*Drosophila prolongata* draft genome and annotation

Introduction

*Drosophila prolongata* is a member of the *melanogaster* species group and *rhopaloa* subgroup native to southeast Asia (Singh & Gupta, 1977; Toda, 1991). The species has a suite of recently evolved male-specific morphological traits (Fig. 1), including relative foreleg size, leg pigmentation, wing pigmentation, reversed size sex dimorphism, and an expanded number of leg chemosensory organs (D. Luecke et al., 2022; Luecke & 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 divergence in sex-specific cuticular hydrocarbon profiles (Kudo et al., 2017; Luo et al., 2019; Setoguchi et al., 2014).

The phylogenetic proximity to the model *D. melanogaster* and available genome sequences for underived sister species *D. rhopaloa* and *D. carrolli* make this species a promising system to study the genetics of dimorphic development, physiology, and behavior (Barmina & Kopp, 2007). A reference genome assembly and annotation for *D. prolongata* will benefit such work. Presented here is a highly complete and contiguous assembly generated by Dovetail Genomics (Cantata Bio. LLC, dovetailgenomics.com) based on long-read PacBio sequencing and Hi-C scaffolding, along with annotation 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 was collected in Vietnam by Dr. Hisaki Takamori. Virgin females were collected by isolating adults within four hours of emergence. Four generations of full sibling 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 plain agar food for at least one day to reduce gut microbes, then collected into 1.5mL tubes and flash-frozen in liquid nitrogen. 50 frozen adult male individuals were sent on dry ice to Dovetail Genomics 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, together 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). An initial assembly based on 1.2M PacBio reads was produced using FALCON with Arrow polishing. A second HiRise assembly was generated with additional HiC sequencing and the HiRise software pipeline.

RNA was extracted using TRIzol (Invitrogen). For foreleg RNA, multiplexed stranded cDNA sequencing libraries were prepared using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England BioLabs) using poly(A) isolation magnetic beads. Libraries were sequenced on the Illumina HiSeq4000 platform by the UC Davis Genomics Center. For head RNA, cDNA sequencing libraries were constructed using the TruSeq Stranded RNA Kit (Illumina) and sequenced on the Illumina HiSeq4000 platform by Novogene (<https://www.novogene.com/us-en/>). Transcripts were assembled using Trinity 2.4.0 with default options for stranded data.

Gene prediction and annotation

Homology-based annotations were generated using Liftoff 1.5.1 with minimap2 2.17 alignment based on the *D. melanogaster* GCF000001215.4 release 6, *D. elegans* GCF000224195.1 2.0, and *D. rhopaloa* GCF000236305.1 2.0 annotations downloaded from FlyBase. Liftoff was run with the copies option and percent identity 0.80. Additional gene models were inferred using MAKER 3.01.02 with BLAST 2.11.0 and repeat masker 4.0.7, using EST evidence from the Trinity transcripts assembled based on leg 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, genometools 1.5.9, and custom Python scripts 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). 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 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 using nucmer alignment with minimum exact match 1000bp and mummerplot for visualization. Annotation statistics were found with genometools 1.5.9. Transcripts were extracted from annotations using gffread 0.9.12, and transcript completeness was assessed using the transcriptome mode of BUSCO.

Results and discussion

Assembly contiguity

The Dovetail assembly HiRise scaffolding method (Fig 2) produced an assembly for *D. prolongata* with higher contiguity than the existing *D. rhopaloa* assembly, approaching the contiguity of the latest *D. melanogaster* reference (Table 1) as measured by N50 or N90. Whole genome alignment between the *D. prolongata* assembly and *D. rhopaloa* reference (Fig 3) shows long stretches of high identity spanning nearly all large scaffolds.

|  |  |  |  |
| --- | --- | --- | --- |
| **Assembly** | *prolongata* | *rhopaloa* | *melanogaster* |
| **Total length (bp)** | 223340102 | 193508231 | 143726002 |
| **Scaffolds** | 414 | 228 | 1870 |
| **N50 (bp)** | 22190000 | 15806012 | 25286936 |
| **L50** | 4 | 5 | 3 |
| **GC%** | 40.11% | 39.87% | 41.67% |
| **BUSCO Complete, Single Copy** | 92.4% (3036) | 98.1% (3221) | 98.5% (3235) |
| **BUSCO Complete, Duplicated** | 6.1% (200) | 0.4% (12) | 0.2% (8) |
| **BUSCO Fragmented** | 0.9% (29) | 0.7% (24) | 0.5% (16) |
| **BUSCO Missing** | 0.6% (20) | 0.8% (28) | 0.8% (26) |

Table 1: Statistics for assembly contiguity and completeness of *D. prolongata* assembly alongside 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. melanogaster*. The whole genome alignment between the *D. prolongata* assembly and the *D. rhopaloa* reference (Fig 3) further shows near complete, highly contiguous coverage of the entire reference with regions of *D. prolongata* scaffolds.

