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

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

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

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

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

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

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

The genome of *Drosophila prolongata*

Keywords: Genome, Drosophila, Sex Dimorphism

Abstract

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

Introduction

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

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

Materials and Methods

Genome line generation

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

Tissue collection

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

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

Sequencing and assembly

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

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

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

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

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

Gene prediction and annotation

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

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

Removal of duplicate scaffolds

BUSCO (Manni et al. 2021) analysis of the Dovetail HiRise using the diptera_ocb10 lineage dataset revealed 200 complete but duplicated benchmark genes (Table S1), indicating potential duplicated regions in the assembly. Scaffolds were assessed for BUSCO benchmark gene content and sorted by the percentage of duplicated BUSCO genes. 53 candidate scaffolds, ranging from 20,819bp to 39,990,007bp, contained at least one duplicated benchmark BUSCO gene (Table S1). Inspection of MUMmer (Marçais et al. 2018) alignments between duplicate-containing candidates and scaffolds with alternate copies of the duplicated benchmark genes showed complete alignment across 27 of the candidate scaffolds (Figure S1). These 27 scaffolds (ranging from 20,819bp to 541,551bp) were considered fully duplicated and split from the assembly and annotation (Files S1, S2, and S3) using SAMtools 1.15.1 (Li et al. 2009). Custom Python pandas 1.1.2 (McKinney 2010), and R 4.0.3 (R Core Team 2020, https://www.R-project.org) for scaffold sorting by BUSCO scores, splitting assembly and annotation, and inspecting genome alignments are available at https://github.com/dluecke/annotation_tools.

Identification of duplicate genes

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

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

Repeat analysis

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

Assembly and annotation evaluation

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

Results and Discussion

Assembly contiguity

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

Assembly	*D. prolongata*	*D. carrolli*	*D. rhopaloa*	*D. melanogaster*
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)

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

Assembly completeness

BUSCO results for assemblies (Table 1) show a comparable degree of completeness for the 3285 genes in the BUSCO dipteran benchmark set between *D. prolongata* assembly and references, 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	***D. prolongata* (%)**	***D. carrolli* (%)**	***D. rhopaloa* (%)**	***D. melanogaster* (%)**
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**

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

Annotation completeness

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

Annotation	*D. prolongata*	*D. carrolli*	*D. rhopaloa*	*D. melanogaster*
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)

Table 3: Statistics for annotation completeness for *D. prolongata* and *D. carrolli* annotations alongside reference annotations *D. rhopaloa* GCF_018152115.1 and *D. melanogaster* GCF_000001215.4. BUSCO statistics are for the 3285 genes in the diptera_odb10 benchmark set.

Potential regional duplications

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

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

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

Data Availability

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

Acknowledgements

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

Literature Cited

Amino, Kai, and Takashi Matsuo. 2023a. "Effects of a Past Contest on the Future Winning Probability in a Hyper-Aggressive Fruit Fly." *Ethology* 129 (8): 380–89. https://doi.org/10.1111/eth.13375.

———. 2023b. "Reproductive Advantage of the Winners of Male-Male Competition in *Drosophila Prolongata*." *Behavioural Processes* 206 (January): 104831. https://doi.org/10.1016/j.beproc.2023.104831.

Benson, Gary. 1999. "Tandem Repeats Finder: A Program to Analyze DNA Sequences." *Nucleic Acids Research* 27 (2): 573–80. https://doi.org/10.1093/nar/27.2.573.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10: 1–9. https://doi.org/10.1186/1471-2105-10-421.

Chin, Chen-Shan, Paul Peluso, Fritz J Sedlazeck, Maria Nattestad, Gregory T Concepcion, Alicia Clum, Christopher Dunn, et al. 2016. "Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing." *Nature Methods* 13 (October): 1050. http://dx.doi.org/10.1038/nmeth.4035.

Flynn, Jullien M., Robert Hubley, Clément Goubert, Jeb Rosen, Andrew G. Clark, Cédric Feschotte, and Arian F. Smit. 2020. "RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families." *Proceedings of the National Academy of Sciences of the United States of America* 117 (17): 9451–57. https://doi.org/10.1073/pnas.1921046117.

Gramates, L. Sian, Julie Agapite, Helen Attrill, Brian R. Calvi, Madeline A. Crosby, Gilberto dos Santos, Joshua L. Goodman, et al. 2022. "FlyBase: A Guided Tour of Highlighted Features." *Genetics* 220 (4). https://doi.org/10.1093/genetics/iyac035.

Gremme, Gordon, Sascha Steinbiss, and Stefan Kurtz. 2013. "Genome Tools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10 (3): 645–56. https://doi.org/10.1109/TCBB.2013.68.

Guo, Longhua, Shasha Zhang, Boris Rubinstein, Eric Ross, and Alejandro Sánchez Alvarado. 2016. "Widespread Maintenance of Genome Heterozygosity in *Schmidtea Mediterranea*." *Nature Ecology & Evolution* 1 (1): 1–10. https://doi.org/10.1038/s41559-016-0019.

Haas, Brian J, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, D Philip, Joshua Bowden, Matthew Brian Couger, et al. 2013. *De Novo Transcript Sequence Recostruction from RNA-Seq: Reference Generation and Analysis with Trinity*. *Nature Protocols*. Vol. 8. 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 Ragoczy,
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.

Luo, Yige, Yunwei Zhang, Jean Pierre Farine, Jean François Ferveur, Santiago Ramírez, and Artyom Kopp. 2019. "Evolution of Sexually Dimorphic Pheromone Profiles Coincides with Increased Number of Male-Specific Chemosensory Organs in *Drosophila Prolongata*." *Ecology and Evolution* 9 (23): 13608–18. https://doi.org/10.1002/ece3.5819.

Manni, Mosè, Matthew R. Berkeley, Mathieu Seppey, Felipe A. Simão, and Evgeny M. Zdobnov. 2021. "BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes." *Molecular Biology and Evolution* 38 (10): 4647–54. https://doi.org/10.1093/molbev/msab199.

