" parameter, which we tuned with simulations (see below). Scripts used for the
614 NCRF approach are located here:
615 https://github.com/jmf422/D_virilis_satellites/tree/master/LongRead_NCRF.
616
617 We did simulations to assess both approaches:
618 https://github.com/jmf422/D_virilis_satellites/tree/master/Simulations. First, we created a
619 simplified mock *D. virilis* genome with a satellite DNA composition estimated from our FISH results.
620 We could not use the genome assembly because it contained very little satellite DNA. Specifically,
621 each chromosome had a centromeric satellite either AAATTAC or AAACTAC followed by the
622 pericentromeric satellite AAACTAC, combined taking up 40% of the genome. The non-satellite DNA
623 portion of the genome was generated randomly with a 40% GC content. We then used PBSim (Ono
624 *et al.* 2013) to simulate PacBio reads and we used these simulated reads for multiple analyses. First,
625 we used them to tune the max divergence parameter of NCRF by running NCRF repeatedly with
626 max divergence parameters ranging 18-30%. We found that the amount of satellites found,
627 particularly the most abundant one, levelled off at 25% max divergence. This is the parameter value
628 we used moving forward. We also used these simulated reads to quantify satellites with both
629 approaches and compare them. Finally, we used these simulated reads to assess strand biases in
630 long read sequencing data (see below).
631
632 **Identification of biases in simple satellites in long read data**

633

634 We suspected that there were biases in the satellite DNA found in the *D. virilis* group PacBio (and
635 Nanopore) data because we found high abundance satellites that had never been found before with
636 other types of data, and so we suspected they were artifactual. These artifactual satellites were
637 found with both kSeek + Phobos and NCRF approaches, but were not found in the simulated data.
638 We tried testing for a strand bias in reads that contained satellite DNA. Using both the summarized
639 output from NCRF and validated with a custom script
640 (LongRead_NCRF/which_strand_pacbio_script.sh), we counted the satellite DNA stretches that
641 originated from each the positive and negative strand. The positive strand is defined as the one that
642 contains the satellite AAACTAC and derivatives (more As than Ts), and the negative strand is the
643 one that contains the reverse complement (e.g. GTAGTTT, more Ts than As). We did this for the
644 three satellites used in the simulated data and real and artefactual satellites found in the PacBio
645 and Nanopore data. Detailed analysis and visualization of the biases is shown here:
646 LongRead_kseek_Phobos/longread_analysis.html
647

648 **Sequencing of polytene and non-polytene tissue**

649

650 To acquire *D. virilis* pure diploid tissue, we dissected male 3$^{rd}$ instar larvae and collected imaginal
651 discs including the eye-antennal disc and wing discs. Approximately 100 larvae were required to get
652 enough DNA (>1 ug). We also collected ~5 adult flies for fly libraries. We used the inbred genome
653 assembly strain 87 for these libraries. DNA was extracted with Qiagen DNeasy blood and tissue kit
654 and PCR-free libraries were prepared. Libraries were run on an Illumina NextSeq with 1 x 150 bp
655 reads, and each sample took up approximately 7% of the flowcell. The other libraries run on this

656 flowcell were from an unrelated project including RNAseq from other species. Reads were analyzed
657 with k-Seek both before and after filtering with Trimmomatic (Bolger et al. 2014). FastQC was run to
658 evaluate the quality of the reads. Scripts are here:
659 https://github.com/jmf422/D_virilis_satellites/tree/master/Polyteny. We also analyzed publicly
660 available *D. melanogaster* data from the same strain and same sequencing platform of embryos
661 (non-polytene), salivary glands (extreme polyteny) from (Yarosh and Spradling 2014), and flies
662 (varied levels of polyteny) from (Gutzwiller *et al.* 2015).
663
664 **Fluorescence in situ hybridization of satellite DNAs**
665
666 We followed the protocol of (Larracuente and Ferree 2015) for satellite DNA FISH. We ordered the
667 following probes from IDT with 5' modifications: $(AAACTAC)_6$ with alexa-488 fluorophore,
668 $(AAACTAT)_6$ with Cyanine5 fluorophore, $(AAATTAC)_6$ with Cyanine3 fluorophore, $(AAACAAC)_6$ with
669 Cyanine3 fluorophore, and $(AAACGAC)_6$ with Cyanine5 fluorophore. We hybridized three probes at
670 a time, to allow for similar probes to compete to result in specific hybridization with the rationale
671 shown in (Beliveau *et al.* 2015). Hybridization temperature was 32°C. We imaged on an Olympus
672 fluorescent microscope and Metamorph capture system at the Cornell Imaging Facility. Composite
673 images were produced with ImageJ.
674
675 **Characterizing TEs linked to satellites and satellites anchored to the genome assembly**
676
677 We extracted the reads identified to have the 7 bp satellites on them from NCRF results, and then
678 we ran RepeatMasker (http://www.repeatmasker.org) on these reads using parameters: "-nolow"