Repeat annotation

The *D. prolongata* genome is moderately repetitive, with transposable elements and simple repeats comprising about 13% of its sequence, of which 38.4% are tandem repeats and 28.3% are LTR retrotransposons (Table S1).

|  |  |
| --- | --- |
| Repeat group | % of the genome |
| DNA transposons | 0.45% |
| Non-LTR retrotransposons (LINEs) | 1.53% |
| LTR retrotransposons | 3.68% |
| SINEs | 0% |
| Low complexity and simple repeats | 0.01% |
| Rolling Circle Transposon (RC) | 0.75% |
| Unknown class TE | 1.55% |
| Tandem repeats by TRF | 5.05% |
| Total | 13.02% |

Supplementary Table 1: Repeat content of *D. prolongata* genome assembly

Annotation completeness

Transcripts extracted from the annotation and assembly show a high degree of completeness between the *D. prolongata* annotation. However this annotation does not reach the completeness of the *D. rhopaloa* and especially *D. melanogaster* references (Table 2), both in terms of gene inclusion and completeness of individual gene models. A higher number of BUSCO dipteran benchmarks are missing in the *D. prolongata* assembly (82) compared to the *D. rhopaloa* (15) or *D. melanogaster* (0) references. Additionally, the *D. prolongata* transcripts are shorter than those from the references, and many more BUSCO dipteran benchmark genes are fragmented in the *D. prolongata* assembly (110) than for the references (both 3). These statistics show the limitations of current algorithmic annotation methods and indicate that care should be used when using gene models from this draft of the *D. prolongata* genome annotation. Despite these limitations, the overall completeness is quite high, with 94.2% of BUSCO benchmark genes covered and a reasonable median transcript length. These resources will provide a good foundation for future genetic studies in *D. prolongata* when used with the limitations of draft annotations in mind, and future iterations of the annotation should aim to improve gene model coverage and completeness.

|  |  |  |  |
| --- | --- | --- | --- |
| **Annotation** | *prolongata* | *rhopaloa* | *melanogaster* |
| **Genes** | 19570 | 15463 | 17559 |
| **Protein Coding Genes** | 16383 | 14607 | 13986 |
| **Exons** | 180792 | 154625 | 190719 |
| **Median Transcript Length (bp)** | 1636 | 1995 | 1954 |
| **Longest Transcript (bp)** | 63866 | 65859 | 71382 |
| **BUSCO Complete** | 94.2% (3093) | 99.4% (3267) | 99.9% (3282) |
| **BUSCO Fragmented** | 3.3% (110) | 0.1% (3) | 0.1% (3) |
| **BUSCO Missing** | 2.5% (82) | 0.5% (15) | 0.0% (0) |

Potential regional duplication

The other major caveat for this assembly and annotation is the extent of identified duplication. This stands out most clearly in the *D. prolongata* assembly BUSCO scores, where 200 benchmark single-copy genes were identified as duplicated compared to 12 for *D. rhopaloa* and 8 for *D. melanogaster*. Additional signatures of duplicated regions are also visible. The total length of the draft assembly is higher than either reference (Table 1), as is the total number of genes in the annotation (Table 2). This suggests that some regions are represented more than once. Some regions of potential duplication are also visible in the whole genome alignment (Fig 3). Identifying and removing duplicate scaffolds is a clear avenue of improvement for future iterations of this genomic reference.

Literature Cited

Barmina, O., & Kopp, A. (2007). Sex-specific expression of a HOX gene associated with rapid morphological evolution. *Developmental Biology*, *311*(2), 277–286. https://doi.org/10.1016/j.ydbio.2007.07.030

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*, *117*(17), 9451–9457. https://doi.org/10.1073/pnas.1921046117

Kudo, A., Shigenobu, S., Kadota, K., Nozawa, M., Shibata, T. F., Ishikawa, Y., & Matsuo, T. (2017). Comparative analysis of the brain transcriptome in a hyper-aggressive fruit fly, Drosophila prolongata. *Insect Biochemistry and Molecular Biology*, *82*, 11–20. https://doi.org/10.1016/j.ibmb.2017.01.006

Luecke, D. M., & Kopp, A. (2019). Sex-specific evolution of relative leg size in Drosophila prolongata results from changes in the intersegmental coordination of tissue growth. *Evolution*, *73*(11), 2281–2294. https://doi.org/10.1111/evo.13847

Luecke, D., Rice, G., & Kopp, A. (2022). Sex-specific evolution of a Drosophila sensory system via interacting cis- and trans-regulatory changes. *Evolution & Development*, *24*(1–2), 37–60. https://doi.org/10.1111/ede.12398

Luo, Y., Zhang, Y., Farine, J.-P., Ferveur, J.-F., Ramírez, S., & Kopp, A. (2019). Evolution of sexually dimorphic pheromone profiles coincides with increased number of male-specific chemosensory organs in Drosophila prolongata. *Ecology and Evolution*, *9*(23), 13608–13618. https://doi.org/10.1002/ece3.5819