Marçais, Guillaume, Arthur L. Delcher, Adam M. Phillippy, Rachel Coston, Steven L. Salzberg, and Aleksey Zimin. 2018. "MUMmer4: A Fast and Versatile Genome Alignment System." *PLoS Computational Biology* 14 (1): 1–14. https://doi.org/10.1371/journal.pcbi.1005944.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." *Proceedings of the 9th Python in Science Conference* 1 (Scipy): 56–61. https://doi.org/10.25080/majora-92bf1922-00a.

Pertea, G, and M Pertea. 2020. "GFF Utilities: GffRead and GffCompare [Version 2; Peer Review: 3 Approved]." *F1000Research* 9 (304): 1–20. https://f1000research.com/articles/9-304/v2.

Putnam, Nicholas H, Brendan O Connell, Jonathan C Stites, Brandon J Rice, Marco Blanchette, Robert Calef, Christopher J Troll, et al. 2016. "Chromosome-Scale Shotgun Assembly Using an in Vitro Method for Long-Range Linkage." *Genome Research* 26: 342–50. https://doi.org/10.1101/gr.193474.115.Freely.

Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42. https://doi.org/10.1093/bioinformatics/btq033.

Setoguchi, Shiori, Hisaki Takamori, Tadashi Aotsuka, Jun Sese, Yukio Ishikawa, and Takashi Matsuo. 2014. "Sexual Dimorphism and Courtship Behavior in *Drosophila Prolongata*." *Journal of Ethology* 32: 91–102. https://doi.org/10.1007/s10164-014-0399-z.

Shumate, Alaina, and Steven L. Salzberg. 2021. "Liftoff: Accurate Mapping of Gene Annotations." *Bioinformatics* 37 (12): 1639–43. https://doi.org/10.1093/bioinformatics/btaa1016.

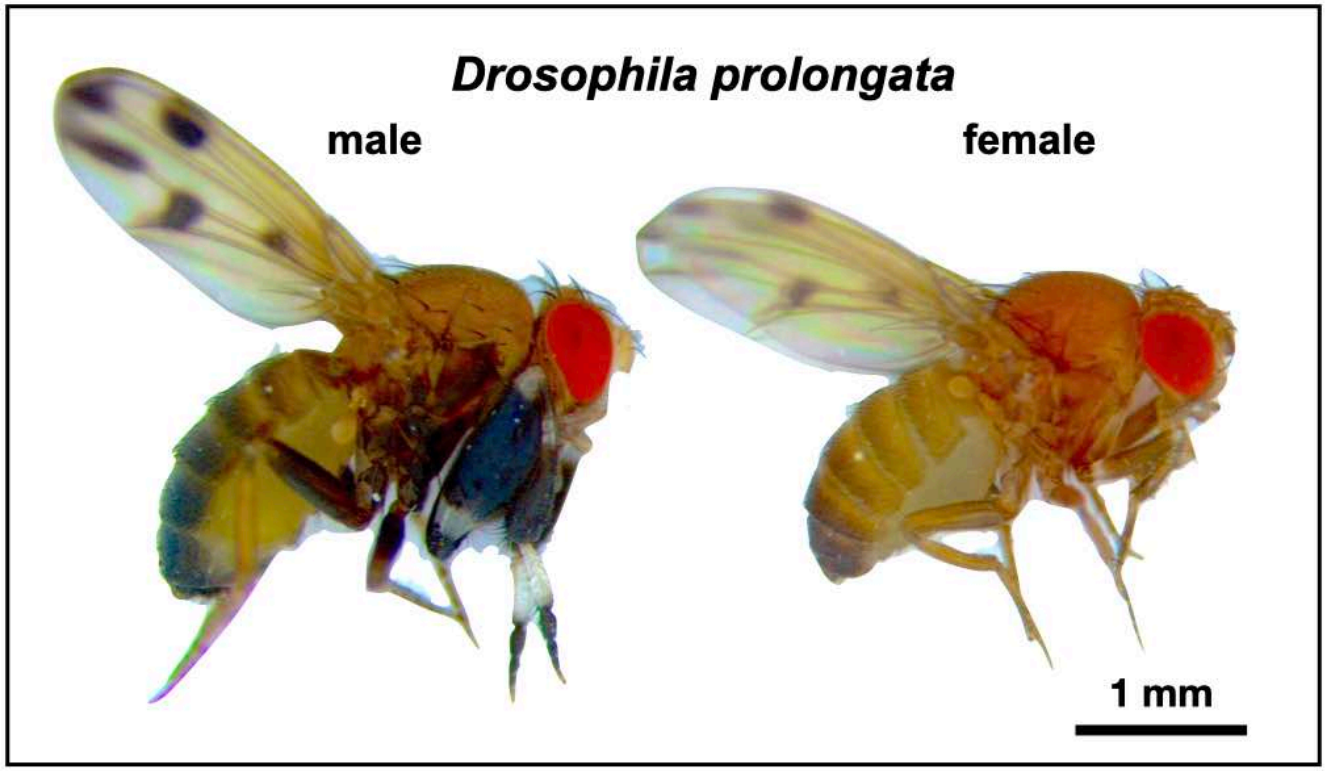
Singh, BK, and JP Gupta. 1977. "Two New and Two Unrecorded Species of the Genus *Drosophila* Fallen (Diptera: Drosophilidae) from Shillong, Meghalaya, India." *Proceedings of the Zoological Society (Calcutta)* 30: 31–38.

Smith, Nicholas M.A., Claire Wade, Michael H. Allsopp, Brock A. Harpur, Amro Zayed, Stephen A. Rose, Jan Engelstädter, Nadine C. Chapman, Boris Yagound, and Benjamin P. Oldroyd. 2019. "Strikingly High Levels of Heterozygosity despite 20 Years of Inbreeding in a Clonal Honey Bee." *Journal of Evolutionary Biology* 32 (2): 144–52. https://doi.org/10.1111/jeb.13397.

