

UC Davis

UC Davis Previously Published Works

Title

Highly contiguous genome assembly of *Drosophila prolongata*—a model for evolution of sexual dimorphism and male-specific innovations

Permalink

<https://escholarship.org/uc/item/9nh4x2vd>

Journal

G3: Genes, Genomes, Genetics, 14(10)

ISSN

2160-1836

Authors

Luecke, David

Luo, Yige

Krzystek, Halina

et al.

Publication Date

2024-10-07

DOI

10.1093/g3journal/jkae155

Peer reviewed

1 Highly Contiguous Genome Assembly of *Drosophila prolongata* - a Model for Evolution of Sexual
2 Dimorphism and Male-specific Innovations

3

4 co-1) David Luecke, Department of Evolution and Ecology, University of California Davis, One
5 Shields Ave Davis CA 95616

6 co-1) Yige Luo, Department of Evolution and Ecology, University of California Davis, One Shields
7 Ave Davis CA 95616

8 3) Halina Krzystek, Biology Department of the University of North Carolina (UNC), 3159 Genome
9 Sciences Building. 250 Bell Tower Drive. Chapel Hill, NC 27599.

10 4) Corbin Jones, Biology Department of the University of North Carolina (UNC), 3159 Genome
11 Sciences Building. 250 Bell Tower Drive. Chapel Hill, NC 27599.

12 5) Artyom Kopp, Department of Evolution and Ecology, University of California Davis, One Shields
13 Ave Davis CA 95616

14

15 Corresponding authors: David Luecke david.luecke@gmail.com, Artyom Kopp akopp@ucdavis.edu

16

17 The genome of *Drosophila prolongata*

18

19 Keywords: Genome, *Drosophila*, Sex Dimorphism

20 **Abstract**

21 *Drosophila prolongata* is a member of the *melanogaster* species group and *rhopaloea* subgroup native
22 to the subtropical highlands of southeast Asia. This species exhibits an array of recently evolved male-
23 specific morphological, physiological, and behavioral traits that distinguish it from its closest relatives,
24 making it an attractive model for studying the evolution of sexual dimorphism and testing theories of
25 sexual selection. The lack of genomic resources has impeded the dissection of the molecular basis of
26 sex-specific development and behavior in this species. To address this, we assembled the genome of *D.*
27 *prolongata* using long-read sequencing and Hi-C scaffolding, resulting in a highly complete and
28 contiguous (scaffold N50 2.2Mb) genome assembly of 220Mb. The repetitive content of the genome is
29 24.6%, the plurality of which are LTR retrotransposons (33.2%). Annotations based on RNA-seq data
30 and homology to related species revealed a total of 19,330 genes, of which 16,170 are protein-coding.
31 The assembly includes 98.5% of Diptera BUSCO genes, including 93.8% present as a single copy.
32 Despite some likely regional duplications, the completeness of this genome suggests that it can be
33 readily used for gene expression, GWAS, and other genomic analyses.

34

35 **Introduction**

36 *Drosophila prolongata* is a member of the *melanogaster* species group and *rhopaloea* subgroup
37 native to southeast Asia (Singh and Gupta 1977; Toda 1991). The species has a suite of recently
38 evolved male-specific morphological traits (Figure 1), including increased foreleg size, leg
39 pigmentation, wing pigmentation, reversed sexual size dimorphism, and an expanded number of leg
40 chemosensory organs (Luecke, Rice, and Kopp 2022; Luecke and Kopp 2019; Luo et al. 2019). These
41 traits are associated with derived behaviors, including male-male grappling and male leg vibration
42 courtship displays, along with increased sexual dimorphism in cuticular hydrocarbon profiles (Amino
43 and Matsuo 2023b; 2023a; Kudo et al. 2015; 2017; Luo et al. 2019; Setoguchi et al. 2014; Takau and
44 Matsuo 2022; Toyoshima and Matsuo 2023).

45 The phylogenetic proximity to the model species *D. melanogaster* and available genome
46 sequences for closely related species *D. rhopaloa* and *D. carrolli* (Kim et al. 2021), which lack these
47 derived traits, make this species a promising system to study the genetics of sexually dimorphic
48 development, physiology, and behavior. A reference genome assembly and annotation for *D.*
49 *prolongata* benefits such work as it would provide insight into the genomic evolutionary patterns
50 associated with the evolution of the novel traits in *D. prolongata*. Presented here is a highly complete
51 and contiguous assembly based on long-read Pacific Biosciences sequencing and Hi-C scaffolding,
52 along with annotations for both *D. prolongata* and *D. carrolli* using *D. melanogaster* sequence
53 homology and gene models based on RNA sequencing evidence and ab initio predictions.

54

55 **Materials and Methods**

56 Genome line generation

57 The isofemale SaPa01 line and BaVi44 line were collected in SaPa and BaVi, Vietnam,
58 respectively, by Dr. Hisaki Takamori in September 2004. Virgin females were collected by isolating
59 adults within four hours of emergence. Four generations of full-sib matings were carried out to produce
60 the genomic strain SaPa_ori_Rep25-2-1-1 (“Sapa_PacBio”). Fly strains were maintained at room
61 temperature on standard cornmeal food provided by the UC Davis Fly Kitchen with filter paper for
62 environment structure and pupariation substrate.

63

64 Tissue collection

65 For genome assembly/scaffolding, adult male flies from the genome strain were moved onto
66 nutrient-free agar media for at least one day to reduce microbial load, then collected into 1.5mL tubes
67 and flash-frozen in liquid nitrogen. Fifty frozen adult male individuals were sent on dry ice to Dovetail
68 Genomics (Cantata Bio. LLC, dovetailgenomics.com) for DNA extraction, sequencing, and assembly.
69 For gene expression data used in annotation, whole forelegs were dissected from carbon dioxide

70 anesthetized males and females of the SaPa01 isofemale line, along with dissected heads from each sex
71 of the genome strain.

72

73 Sequencing and assembly

74 All genomic DNA extraction, sequencing, and assembly were carried out by Dovetail Genomics
75 (Cantata Bio LLC, Scotts Valley, CA, USA). Genomic DNA was extracted with the Qiagen HMW
76 genomic extraction kit (Qiagen, Germantown, MD, USA). DNA samples were quantified using a Qubit
77 2.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). The PacBio SMRTbell library (~20kb) for
78 PacBio Sequel was constructed using SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park,
79 CA, USA) using the manufacturer-recommended protocol. The library was bound to polymerase using
80 the Sequel II Binding Kit 2.0 (PacBio) and loaded onto PacBio Sequel II. Sequencing was performed
81 on PacBio Sequel II 8M SMRT cells, generating 16 gigabases of data. An initial assembly based on
82 1.2M PacBio reads was produced using FALCON (Chin et al. 2016) with Arrow polishing.

83 A Dovetail HiC library was prepared similarly as described previously (Lieberman-Aiden et al.
84 2009). Briefly, for each library, chromatin was fixed in place with formaldehyde in the nucleus and
85 then extracted. Fixed chromatin was digested with DpnII, the 5' overhangs filled in with biotinylated
86 nucleotides, and free blunt ends were ligated subsequently. After ligation, crosslinks were reversed, and
87 the DNA was purified from protein. Purified DNA was treated to remove biotin that was not internal to
88 ligated fragments. The DNA was then sheared to ~350 bp mean fragment size, and sequencing libraries
89 were generated using NEBNext Ultra enzymes and Illumina-compatible adapters. Biotin-containing
90 fragments were isolated using streptavidin beads before PCR enrichment of each library. The libraries
91 were sequenced on an Illumina HiSeq X to a target depth of 30x coverage.

92 The input *de novo* assembly and Dovetail HiC library reads were used as input data for HiRise,
93 a software pipeline designed specifically for using proximity ligation data to scaffold genome
94 assemblies (Putnam et al. 2016). Dovetail HiC library sequences were aligned to the draft input

95 assembly using a modified SNAP read mapper (<http://snap.cs.berkeley.edu>). The separations of
96 Dovetail HiC read pairs mapped within draft scaffolds were analyzed by HiRise to produce a likelihood
97 model for genomic distance between read pairs, and the model was used to identify and break putative
98 misjoins, to score prospective joins, and make joins above a threshold. A second HiRise assembly was
99 generated with additional HiC sequencing and the HiRise software pipeline.

100 RNA was extracted using TRIzol (Invitrogen, Waltham, MA, USA). For foreleg RNA,
101 multiplexed stranded cDNA sequencing libraries were prepared using the NEBNext Ultra Directional
102 RNA Library Prep Kit for Illumina (New England BioLabs, Ipswich, MA, USA) using poly(A)
103 isolation magnetic beads. Libraries were sequenced on the Illumina HiSeq4000 platform by the UC
104 Davis Genome Center. For head RNA, cDNA sequencing libraries were constructed using the TruSeq
105 Stranded RNA Kit (Illumina, San Diego, CA) and sequenced on the Illumina HiSeq4000 platform by
106 Novogene (<https://www.novogene.com/us-en/>). Raw RNA-seq reads and assembled genome can be
107 accessed with NCBI BioProject PRJNA1057277. Transcripts were assembled using Trinity 2.4.0 (Haas
108 et al. 2013) with default options for stranded data.