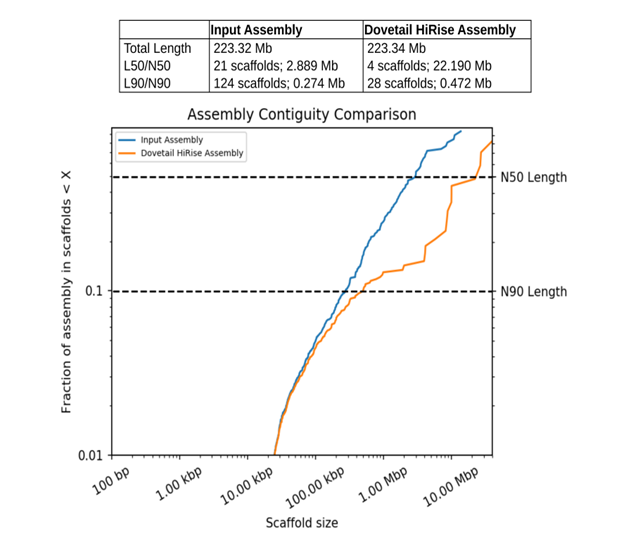
Setoguchi, S., Takamori, H., Aotsuka, T., Sese, J., Ishikawa, Y., & Matsuo, T. (2014). Sexual dimorphism and courtship behavior in Drosophila prolongata. *Journal of Ethology*, *32*(2), 91–102. https://doi.org/10.1007/s10164-014-0399-z

Singh, B. K., & Gupta, J. P. (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*(1–2), pp. 31–38.

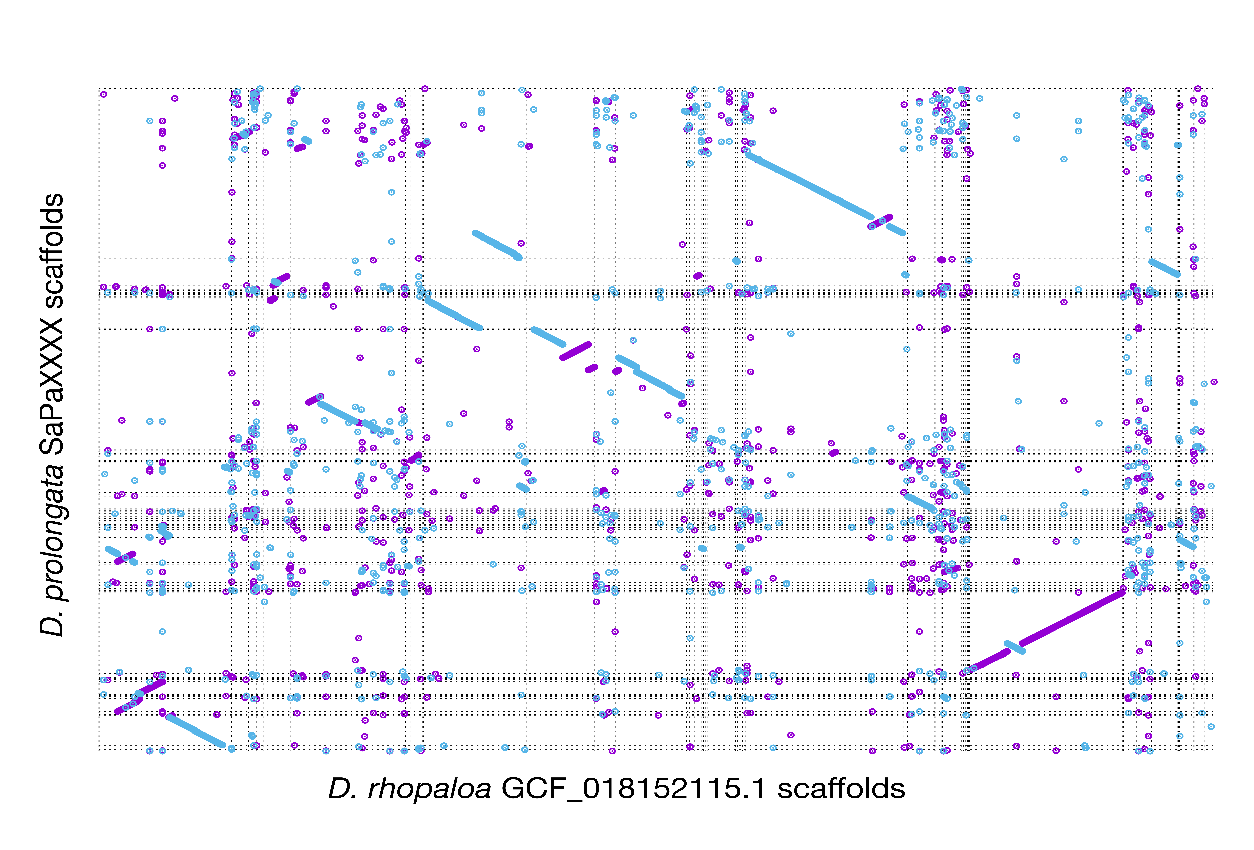
Toda, M. J. (1991). Drosophilidae (Diptera) in Myanmar (Burma) VII. The Drosophila melanogaster species-group, excepting the D. montium species-subgroup. *Oriental Insects*, *25*(1), 69–94. https://doi.org/10.1080/00305316.1991.10432216



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

**

2the a

**

3