Takau, Ayumi, and Takashi Matsuo. 2022. "Contribution of Visual Stimuli to Mating and Fighting Behaviors of *Drosophila Prolongata*." *Entomological Science* 25 (4). https://doi.org/10.1111/ens.12529.

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

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



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

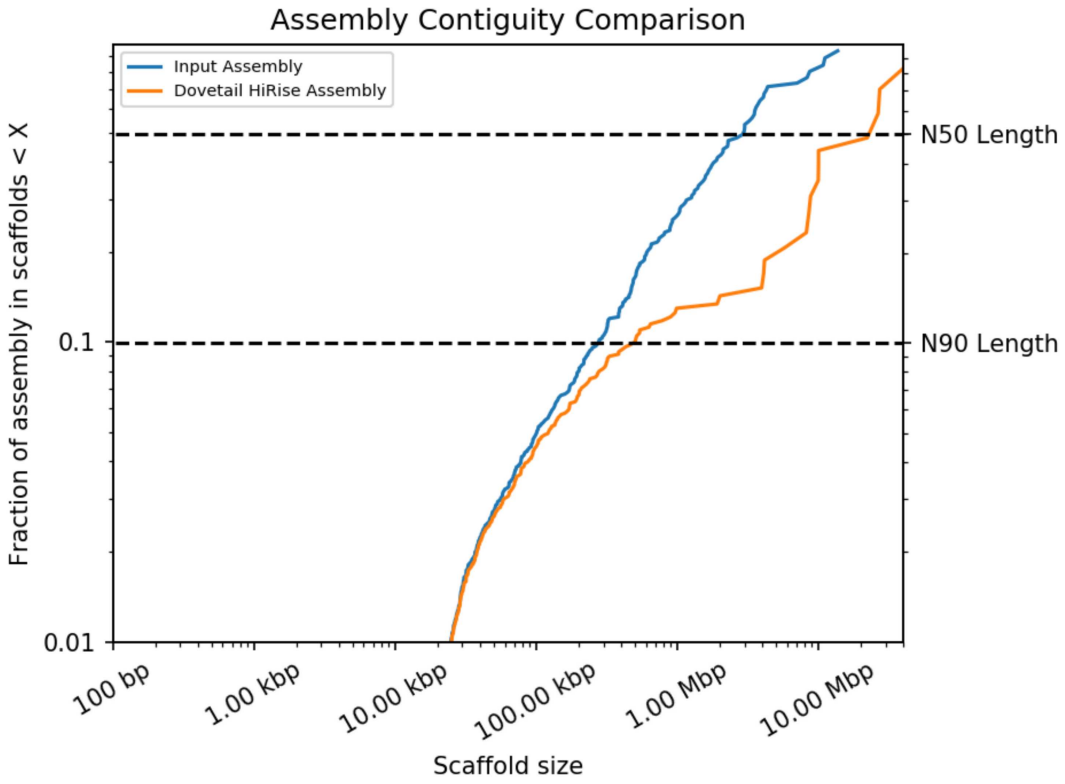


Figure 2. Dovetail assembly process generates high contiguity assembly. Comparison between initial PacBio FALCON with Arrow polished assembly ("Input Assembly") and final assembly generated by Dovetail HiC scaffolding method ("HiRise Assembly"), provided by Dovetail genomics. Each curve shows the fraction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

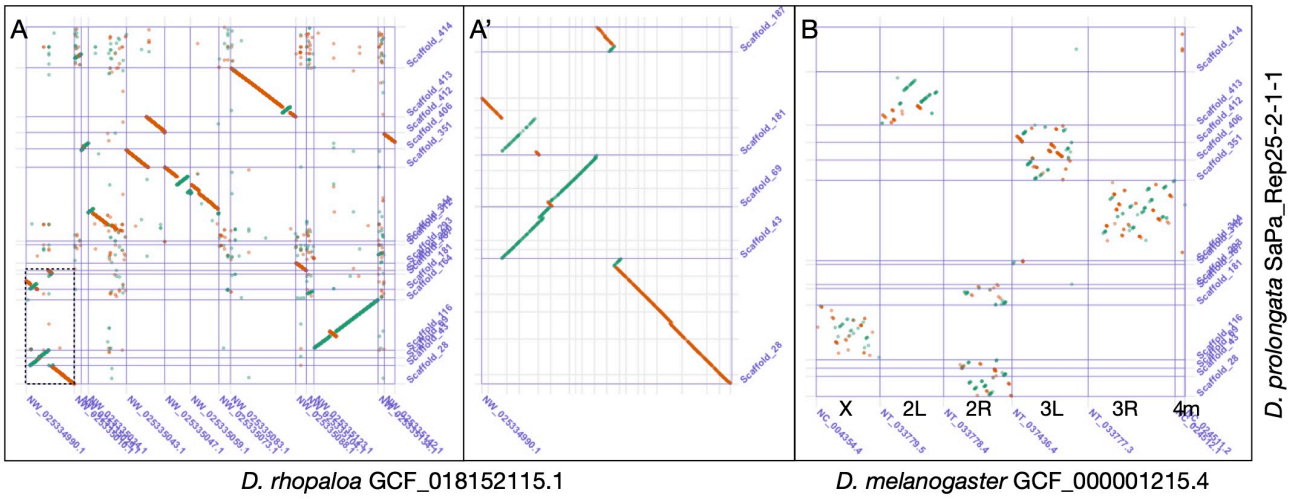
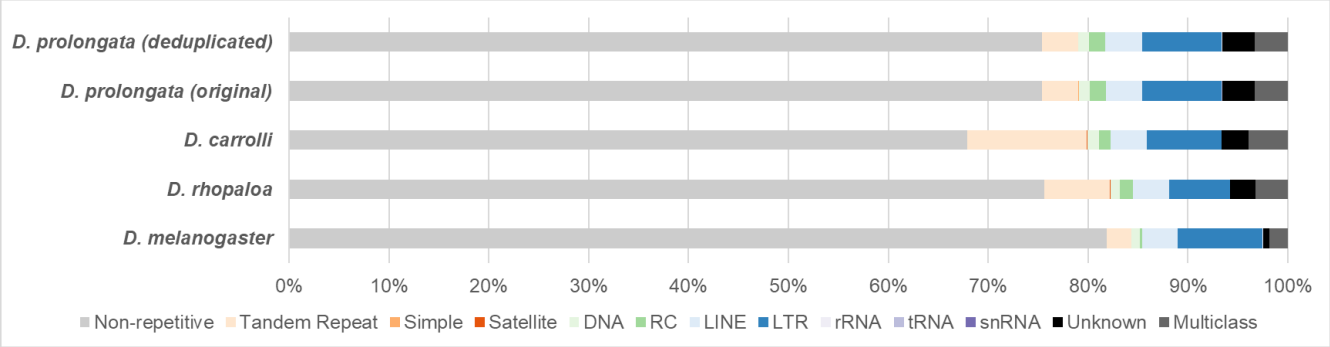