109

110 Gene prediction and annotation

111 Homology-based annotations were generated using Liftoff 1.5.1 (Shumate and Salzberg 2021)
112 with minimap2 2.17 (Li 2018) alignment based on the *D. melanogaster* GCF000001215.4 release 6
113 (Hoskins et al. 2015) *D. elegans* GCF000224195.1 2.0, and *D. rhopaloa* GCF000236305.1 2.0 (Kim et
114 al. 2021) annotations downloaded from FlyBase (Gramates et al. 2022). Liftoff was run with the copies
115 option and percent identity 0.80. Additional gene models were inferred using MAKER 3.01.02 (Holt
116 and Yandell 2011) with BLAST 2.11.0 (Camacho et al. 2009) and repeat masker 4.0.7, using EST
117 evidence from the Trinity transcripts assembled based on foreleg and head RNA and protein homology
118 evidence based on the combined protein sets from the *D. melanogaster* and *D. elegans* annotations also
119 used for Liftoff. The annotations from different sources were then combined using gffcompare 10.4

120 (Pertea and Pertea 2020), genomertools 1.5.9 (Gremme, Steinbiss, and Kurtz 2013), and custom Python
121 3.7.6 scripts available at https://github.com/dluecke/annotation_tools.

122

123 Removal of duplicate scaffolds

124 BUSCO (Manni et al. 2021) analysis of the Dovetail HiRise using the diptera_ ocb10 lineage
125 dataset revealed 200 complete but duplicated benchmark genes (Table S1), indicating potential
126 duplicated regions in the assembly. Scaffolds were assessed for BUSCO benchmark gene content and
127 sorted by the percentage of duplicated BUSCO genes. 53 candidate scaffolds, ranging from 20,819bp
128 to 39,990,007bp, contained at least one duplicated benchmark BUSCO gene (Table S1). Inspection of
129 MUMmer (Marçais et al. 2018) alignments between duplicate-containing candidates and scaffolds with
130 alternate copies of the duplicated benchmark genes showed complete alignment across 27 of the
131 candidate scaffolds (Figure S1). These 27 scaffolds (ranging from 20,819bp to 541,551bp) were
132 considered fully duplicated and split from the assembly and annotation (Files S1, S2, and S3) using
133 SAMtools 1.15.1 (Li et al. 2009). Custom Python pandas 1.1.2 (McKinney 2010), and R 4.0.3 (R Core
134 Team 2020, <https://www.R-project.org>) for scaffold sorting by BUSCO scores, splitting assembly and
135 annotation, and inspecting genome alignments are available at
136 https://github.com/dluecke/annotation_tools.

137

138 Identification of duplicate genes

139 The remaining duplicated genes in the *D. prolongata* deduplicated annotation were identified
140 using reciprocal BLAST. Strand oriented regions corresponding to all “gene” features in both *D.*
141 *prolongata* and *D. rhopaloa* annotations were extracted from their respective assemblies using
142 bedtools 2.29.2 (Quinlan and Hall 2010). *D. prolongata* gene regions were searched against all *D.*
143 *prolongata* and all *D. rhopaloa* gene regions using blastn 2.14.1 (Camacho et al. 2009). BLAST results
144 were combined and sorted by match alignment bit score, then duplicate status was assigned to pairs of

145 *D. prolongata* genes if both regions had higher match scores with the corresponding *D. prolongata*
146 region than to any gene region from *D. rhopaloa*. Custom Bash and Python scripts used in this process
147 are available at https://github.com/dluecke/annotation_tools.

148

149 Repeat analysis

150 Tandem repeats were annotated with Tandem Repeat Finder 4.09.1 (Benson 1999). A *de novo*
151 library of classified repetitive element models was created using RepeatModeler 2.0 (Flynn et al.
152 2020). To reduce the run-to-run variations, repeat classification was based on five independent
153 RepeatModeler runs with the following random seeds: 1681089287, 1687990919, 1683413925,
154 1683532158, and 1683532058. Custom R and Bash scripts are available at [https://github.com/yige-](https://github.com/yige-luo/Repeat_analysis)
155 [luo/Repeat_analysis](https://github.com/yige-luo/Repeat_analysis).

156

157 Assembly and annotation evaluation

158 Assembly contiguity statistics were provided by Dovetail. Reference annotations *D.*
159 *melanogaster* GCF_000001215.4 and *D. rhopaloa* GCF_018152115.1 were downloaded from the
160 NCBI genomes database. Assembly completeness was assessed with BUSCO 5.3.2 (Manni et al. 2021)
161 using the diptera_ocb10 lineage dataset, HMMER 3.1b2, and Mmseqs 5.34c21f2. Whole genome
162 alignment between *D. prolongata* and *D. rhopaloa* assemblies was performed with MUMmer 4.0.0
163 (Marçais et al. 2018) using nucmer alignment with a minimum exact match 1000bp for alignment with
164 *D. rhopaloa* and 500bp for *D. melanogaster* alignment, and mummerplot plus custom Bash and R
165 scripts (https://github.com/dluecke/annotation_tools) for visualization. Annotation statistics were found
166 with genomertools 1.5.9 (Gremme, Steinbiss, and Kurtz 2013). Transcripts were extracted from
167 annotations using gffread 0.9.12 (Pertea and Pertea 2020), and transcript completeness was assessed
168 using the transcriptome mode of BUSCO.

169

170 **Results and Discussion**

171 Assembly contiguity

172 The Dovetail HiRise assembly scaffolding method (Figure 2) produced an assembly for *D.*
173 *prolongata* with higher contiguity than the existing *D. rhopaloa* and *D. carrolli* assemblies,
174 approaching the contiguity of the latest *D. melanogaster* reference (Table 1) as measured by N50.
175 Whole genome alignments of the *D. prolongata* assembly to *D. rhopaloa* and *D. melanogaster*
176 references (Figure 3A) show long stretches of high identity with *D. rhopaloa* spanning nearly all large
177 scaffolds.

178

Assembly	<i>D. prolongata</i>	<i>D. carrolli</i>	<i>D. rhopaloa</i>	<i>D. melanogaster</i>
Total length (bp)	220759777	231219246	193508231	143726002
Scaffolds	387	338	228	1870
N50 (bp)	22190323	14004682	15806012	25286936
L50	4	5	5	3
GC%	40.11%	39.52%	39.87%	41.67%
BUSCO Complete, Single Copy	93.7% (3078)	97.8% (3214)	98.1% (3221)	98.5% (3235)
BUSCO Complete, Duplicated	4.8% (158)	0.4% (13)	0.4% (12)	0.2% (8)
BUSCO Fragmented	0.9% (29)	0.6% (19)	0.7% (24)	0.5% (16)
BUSCO Missing	0.6% (20)	1.2% (39)	0.8% (28)	0.8% (26)

179 Table 1: Statistics for assembly contiguity and completeness of *D. prolongata* assembly alongside
180 previously published *D. carrolli* GCA_018152295.1 assembly (Kim et al. 2021), reference assemblies
181 *D. rhopaloa* GCF_018152115.1 and *D. melanogaster* GCF_000001215.4. BUSCO statistics are for the
182 3285 genes in the diptera_odb10 benchmark set.

183

184 Assembly completeness

185 BUSCO results for assemblies (Table 1) show a comparable degree of completeness for the
186 3285 genes in the BUSCO dipteran benchmark set between *D. prolongata* assembly and references,
187 with 3236 complete for *D. prolongata*, 3233 complete for *D. rhopaloa*, and 3243 complete for *D.*

188 *melanogaster*. The whole genome alignments between the *D. prolongata* assembly and the *D. rhopaloa*
189 (Figure 3A) and *D. melanogaster* references (Figure 3B) further show near complete highly contiguous
190 coverage of the entire reference with regions of *D. prolongata* scaffolds, corresponding to all five
191 major chromosome arms in the *D. melanogaster* genome.

192

193 Repeat annotation

194 The *D. prolongata* genome exhibits a moderate level of repeat content (24.6%) comparable to
195 the other species (Figure 4). The vast majority (37/40) of classified repeat families are not specific to
196 *D. prolongata*, except for two Long Interspersed Nuclear Element (LINE) retrotransposons, RTE-BovB
197 and L1, and one DNA transposon, Crypton-V (Table S2). We note, however, that further evidence is
198 required to test whether these repeat families have evolved in *D. prolongata*, as all of them have only
199 one identified member in one out of five RepeatModeler runs. Among the repetitive elements of *D.*
200 *prolongata*, the most prominent repeat classes are Long Terminal Repeats retrotransposons (LTR,
201 32.2%), LINE (15.1%) and Tandem Repeats (14.6%, Table 2). A breakdown of repeat content by
202 scaffolds across four species can be found in Table S3.