679 and "-species Drosophila". All reads had at least 500 bp of tandem satellite on them according to
680 NCRF default parameters, and to avoid spurious identification of TEs from semi-repetitive
681 fragments, we described only TEs in reads that also had at least 500 bp of a TE identified from
682 RepeatMasker. We also BLASTed the same satellite reads to the genome assembly to evaluate if
683 satellite reads could be anchored to the genome assembly. Analysis scripts are here:
684 https://github.com/jmf422/D_virilis_satellites/tree/master/TEs_satellites.
685
686 **Sequencing of multiple *D. virilis* group strains**
687
688 We obtained as many strains of *D. virilis* that have information about where they were collected as
689 possible. This included 12 strains as live stocks we obtained either from stocks in our lab or from the
690 Drosophila species stock center (Table S1). We also prepared a female library for strain 87 for when
691 we wanted to differentiate Y-specific satellite patterns. We also obtained five strains of *D.*
692 *novamexicana* and eight strains of *D. americana*. All were obtained from live stocks and the inbred
693 genome strains were included for both species as well (strain 14 and G96, respectfully). For *D.*
694 *americana*, we included four strains that have the chromosome X-4 fusion and four strains that do
695 not have it, based on communication with the Bryant McAllister lab. DNA was extracted as above
696 from five flies each and samples were prepared identically as above and sequenced on 50% of 3
697 flowcells of Illumina NextSeq 1 x 150 bp reads. We dispersed the samples from each species
698 between multiple flowcells. Our samples took up only half the flowcell with the other half being
699 occupied by a RNAseq libraries from an unrelated project.
700 All scripts used to analyze these data are located here:
701 https://github.com/jmf422/D_virilis_satellites/tree/master/Intra_inter_species_sequencing.

Reads were evaluated with FastQC and not filtered for quality based on the potential bias of Illumina quality scores on satellites. We used k-Seek to quantify satellites. We tried several normalization strategies but decided the most appropriate was a mapping-free normalization. We estimated average depth by dividing the total number of bases sequenced by the estimated genome size by flow cytometry (Bosco *et al.* 2007). We believe this was the best option in this case because: 1) we were concerned about a mapping bias because for each species the strain that the genome assembly was made from may have more reads map to it; 2) after masking the genome from the 7mer satellites and also excluding the X and Y contigs (because we had male and female strains, and the Y chromosome contained more low GC regions) - there was little difference in coverage based on GC content. We include results when we used a mapping based GC normalization in the sub-directory "AlternativeNormalization".

We used NCRF with modified parameters (minlength=100, maxdiv=10) to characterize the average sequence identity of satellite arrays from the Illumina data. To quantify DINE-1 satellites across *D. virilis* strains, we produced a library of DINE-1 satellite variants based on our PacBio genome analysis. We then mapped Illumina reads to this library and normalized the number of reads that mapped to any sequence in the library by the estimated depth. We also analyzed Illumina DNA sequencing reads of *D. montana* (Parker et al. 2018) and *D. lummei* (Ahmed-Braimah *et al.* 2017) with k-Seek to identify the most abundant satellites and whether or not the AAACTAC satellite family was present.

**Acknowledgments**

724 We thank Yasir Ahmed-Braimah for helpful discussions and advice for some analyses. We also thank
725 Bryant McAllister for providing *D. americana* strains along with their fusion status. We are grateful
726 for Elissa Cosgrove's help with some computational trouble-shooting and Asha Jain's help in
727 preparing DNA sequencing libraries. We also thank Danny Miller for useful discussions and for
728 providing the raw Nanopore reads from his study. Sarah Kingan, Jane Landolin, and Greg Young
729 from Pacific Biosciences were very helpful in exploring the causes for the artefactual repeats and in
730 producing the HiFi data. We thank Amanda Larracuente for advice on FISH protocols and the Cornell
731 Imaging Facility for use of their microscope. This project was funded by NIH grant number:
732 GM116113 to RAW, ML, and AGC and GM119125 to AGC and Daniel Barbash. JMF was supported
733 by an NSERC PGS D fellowship.
734
735
736 **References**

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

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

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

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

746 Research 27: 573–580.

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

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

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

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

756 Casacuberta E., and M.-L. Pardue, 2003 HeT-A elements in *Drosophila virilis*: retrotransposon
757 telomeres are conserved across the Drosophila genus. Proc. Natl. Acad. Sci. U. S. A. 100:
758 14091–14096.