Figure 3. Whole genome alignments between major scaffolds of *D. prolongata* (>1Mb) assembly and reference assemblies. Sense matches are shown in green, and anti-sense matches in orange. (A) Alignment to *D. rhopaloa* reference based on minimum 1000bp matches, showing reference scaffolds >2.5Mb as ordered in assembly; boxed area is expanded in panel A'. (A') Zoom on portion of alignment A, showing regional duplication and inversion. (B) Alignment to major chromosome arms from *D. melanogaster* assembly, based on minimum 500bp matches. Large stretches of contiguity with limited large inversions are evident between *D. prolongata* and *D. rhopaloa* (A), while conservation of each chromosome arm's content along with considerable intra-arm rearrangement is seen between *D. prolongata* and *D. melanogaster* (B). A duplication spanning an inversion is evident between Scaffold_43 and Scaffold_181 (A').

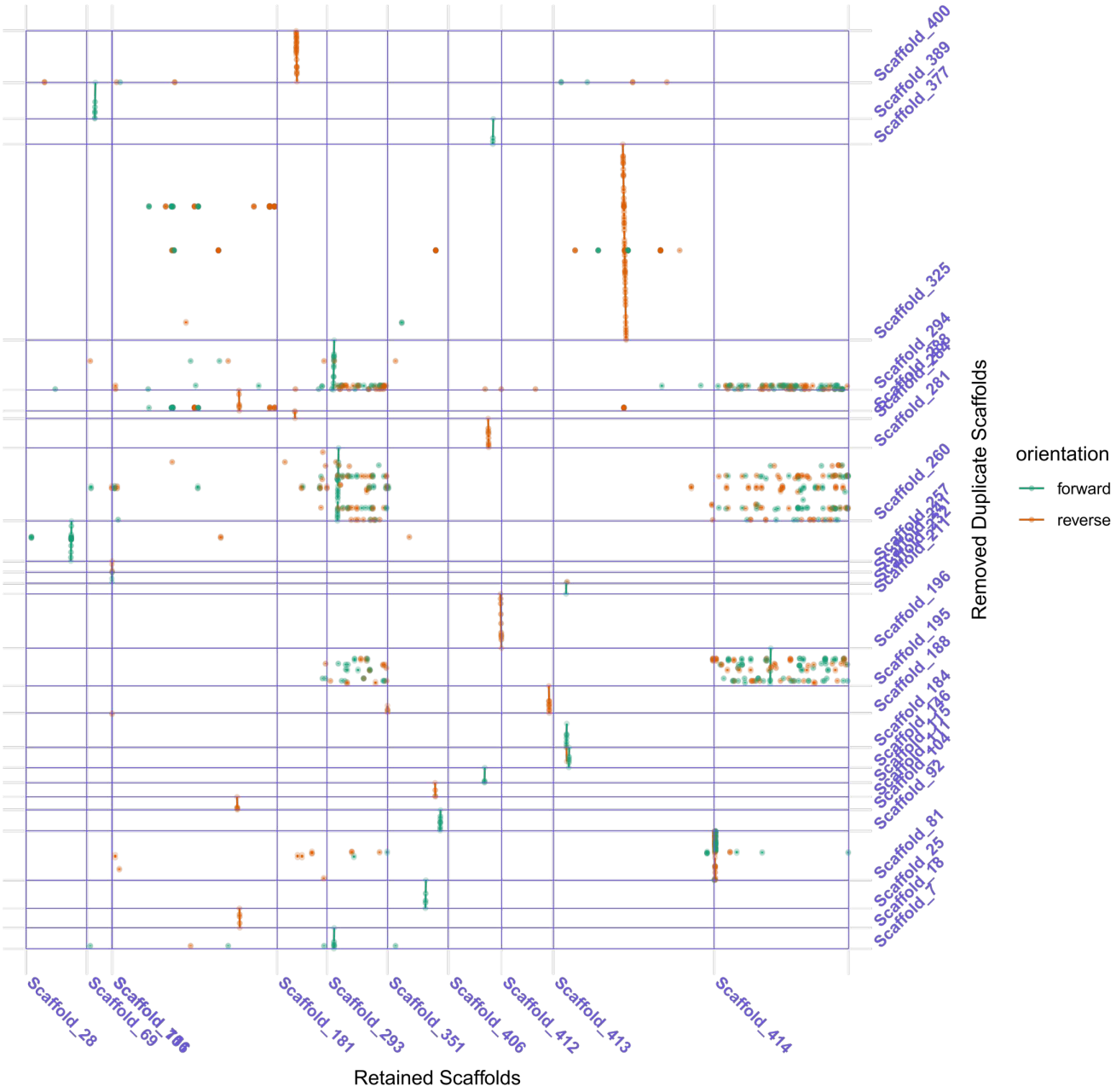