203 Compared with most long (>1Mb) scaffolds, intermediate-sized scaffolds in *D. prolongata*
204 assembly tend to show higher repeat content (Figure S2, Figure S3). Exceptions are found in scaffolds
205 414, scaffold 293, scaffold 164 and scaffold 280 (Figure S2), where LTR and LINE are
206 overrepresented, reminiscent of the repeat profiles of several primary scaffolds in closely related
207 species *D. carrolli* and *D. rhopaloa* (Figure S4, Figure S5), as well as the Y chromosome in *D.*
208 *melanogaster* (Figure S6).

209

Repeat Class	<i>D. prolongata</i> (%)	<i>D. carrolli</i> (%)	<i>D. rhopaloa</i> (%)	<i>D. melanogaster</i> (%)
Tandem Repeat	3.627	12.003	6.601	2.421
Simple	0.008	0.007	0.008	0.007
Satellite	0.019	0.017	0.012	0.031

DNA	1.067	1.224	0.971	0.877
RC	1.595	1.122	1.274	0.218
LINE	3.727	3.626	3.612	3.526
LTR	7.939	7.505	6.141	8.525
rRNA	0.061	0.014	0.000	0.040
snRNA	0.000	0.001	0.004	0.000
tRNA	0.005	0.001	0.004	0.005
Unknown	3.276	2.686	2.519	0.575
Multiclass	3.311	3.926	3.260	1.896
Total	24.636	32.131	24.406	18.121

210 Table 2: Repeat content of genome assemblies of *D. prolongata* and three reference species.

211

212 Annotation completeness

213 Transcripts extracted from the annotation and assembly show that the *D. prolongata* and *D.*
214 *carrolli* annotations have a high degree of completeness. However, they do not match the completeness
215 of the *D. rhopaloa* and especially *D. melanogaster* references (Table 3), both in terms of gene inclusion
216 and completeness of individual gene models. A higher number of BUSCO dipteran benchmark genes
217 are missing in the *D. prolongata* (95) and *D. carrolli* (115) annotations compared to the *D. rhopaloa*
218 (15) or *D. melanogaster* (0) references. Additionally, the transcripts in the *D. prolongata* and *D.*
219 *carrolli* annotations are shorter than those from the references, and many more BUSCO dipteran
220 benchmark genes are fragmented in the *D. prolongata* (109) and *D. carrolli* (89) annotations than for
221 *D. rhopaloa* and *D. melanogaster* (both 3). These statistics show the limitations of current algorithmic
222 annotation methods and indicate that care should be used when using gene models from these draft
223 annotations. Despite these limitations, the overall completeness is quite high, with 93.8% of BUSCO
224 benchmark genes covered in both *D. prolongata* and *D. carrolli* annotations, and comparable median
225 transcript lengths in both. These gene models will provide a good foundation for future genetic studies
226 in *D. prolongata* and relatives when used with the limitations of draft annotations in mind. Future
227 iterations of the annotations, when informed by more transcriptome data, will improve gene model
228 coverage and completeness.

Annotation	<i>D. prolongata</i>	<i>D. carrolli</i>	<i>D. rhopaloa</i>	<i>D. melanogaster</i>
Genes	19330	16346	15463	17559
Protein Coding Genes	16170	13159	14607	13986
Exons	178992	168247	154625	190719
Median Transcript Length (bp)	1635	1758	1995	1954
Longest Transcript (bp)	63866	63847	65859	71382
BUSCO Complete	93.8% (3081)	93.8% (3081)	99.4% (3267)	99.9% (3282)
BUSCO Fragmented	3.3% (109)	2.7% (89)	0.1% (3)	0.1% (3)
BUSCO Missing	2.9% (95)	3.5% (115)	0.5% (15)	0.0% (0)

230 Table 3: Statistics for annotation completeness for *D. prolongata* and *D. carrolli* annotations alongside

231 reference annotations *D. rhopaloa* GCF_018152115.1 and *D. melanogaster* GCF_000001215.4.

232 BUSCO statistics are for the 3285 genes in the diptera_odb10 benchmark set.

233

234 Potential regional duplications

235 The other major caveat for this assembly and annotation is the extent of identified duplication,
236 even after removing duplicate scaffolds. This stands out most clearly in the *D. prolongata* assembly
237 BUSCO scores, where 158 benchmark single-copy genes were identified as duplicated compared to 12
238 for *D. rhopaloa* and 8 for *D. melanogaster* (Table 1). Additional signals of duplicated regions include
239 the total length of the draft assembly and total gene number in the annotation, which are both higher
240 than in the *D. melanogaster* and *D. rhopaloa* references (Tables 1 and 3), and duplicated regions visible
241 in the whole genome alignment (Figure 3). This suggests some genome regions are represented more
242 than once in the assembly, in addition to any true *D. prolongata*-specific duplication events. Our
243 duplicate gene labeling method identified 945 of 19330 genes (4.89%, close to the BUSCO duplicate
244 frequency); these results are included in Table S4, with a list of duplicated genes on Sheet 1 and the
245 regions and relationships between pairs on Sheet 2; care should be taken when working with these
246 genes and regions. We note that all major (>1Mb) scaffolds in *D. prolongata* have duplicated BUSCO

247 genes even after removal of the fully duplicate scaffolds (Table S1, Figure S3). In contrast, removed
248 scaffolds tend to be intermediate in size and have less repeat content (Figure S7). Remaining BUSCO
249 duplications per scaffold for the final assembly are provided in Sheet 2 of Table S1.

250 Duplication artifacts often result from heterozygosity persisting through inbreeding (Guo et al.
251 2016; Kardos et al. 2018; Smith et al. 2019). Segregating inversions, in particular, can capture stretches
252 of heterozygosity and cause the assembler to split haplotypes into separate scaffolds. Consistent with
253 this explanation, the largest remaining duplication candidate visible in the whole genome alignment
254 spans a segregating inversion (Figure 3A'). Sorting biologically real from artifactual duplicates is a key
255 area of improvement for future *D. prolongata* assemblies.

256

257 **Data Availability**

258 The final deduplicated assembly for this Whole Genome Shotgun project has been deposited at
259 DDBJ/ENA/GenBank under the accession JAYMZC000000000; the version described in this paper is
260 version JAYMZC010000000. All sequence data used for genome annotation have been deposited in the
261 NCBI Sequence Read Archive under BioProject PRJNA1057277. Genome annotation files for *D.*
262 *prolongata* and *D. carrolli*, the Dovetail Falcon and HiRise assemblies (containing duplicate scaffolds),
263 sequence file for removed duplicate scaffolds, and all sequence and information files provided by
264 Dovetail have been uploaded to Dryad (URL TBD).

265

266 **Acknowledgements**

267 We thank Dr. Hisaki Takamori for providing the original SaPa01 and BaVi44 strain of *D.*
268 *prolongata*. This work was supported by NIH grant R35 GM122592 to A.K., by NSF award 1601130 to
269 D.L., by the UC Davis Center for Population Biology Pengelley Award to D. L., by UC Davis Center
270 for Population Biology Research Award to Y. L., and by funds from the UNC School of Medicine
271 Strategic Plan to CDJ.

272

273 Literature Cited

- 274 Amino, Kai, and Takashi Matsuo. 2023a. “Effects of a Past Contest on the Future Winning Probability
275 in a Hyper-Aggressive Fruit Fly.” *Ethology* 129 (8): 380–89. <https://doi.org/10.1111/eth.13375>.
- 276 ———. 2023b. “Reproductive Advantage of the Winners of Male-Male Competition in *Drosophila*
277 *Prolongata*.” *Behavioural Processes* 206 (January): 104831.
278 <https://doi.org/10.1016/j.beproc.2023.104831>.
- 279 Benson, Gary. 1999. “Tandem Repeats Finder: A Program to Analyze DNA Sequences.” *Nucleic Acids*
280 *Research* 27 (2): 573–80. <https://doi.org/10.1093/nar/27.2.573>.
- 281 Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin
282 Bealer, and Thomas L Madden. 2009. “BLAST+: Architecture and Applications.” *BMC*
283 *Bioinformatics* 10: 1–9. <https://doi.org/10.1186/1471-2105-10-421>.
- 284 Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia
285 Clum, Christopher Dunn, et al. 2016. “Phased Diploid Genome Assembly with Single-Molecule
286 Real-Time Sequencing.” *Nature Methods* 13 (October): 1050.
287 <http://dx.doi.org/10.1038/nmeth.4035>.
- 288 Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte,
289 and Arian F. Smit. 2020. “RepeatModeler2 for Automated Genomic Discovery of Transposable
290 Element Families.” *Proceedings of the National Academy of Sciences of the United States of*
291 *America* 117 (17): 9451–57. <https://doi.org/10.1073/pnas.1921046117>.
- 292 Gramates, L. Sian, Julie Agapite, Helen Attrill, Brian R. Calvi, Madeline A. Crosby, Gilberto dos
293 Santos, Joshua L. Goodman, et al. 2022. “FlyBase: A Guided Tour of Highlighted Features.”
294 *Genetics* 220 (4). <https://doi.org/10.1093/genetics/iyac035>.
- 295 Gremme, Gordon, Sascha Steinbiss, and Stefan Kurtz. 2013. “Genome Tools: A Comprehensive
296 Software Library for Efficient Processing of Structured Genome Annotations.” *IEEE/ACM*
297 *Transactions on Computational Biology and Bioinformatics* 10 (3): 645–56.
298 <https://doi.org/10.1109/TCBB.2013.68>.
- 299 Guo, Longhua, Shasha Zhang, Boris Rubinstein, Eric Ross, and Alejandro Sánchez Alvarado. 2016.
300 “Widespread Maintenance of Genome Heterozygosity in *Schmidtea Mediterranea*.” *Nature*
301 *Ecology & Evolution* 1 (1): 1–10. <https://doi.org/10.1038/s41559-016-0019>.
- 302 Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, D Philip, Joshua Bowden,
303 Matthew Brian Couger, et al. 2013. *De Novo Transcript Sequence Reconstruction from RNA-Seq:*
304 *Reference Generation and Analysis with Trinity. Nature Protocols*. Vol. 8.
305 <https://doi.org/10.1038/nprot.2013.084.De>.