759 Chang C.-H., and A. M. Larracuente, 2019 Heterochromatin-Enriched Assemblies Reveal the
760 Sequence and Organization of the *Drosophila melanogaster* Y Chromosome. Genetics 211:
761 333–348.

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

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

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

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

771 Drosophila 12 Genomes Consortium, A. G. Clark, M. B. Eisen, D. R. Smith, C. M. Bergman, *et al.*,
772 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.

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

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

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

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

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

784 Gall J. G., and D. D. Atherton, 1974 Satellite DNA sequences in Drosophila virilis. J. Mol. Biol. 85:
785 633–664.

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

788 Gregory T. R., and J. S. Johnston, 2008 Genome size diversity in the family Drosophilidae. Heredity
789 101: 228–238.

790 Gutzwiller F., C. R. Carmo, D. E. Miller, D. W. Rice, I. L. G. Newton, *et al.*, 2015 Dynamics of
791 *Wolbachia pipientis* Gene Expression Across the Drosophila melanogaster Life Cycle. G3 5:
792 2843–2856.

793 Harris R. S., M. Cechova, K.D. Makova. 2019 Noise-cancelling repeat finder: uncovering tandem
794 repeats in error-prone long-read sequencing data. Bioinformatics: btz484

795 Heikkinen E., V. Launonen, E. Muller, L. Bachmann 1995 The pvB370 *BamHI* Satellite DNA Family of
796 the *Drosophila virilis* Group and Its Evolutionary Relation to Mobile Dispersed Genetic pDv
797 Elements. J Mol Evol 41:604-614.

798 Henikoff S., K. Ahmad, H.S. Malik. 2001 The Centromere Paradox: Stable Inheritance with Rapidly
799 Evolving DNA. Science 293: 1098–1102.

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

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

805 Jagannathan M., R. Cummings, and Y. M. Yamashita, 2018 A conserved function for pericentromeric

806 satellite DNA. Elife 7. https://doi.org/10.7554/eLife.34122

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

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

812 Larracuente A. M., and P. M. Ferree, 2015 Simple Method for Fluorescence DNA In Situ
813 Hybridization to Squashed Chromosomes. Journal of Visualized Experiments. 95: e52288.
814

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

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

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

821 Malik H.S. and J.J. Bayes, 2006. Genetic conflicts during meiosis and the evolutionary origins of
822 centromere complexity. Biochem Soc Trans. 34: 569-573.
823

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

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

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

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

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

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

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

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

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

845 Pavlek M., Y. Gelfand, M. Plohl, and N. Meštrović, 2015 Genome-wide analysis of tandem repeats in
846 *Tribolium castaneum* genome reveals abundant and highly dynamic tandem repeat families

847 with satellite DNA features in euchromatic chromosomal arms. DNA Res. 22: 387–401.

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

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

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

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

858 Stewart B. S., Y. H. Ahmed-Braimah, D. G. Cerne, B. F. McAllister, Female meiotic drive preferentially
859 segregates derived metacentric chromosomes in Drosophila.
860

861 Throckmorton LH: *The Genetics and Biology of Drosophila Volume 3b*. Academic Press: New York;
862 1982.

863 Walsh J. B., 1987 Persistence of tandem arrays: implications for satellite and simple-sequence
864 DNAs. Genetics 115: 553–567.

865 Wei K. H.-C., J. K. Grenier, D. A. Barbash, and A. G. Clark, 2014 Correlated variation and population
866 differentiation in satellite DNA abundance among lines of *Drosophila melanogaster*. Proc. Natl.

867 Acad. Sci. U. S. A. 111: 18793–18798.

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

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

872 Zelentsova E. S., R. P. Vashakidze, A. S Krayev, M. B. Evgen'ev 1986 Dispersed repeats in *Drosophila*
873 *virilis*: Elements mobilized by interspecific hybridization. Chromosoma 93: 469-476.