414

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

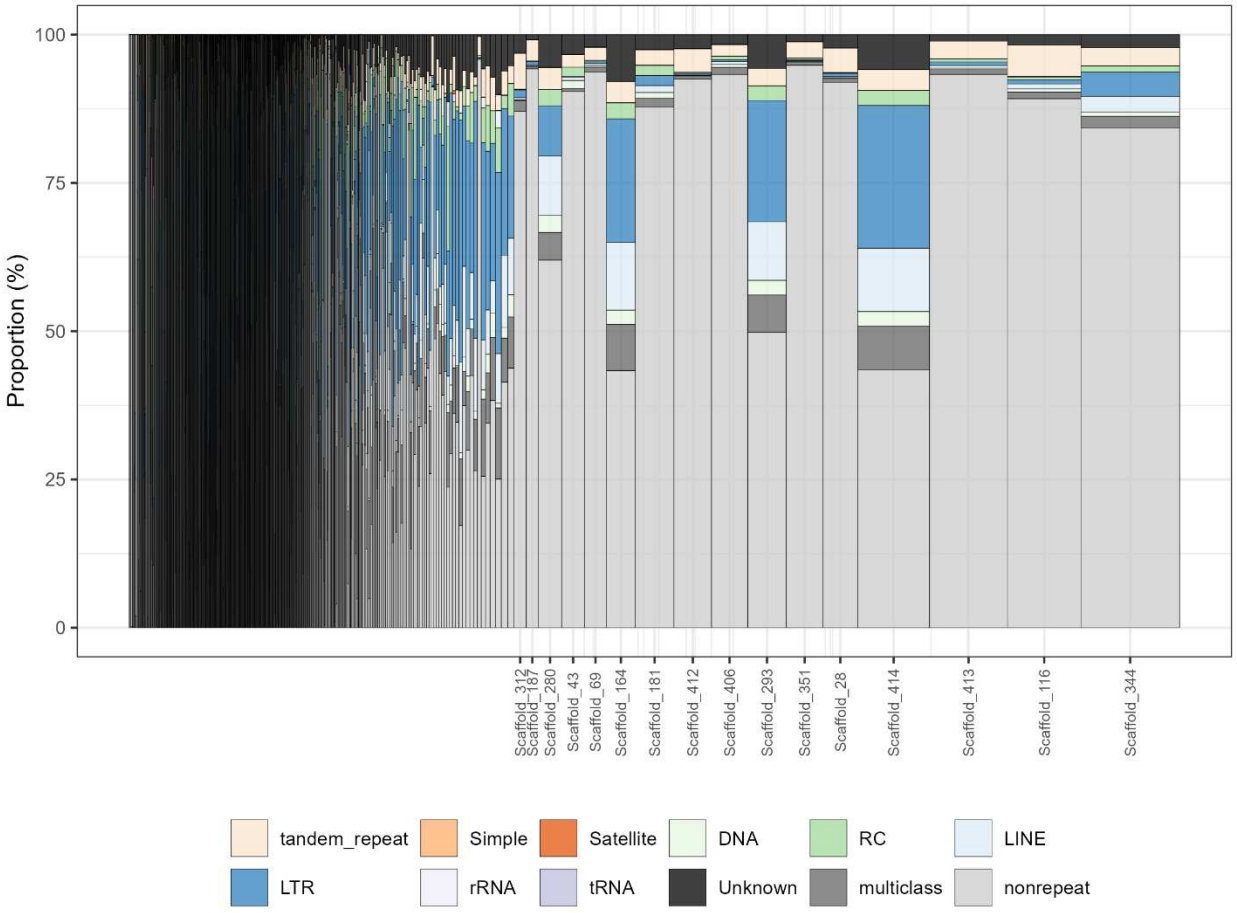


Figure S2. Stacked bar plots showing the distribution of repeat content by scaffolds in *D. prolongata* genome assembly (deduplicated). Widths of bars are proportional to the square root of scaffold/chromosome lengths. Scaffold names are displayed for those of length 1Mb or greater; see Figure S3 for results from smaller scaffolds. Repeat contents are color coded as follows. Low-complexity region: orange palette, DNA transposon: green palette, retrotransposon: blue palette, RNA: purple palette. Abbreviations for each repeat class are as follows. RC: Rolling Circle transposons, LINE: Long-Interspersed Nuclear Element, LTR: Long-Terminal-Repeats retrotransposon, snRNA: small-nuclear RNA, Unknown: unknown class of repeats/transposons, multiclass: sequences belonging to more than one repeat class, nonrepeat: non-repetitive DNA sequence.