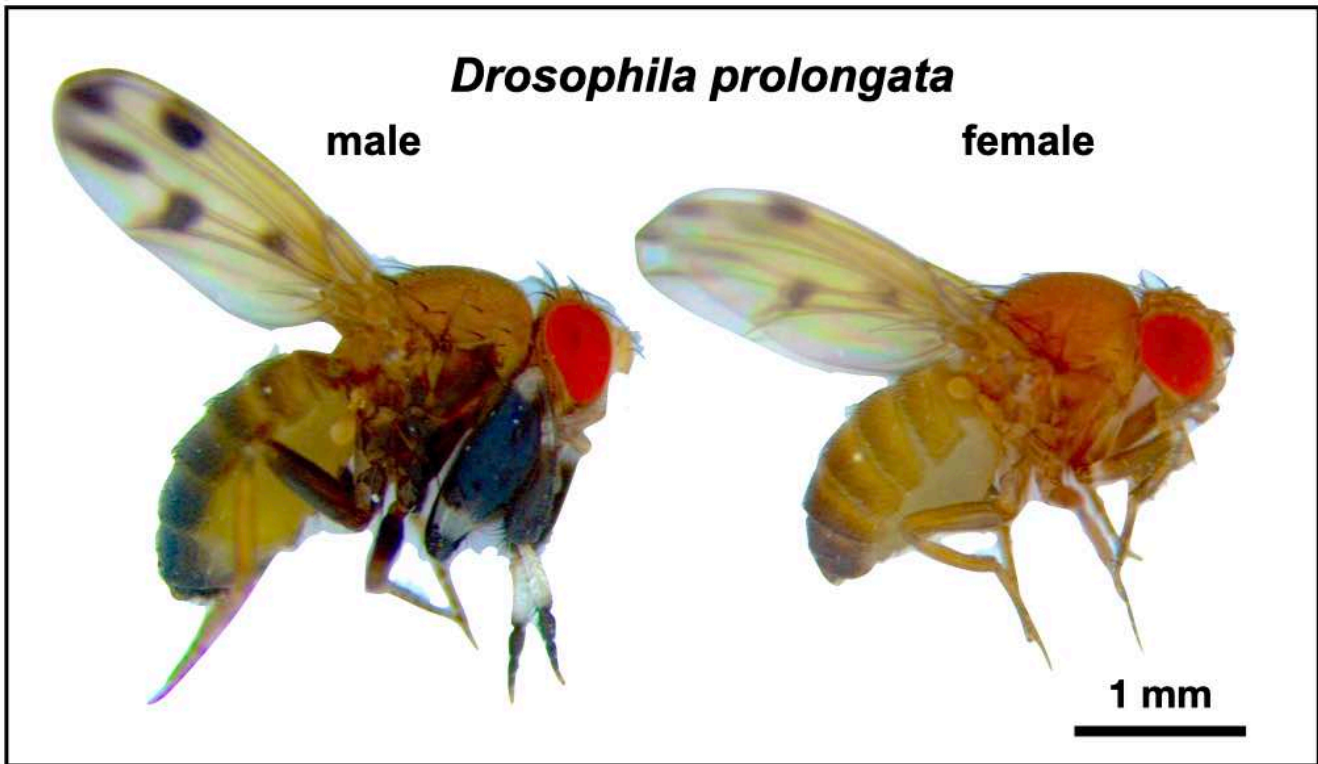
- 306 Holt, Carson, and Mark Yandell. 2011. “MAKER2: An Annotation Pipeline and Genome-Database
307 Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12 (1).
308 <https://doi.org/10.1186/1471-2105-12-491>.
- 309 Hoskins, Roger A., Joseph W. Carlson, Kenneth H. Wan, Soo Park, Ivonne Mendez, Samuel E. Galle,
310 Benjamin W. Booth, et al. 2015. “The Release 6 Reference Sequence of the *Drosophila*
311 *Melanogaster* Genome.” *Genome Research* 25 (3): 445–58. <https://doi.org/10.1101/gr.185579.114>.
- 312 Kardos, Marty, Mikael Åkesson, Toby Fountain, Øystein Flagstad, Olof Liberg, Pall Olason, Håkan
313 Sand, Petter Wabakken, Camilla Wikenros, and Hans Ellegren. 2018. “Genomic Consequences of
314 Intensive Inbreeding in an Isolated Wolf Population.” *Nature Ecology and Evolution* 2 (1): 124–
315 31. <https://doi.org/10.1038/s41559-017-0375-4>.
- 316 Kim, Bernard Y., Jeremy R. Wang, Danny E. Miller, Olga Barmina, Emily Delaney, Ammon
317 Thompson, Aaron A. Comeault, et al. 2021. “Highly Contiguous Assemblies of 101 *Drosophilid*
318 Genomes.” *ELife* 10: 1–33. <https://doi.org/10.7554/eLife.66405>.
- 319 Kudo, Ayumi, Shuji Shigenobu, Koji Kadota, Masafumi Nozawa, Tomoko F. Shibata, Yukio Ishikawa,
320 and Takashi Matsuo. 2017. “Comparative Analysis of the Brain Transcriptome in a Hyper-
321 Aggressive Fruit Fly, *Drosophila Prolongata*.” *Insect Biochemistry and Molecular Biology* 82:
322 11–20. <https://doi.org/10.1016/j.ibmb.2017.01.006>.
- 323 Kudo, Ayumi, Hisaki Takamori, Hideaki Watabe, Yukio Ishikawa, and Takashi Matsuo. 2015.
324 “Variation in Morphological and Behavioral Traits among Isofemale Strains of *Drosophila*
325 *Prolongata* (Diptera: Drosophilidae).” *Entomological Science* 18 (2): 221–29.
326 <https://doi.org/10.1111/ens.12116>.
- 327 Li, Heng. 2018. “Minimap2: Pairwise Alignment for Nucleotide Sequences.” *Bioinformatics* 34 (18):
328 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
- 329 Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo
330 Abecasis, and Richard Durbin. 2009. “The Sequence Alignment/Map Format and SAMtools.”
331 *Bioinformatics* 25 (16): 2078–79. <https://doi.org/10.1093/bioinformatics/btp352>.
- 332 Lieberman-Aiden, Erez, Nynke L. Van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragozy,
333 Agnes Telling, Ido Amit, et al. 2009. “Comprehensive Mapping of Long-Range Interactions
334 Reveals Folding Principles of the Human Genome.” *Science* 326 (5950): 289–93.
335 <https://doi.org/10.1126/science.1181369>.
- 336 Luecke, David, and Artyom Kopp. 2019. “Sex-Specific Evolution of Relative Leg Size in *Drosophila*
337 *Prolongata* Results from Changes in the Intersegmental Coordination of Tissue Growth.”
338 *Evolution* 73 (11): 2281–94. <https://doi.org/10.1111/evo.13847>.
- 339 Luecke, David, Gavin Rice, and Artyom Kopp. 2022. “Sex-Specific Evolution of a *Drosophila* Sensory
340 System via Interacting *Cis*- and *Trans*-Regulatory Changes.” *Evolution and Development* 24 (1–
341 2): 37–60. <https://doi.org/10.1111/ede.12398>.