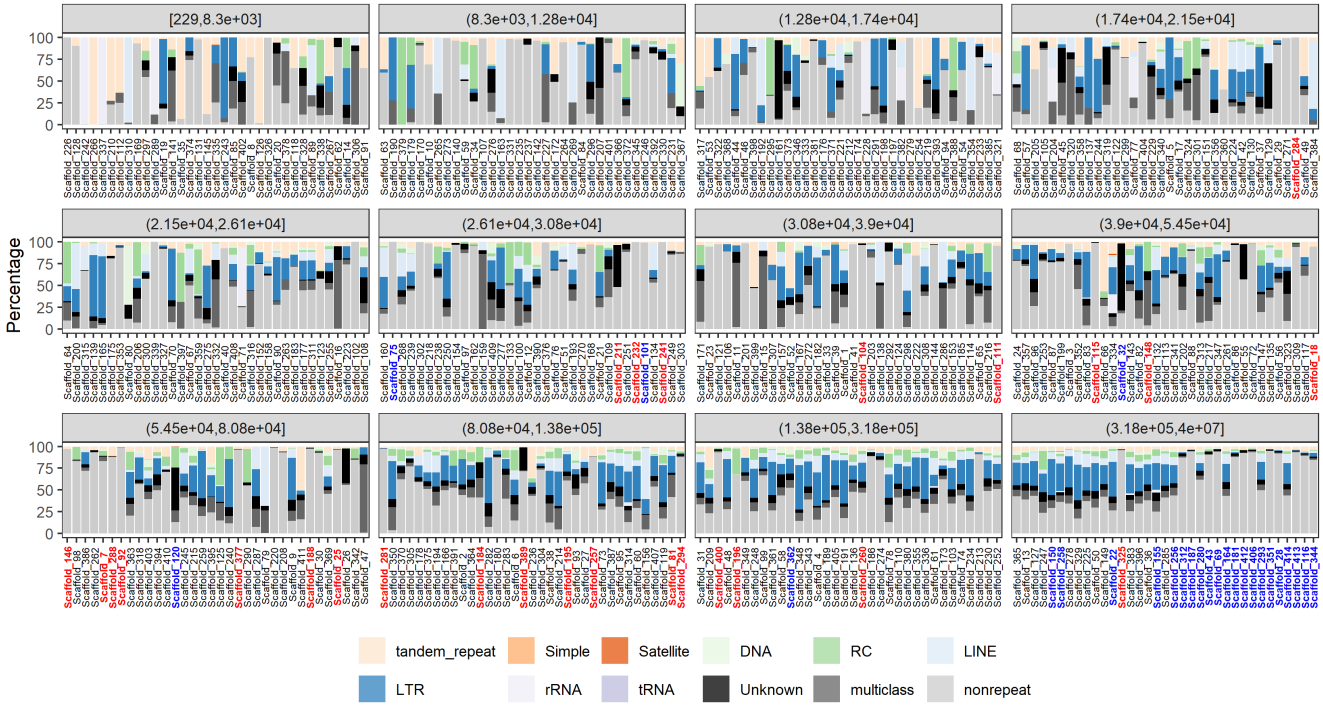
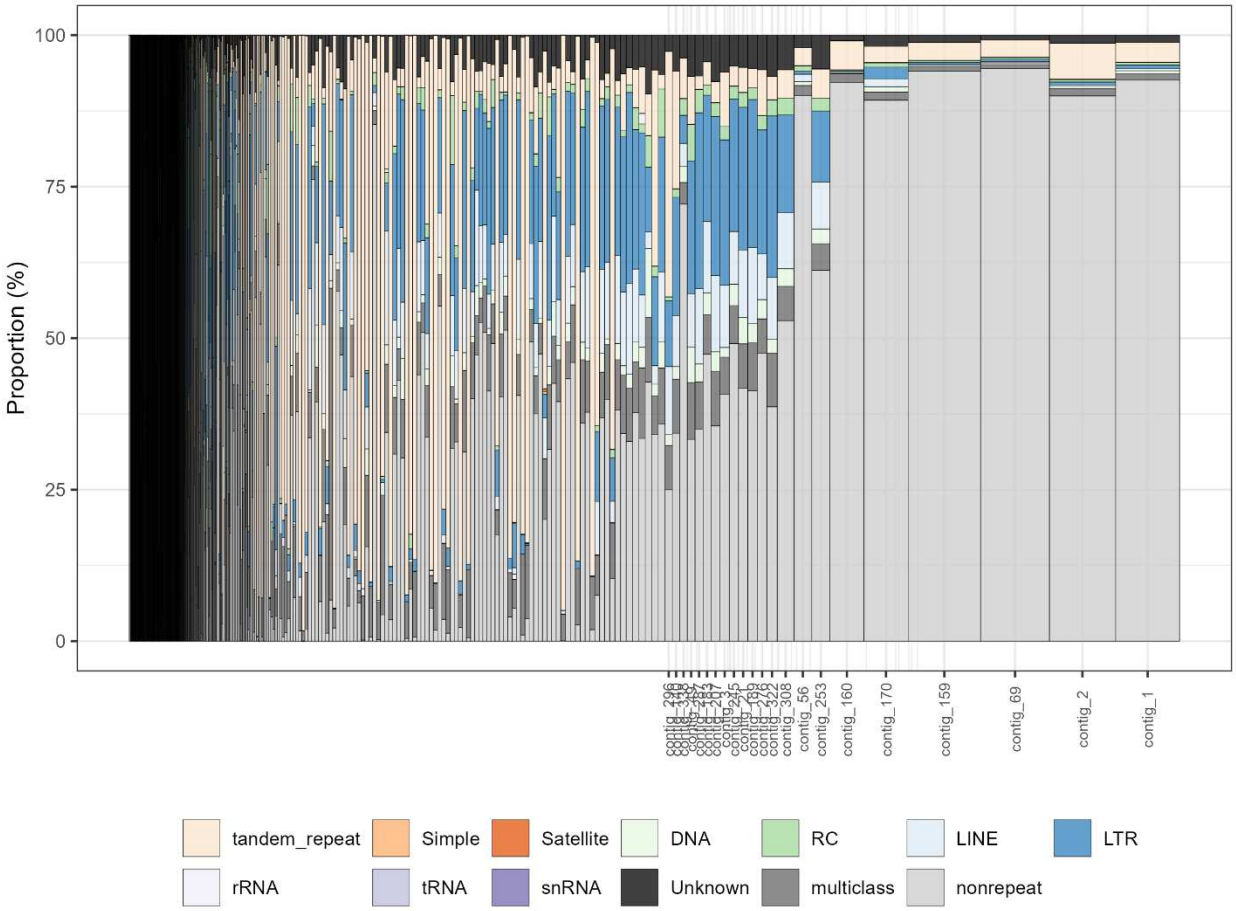


Figure S3. Stacked bar plots showing the distribution of repeat content by scaffolds (partitioned by scaffold length bins) in *D. prolongata* genome assembly. Scaffold names are ordered by their corresponding lengths. Repeat contents are color coded as Fig. S2, with the exception that removed scaffolds have names colored red, and retained members of duplicate scaffold pairs are colored in blue.



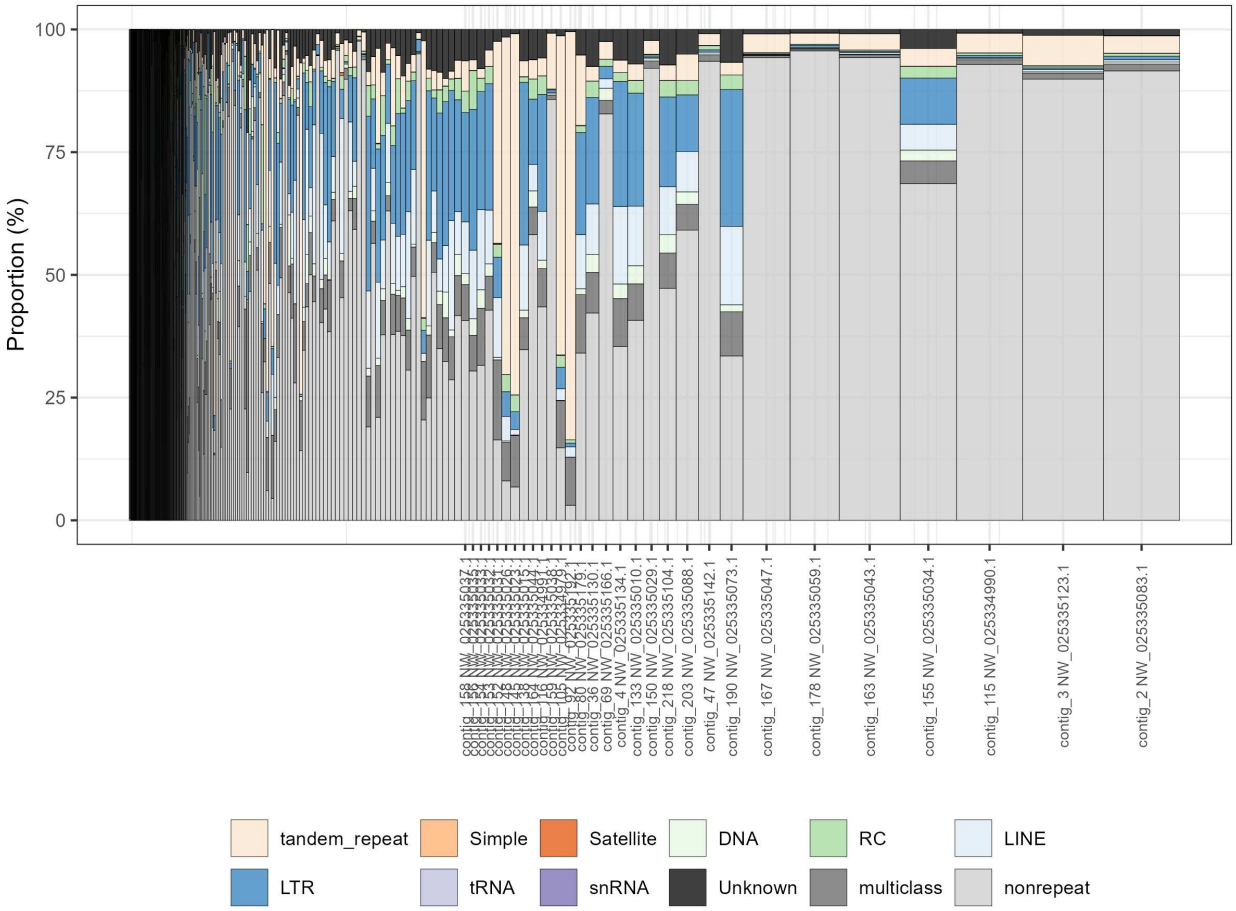
448

449 Figure S4. Stacked bar plots showing the distribution of repeat content by scaffolds in *D. carrolli*

450 genome assembly. Widths of bars are proportional to the square root of scaffold/chromosome lengths.