- 342 Luo, Yige, Yunwei Zhang, Jean Pierre Farine, Jean François Ferveur, Santiago Ramírez, and Artyom
343 Kopp. 2019. “Evolution of Sexually Dimorphic Pheromone Profiles Coincides with Increased
344 Number of Male-Specific Chemosensory Organs in *Drosophila Prolongata*.” *Ecology and*
345 *Evolution* 9 (23): 13608–18. <https://doi.org/10.1002/ece3.5819>.
- 346 Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021.
347 “BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper
348 Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes.” *Molecular*
349 *Biology and Evolution* 38 (10): 4647–54. <https://doi.org/10.1093/molbev/msab199>.
- 350 Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and
351 Aleksey Zimin. 2018. “MUMmer4: A Fast and Versatile Genome Alignment System.” *PLoS*
352 *Computational Biology* 14 (1): 1–14. <https://doi.org/10.1371/journal.pcbi.1005944>.
- 353 McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” *Proceedings of the 9th*
354 *Python in Science Conference* 1 (Scipy): 56–61. <https://doi.org/10.25080/majora-92bf1922-00a>.
- 355 Perteua, G, and M Perteua. 2020. “GFF Utilities: GffRead and GffCompare [Version 2; Peer Review: 3
356 Approved].” *F1000Research* 9 (304): 1–20. <https://f1000research.com/articles/9-304/v2>.
- 357 Putnam, Nicholas H, Brendan O Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert
358 Calef, Christopher J Troll, et al. 2016. “Chromosome-Scale Shotgun Assembly Using an in Vitro
359 Method for Long-Range Linkage.” *Genome Research* 26: 342–50.
360 <https://doi.org/10.1101/gr.193474.115>.Freely.
- 361 Quinlan, Aaron R., and Ira M. Hall. 2010. “BEDTools: A Flexible Suite of Utilities for Comparing
362 Genomic Features.” *Bioinformatics* 26 (6): 841–42. <https://doi.org/10.1093/bioinformatics/btq033>.
- 363 Setoguchi, Shiori, Hisaki Takamori, Tadashi Aotsuka, Jun Sese, Yukio Ishikawa, and Takashi Matsuo.
364 2014. “Sexual Dimorphism and Courtship Behavior in *Drosophila Prolongata*.” *Journal of*
365 *Ethology* 32: 91–102. <https://doi.org/10.1007/s10164-014-0399-z>.
- 366 Shumate, Alaina, and Steven L. Salzberg. 2021. “Liftoff: Accurate Mapping of Gene Annotations.”
367 *Bioinformatics* 37 (12): 1639–43. <https://doi.org/10.1093/bioinformatics/btaa1016>.
- 368 Singh, BK, and JP Gupta. 1977. “Two New and Two Unrecorded Species of the Genus *Drosophila*
369 *Fallen* (Diptera: Drosophilidae) from Shillong, Meghalaya, India.” *Proceedings of the Zoological*
370 *Society (Calcutta)* 30: 31–38.
- 371 Smith, Nicholas M.A., Claire Wade, Michael H. Allsopp, Brock A. Harpur, Amro Zayed, Stephen A.
372 Rose, Jan Engelstädter, Nadine C. Chapman, Boris Yagound, and Benjamin P. Oldroyd. 2019.
373 “Strikingly High Levels of Heterozygosity despite 20 Years of Inbreeding in a Clonal Honey Bee.”
374 *Journal of Evolutionary Biology* 32 (2): 144–52. <https://doi.org/10.1111/jeb.13397>.
- 375 Takau, Ayumi, and Takashi Matsuo. 2022. “Contribution of Visual Stimuli to Mating and Fighting
376 Behaviors of *Drosophila Prolongata*.” *Entomological Science* 25 (4).
377 <https://doi.org/10.1111/ens.12529>.

378 Toda, M. J. 1991. “Drosophilidae (Diptera) in Myanmar (Burma) VII. The *Drosophila Melanogaster*
379 Species-Group, Excepting the *D. Montium* Species-Subgroup.” *Oriental Insects* 25 (1): 69–94.
380 <https://doi.org/10.1080/00305316.1991.10432216>.

381 Toyoshima, Naoki, and Takashi Matsuo. 2023. “Fight Outcome Influences Male Mating Success in
382 *Drosophila Prolongata*.” *Journal of Ethology* 41 (2): 119–27. [https://doi.org/10.1007/s10164-023-](https://doi.org/10.1007/s10164-023-00778-1)
383 00778-1.
384

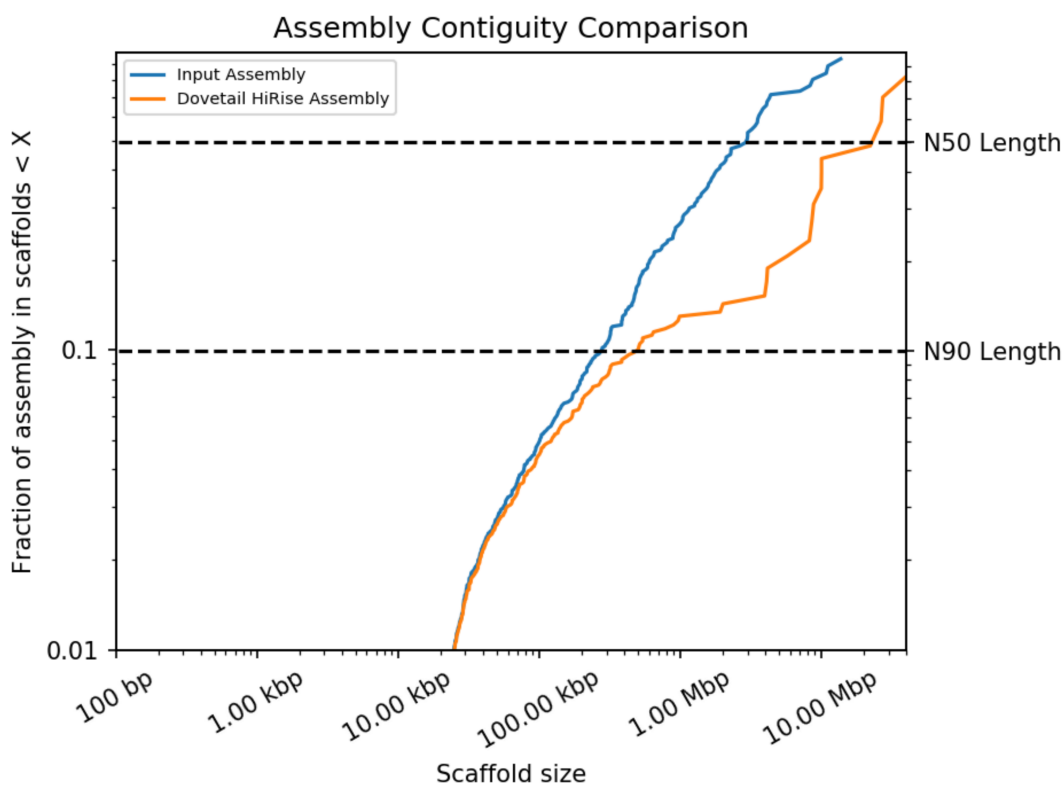
385



386 Figure 1. *Drosophila prolongata* has a suite of recently evolved male-specific traits, ideal for studying
387 the evolution of sexual dimorphism. Most noticeable is the size and pigmentation banding of front legs
388 in males. Other sexually dimorphic characteristics include wing spots, eye shape, pigmentation, and
389 increased length of second and third legs.

390

	Input Assembly	Dovetail HiRise Assembly
Total Length	223.32 Mb	223.34 Mb
L50/N50	21 scaffolds; 2.889 Mb	4 scaffolds; 22.190 Mb
L90/N90	124 scaffolds; 0.274 Mb	28 scaffolds; 0.472 Mb



391 Figure 2. Dovetail assembly process generates high contiguity assembly. Comparison between initial
392 PacBio FALCON with Arrow polished assembly (“Input Assembly”) and final assembly generated by
393 Dovetail HiC scaffolding method (“HiRise Assembly”), provided by Dovetail genomics. Each curve
394 shows the fraction of the total length of the assembly in scaffolds of a given length or smaller. Scaffolds
395 shorter than 1kb are excluded.

396

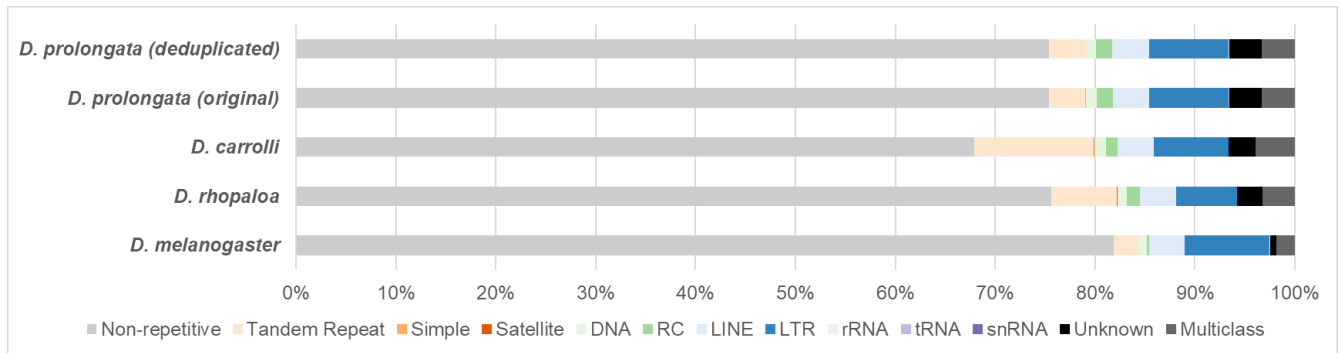
397

398

399

400

401

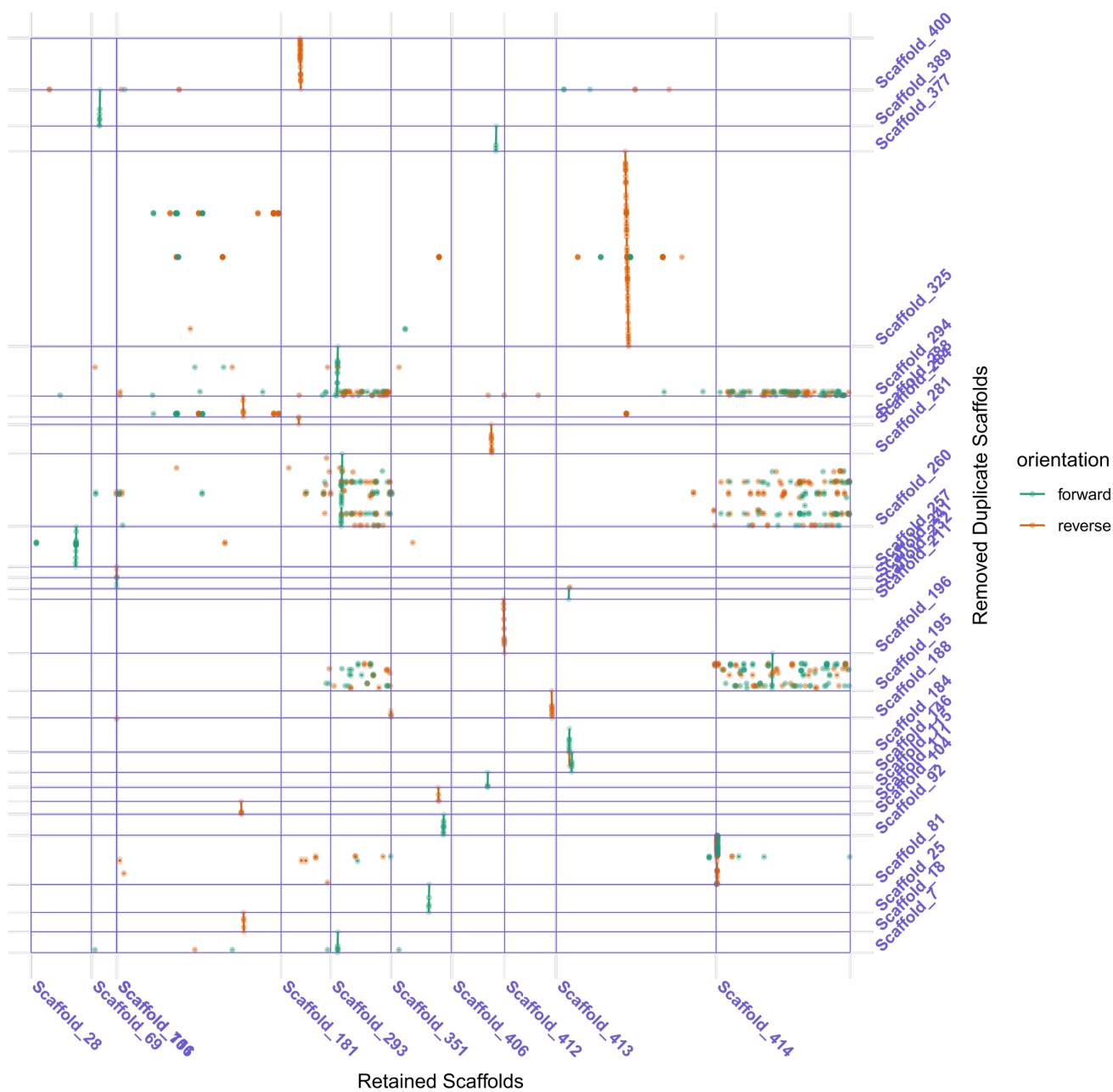


414

415 Figure 4. Genome-wide repeat content of *D. prolongata* (before and after de-duplication) and related
416 species. Repeat contents are color coded as follows. Low-complexity regions (Tandem repeats, simple
417 repeats, Satellite): orange palette, DNA transposons (DNA, RC): green palette, retrotransposons (LINE,
418 LTR): blue palette, RNA: purple palette. Abbreviations for each repeat class are as follows. RC:
419 Rolling Circle transposons, LINE: Long-Interspersed Nuclear Element, LTR: Long-Terminal-Repeats
420 retrotransposon, snRNA: small-nuclear RNA, Unknown: unknown class of repeats/transposons,
421 Multiclass: sequences belonging to more than one repeat class.