451 Repeat contents are color coded as Fig. S2.

452



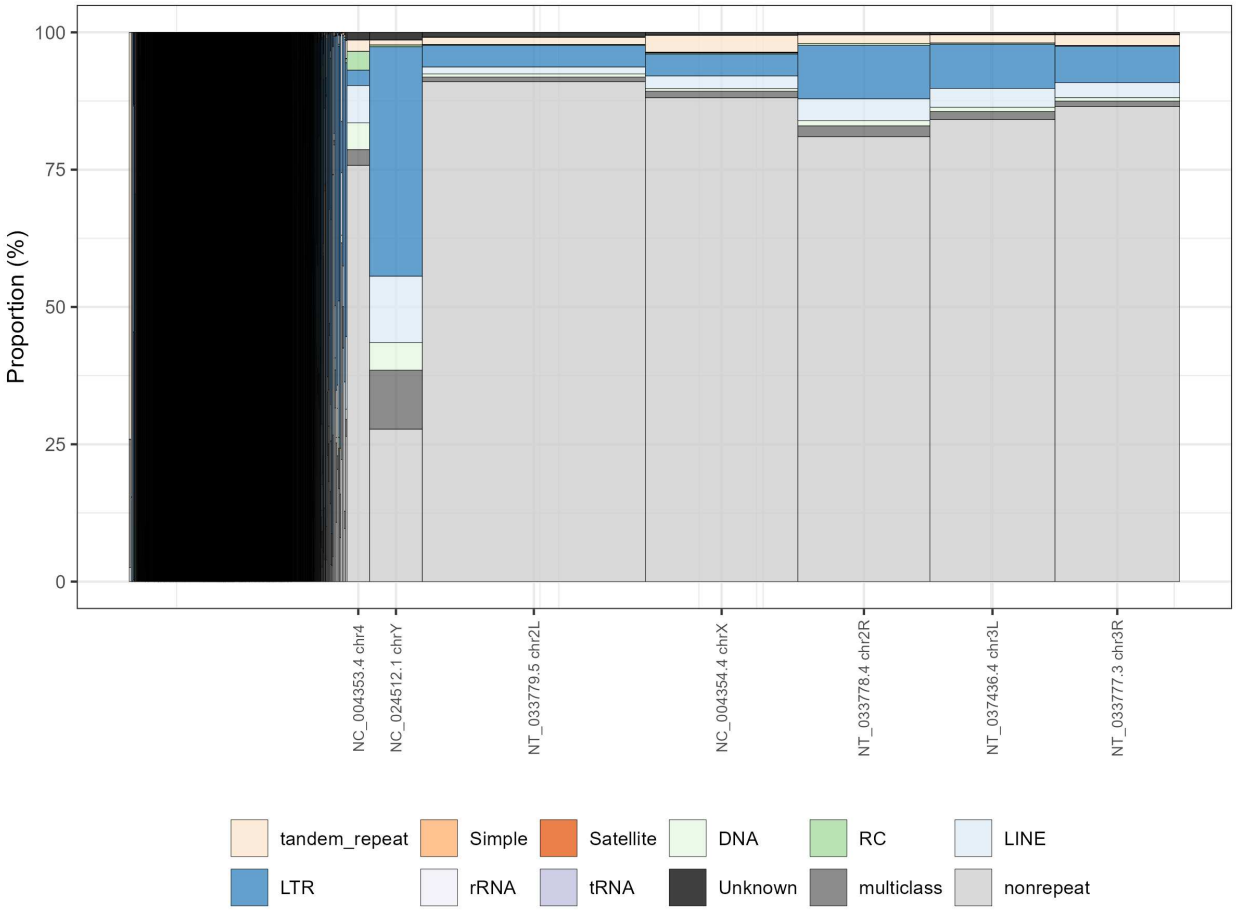
453

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



458

459 Figure S6. Stacked bar plots showing the distribution of repeat content by scaffolds in *D. melanogaster*

460 genome assembly (GCF_000001215.4_Release_6_plus_ISO1). Widths of bars are proportional to the

461 square root of scaffold/chromosome lengths. Repeat contents are color coded as Fig. S2.

462

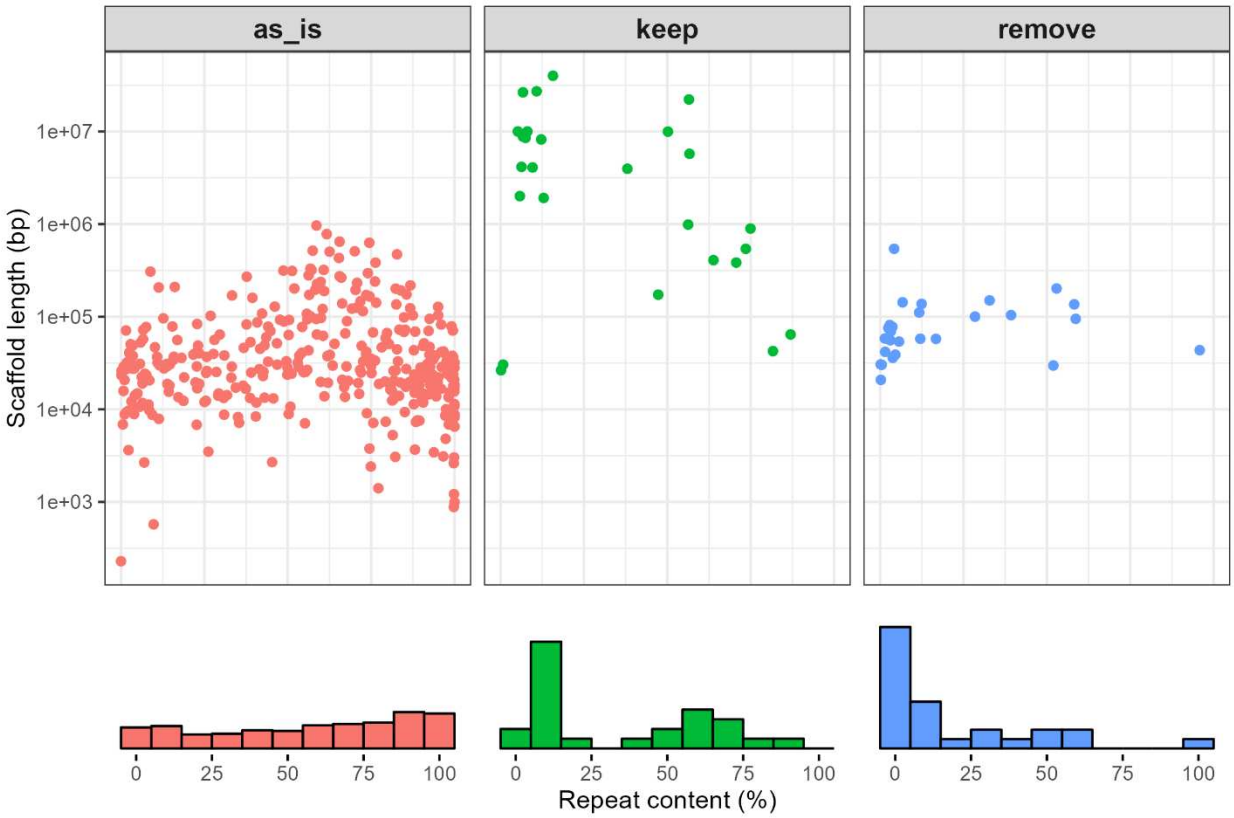


Figure S7. Scatter plots showing the distribution of repeat profiles by scaffolds under each category in the complete *D. prolongata* genome assembly. Frequency histograms of repeat content are displayed at the bottom. X-axis is the repeat content (%), and the y-axis is the corresponding scaffold length. Scaffolds with no BUSCO duplicates are colored in red (as_is), retained scaffolds with BUSCO duplicates in green (keep), and removed scaffolds with BUSCO duplicates in blue (remove).