422

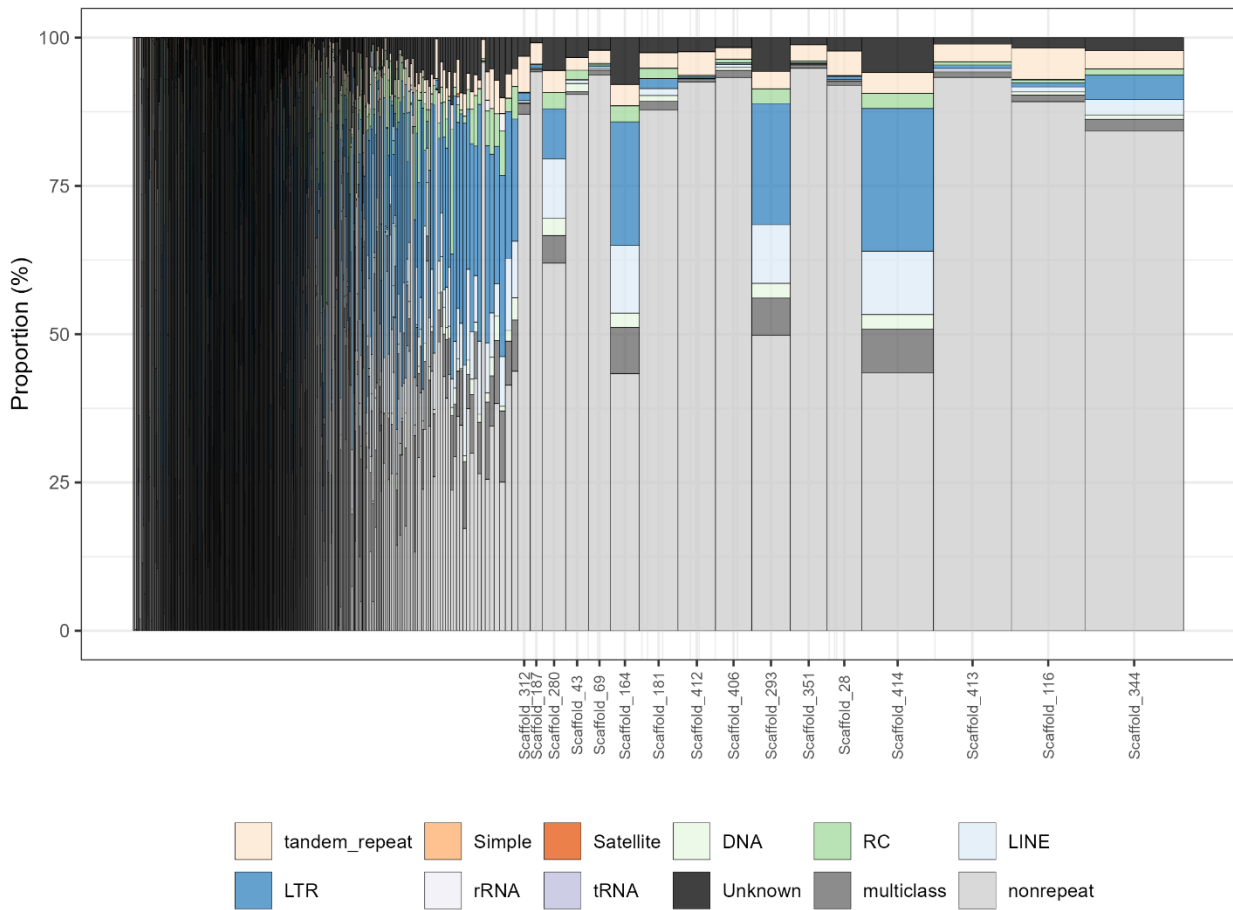
423

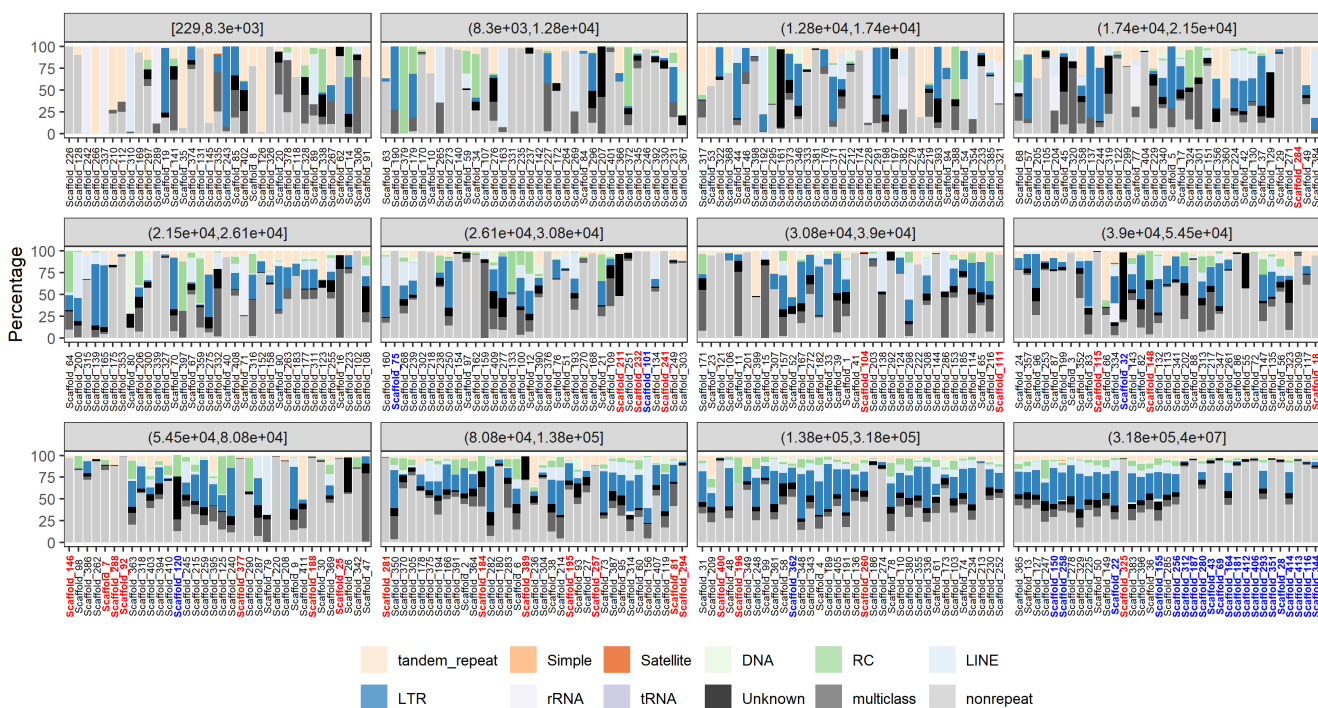


424 Figure S1. Pairwise MUMmer alignments between 27 duplicate scaffolds and sister scaffolds. Straight
425 lines show alignment between duplicate scaffolds (y-axis) and sister scaffolds (x-axis), with alignment
426 boundaries indicated by flanking points. Sense alignment between scaffolds is shown in green, and
427 antisense alignment is in orange.

428

429





442

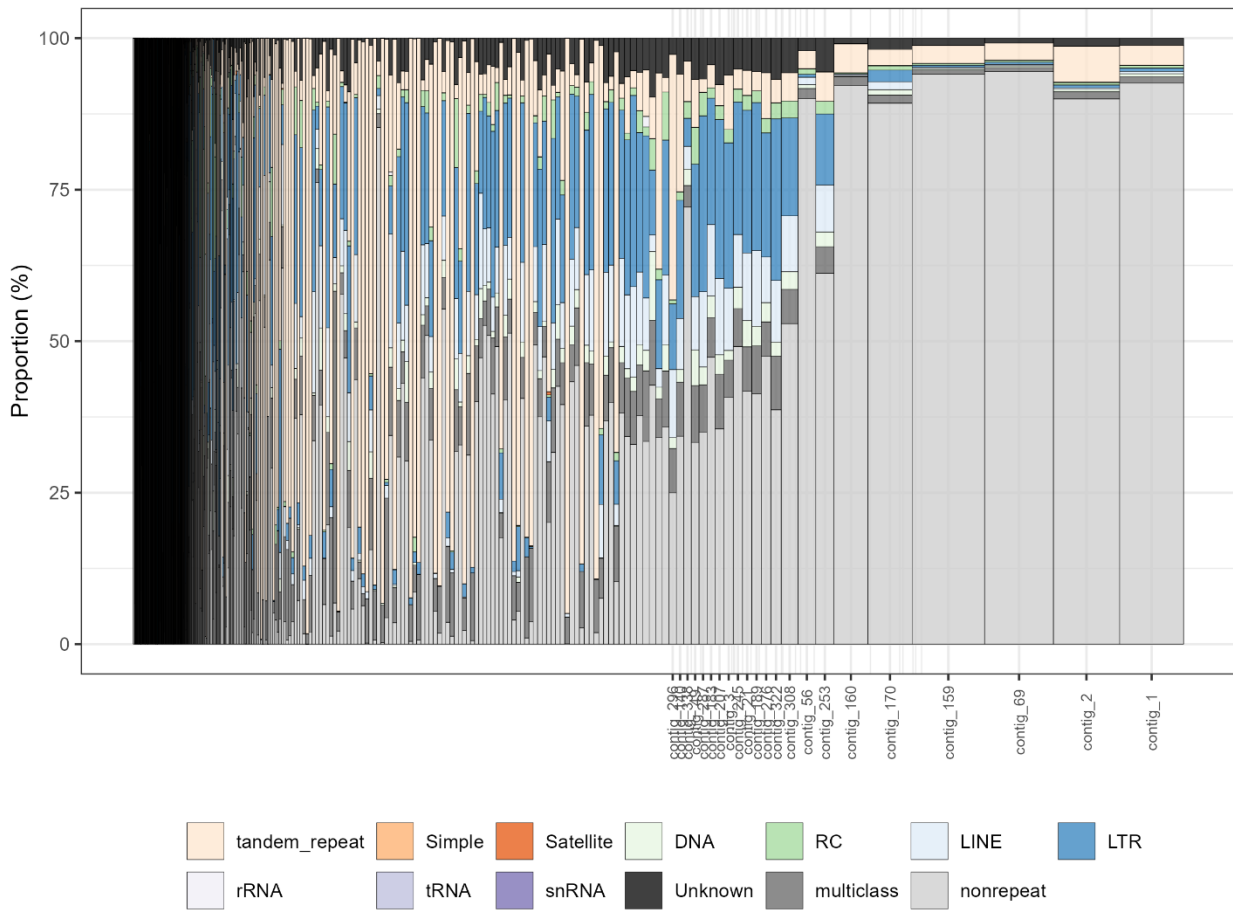
443 Figure S3. Stacked bar plots showing the distribution of repeat content by scaffolds (partitioned by

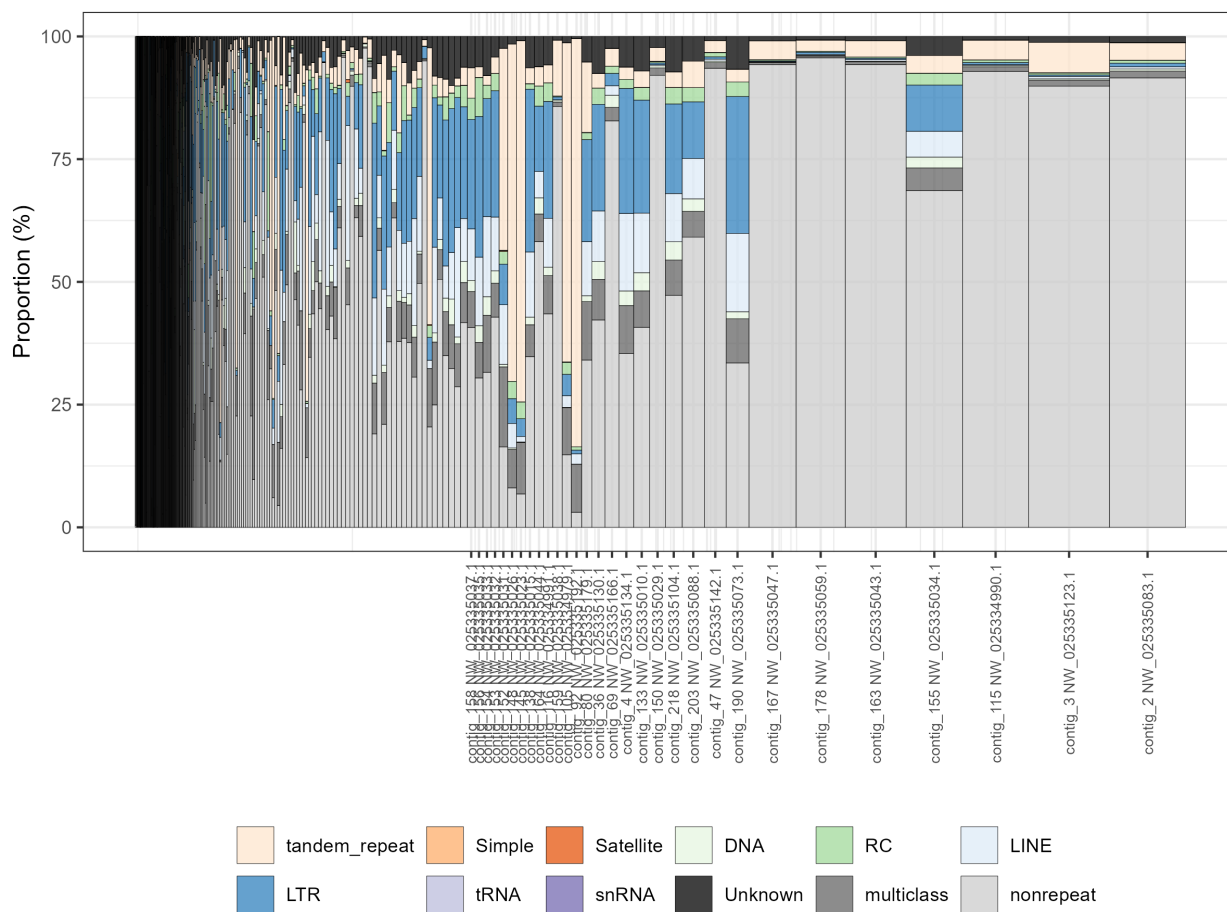
444 scaffold length bins) in *D. prolongata* genome assembly. Scaffold names are ordered by their

445 corresponding lengths. Repeat contents are color coded as Fig. S2, with the exception that removed

446 scaffolds have names colored red, and retained members of duplicate scaffold pairs are colored in blue.

447

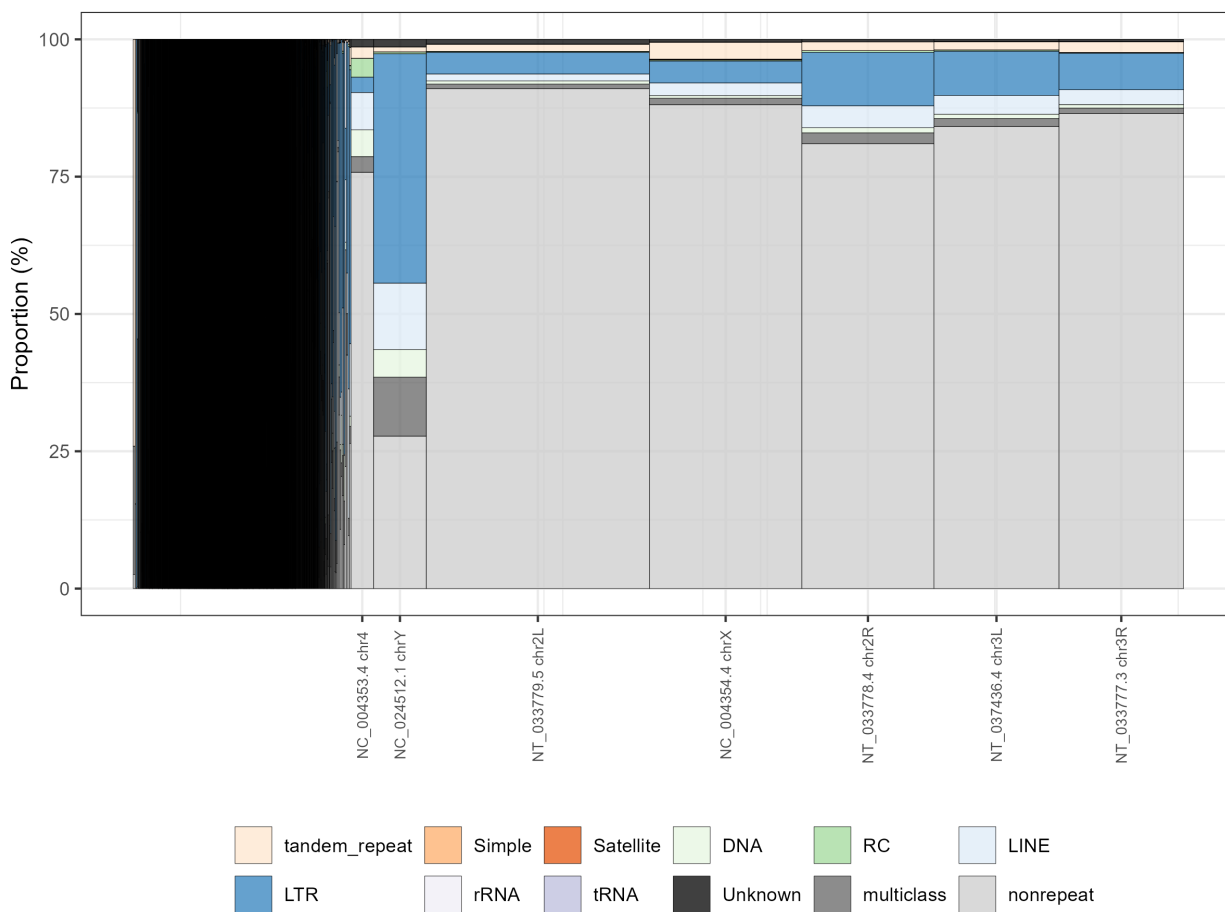


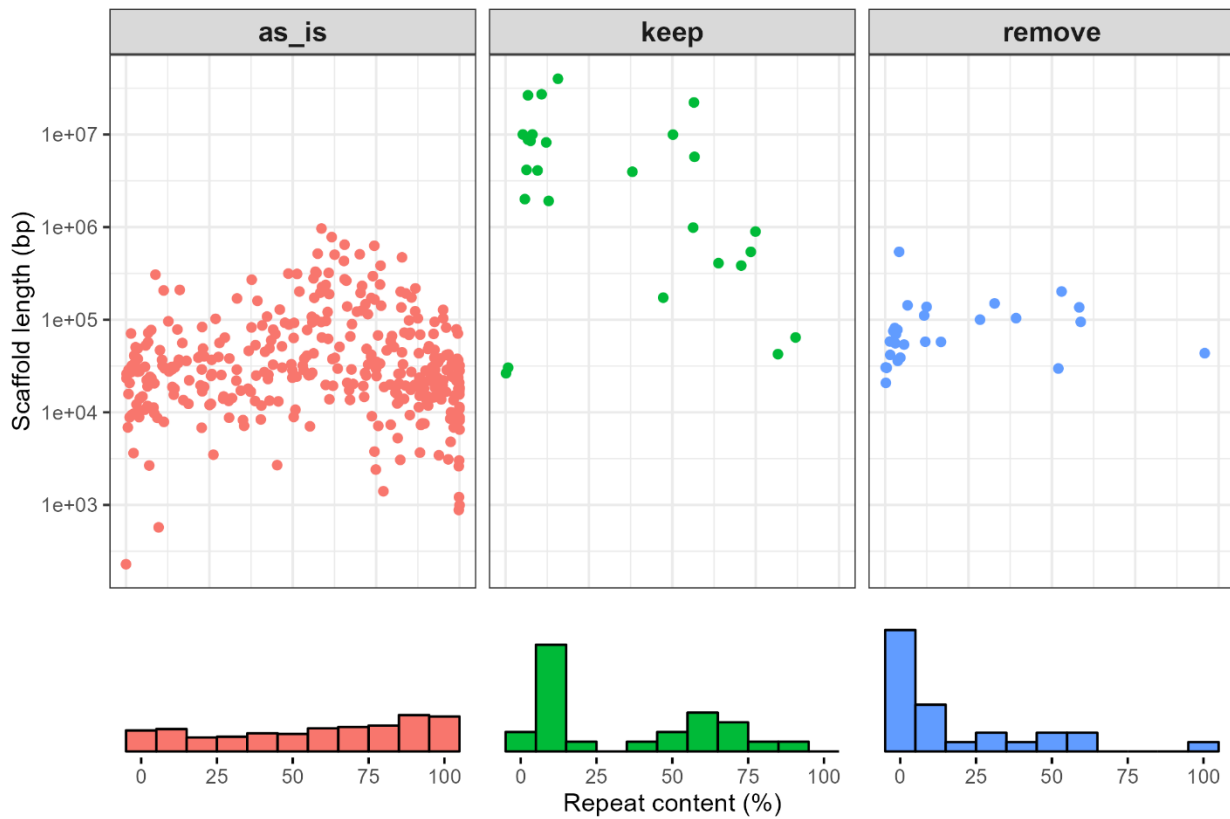


453

454 Figure S5. Stacked bar plots showing the distribution of repeat content by scaffolds in *D. rhopaloa*
455 genome assembly (GCF_018152115.1_ASM1815211v1). Widths of bars are proportional to the square
456 root of scaffold/chromosome lengths. Repeat contents are color coded as Fig. S2.

457





463

464 Figure S7. Scatter plots showing the distribution of repeat profiles by scaffolds under each category in
465 the complete *D. prolongata* genome assembly. Frequency histograms of repeat content are displayed at

466 the bottom. X-axis is the repeat content (%), and the y-axis is the corresponding scaffold length.

467 Scaffolds with no BUSCO duplicates are colored in red (as_is), retained scaffolds with BUSCO

468 duplicates in green (keep), and removed scaffolds with BUSCO duplicates in blue (remove).

469