Long-read genome assembly of the Japanese parasitic wasp *Copidosoma floridanum* (Hymenoptera: Encyrtidae)

Running title: Genome assembly of Japanese *C. floridanum*

Kouhei Toga [1,2], Takuma Sakamoto [3], Miyuki Kanda [1,4], Keita Tamura [1], Keisuke Okuhara [2,4], Hiroko Tabunoki [3,5], Hidemasa Bono [1,2]*

[1]Laboratory of BioDX, PtBio Co-Creation Research Center, Genome Editing Innovation Center, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan

[2]Laboratory of Genome Informatics, Graduate School of Integrated Sciences for Life, Hiroshima University, 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan

[3]Department of Science of Biological Production, Graduate School of Agriculture, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo, 183-8509, Japan

[4]Research and Development Department, PtBio Inc., 3-10-23 Kagamiyama, Higashi-Hiroshima city, Hiroshima, 739-0046, Japan

[5]Cooperative Major in Advanced Health Science, Graduate School of Bio-Applications and System Engineering, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo, 183-8509, Japan

* Correspondence:
Hidemasa Bono
Email: bonohu@hiroshima-u.ac.jp
Tel.: +81 82 424 4013

Abstract

Copidosoma floridanum is a cosmopolitan species and an egg-larval parasitoid of the Plusiine moth. *C. floridanum* has a unique development mode called polyembryony, in which thousands of genetically identical embryos are produced from a single egg. Some embryos develop into sterile soldier larvae, and their developmental patterns differ between the US and Japanese *C. floridanum* strains. Genome sequencing can accelerate our understanding of the molecular bases underlying polyembryony, including the production of soldier castes. However, only the genome sequence of the US strain has been reported. In the present study, we determined the genome sequence of the Japanese strain using Pacific Biosciences high-fidelity reads and generating a highly contiguous assembly (552.7 Mb, N50: 17.9 Mb). Gene prediction and annotation identified 13,886 transcripts derived from 10,786 gene models. Among these, 149 transcripts were not identified in the US strain gene model. Differences in the amino acid sequence of the *vasa* gene were observed between the US and Japanese strains. The genome assemblies constructed in this study will facilitate our understanding of intraspecific variation in the developmental patterns of *C. floridanum*.

Keywords: *Copidosoma floridanum*, Pacific Biosciences high-fidelity reads, genome assembly

Introduction

Copidosoma floridanum (Hymenoptera: Encyrtidae) is a worldwide distributed egg-larval parasitoid of the Plusiinae moth (Lepidoptera, Noctuidae, Plusiinae)[1]. This species reproduces by polyembryony, producing >2,000 genetically identical individuals from a single egg[2]. Polyembryony has evolved independently in the four Hymenoptera families (Encyrtidae, Platygastridae, Braconidae, and Dryinidae); however, encyrtid wasps have the most extreme form regarding clone size[3]. In some encyrtid wasps, the embryos develop into sterile soldier larvae with large mandibles[4,5]. The phenomenon of solitary individuals forming cooperative groups with a reproductive division of labor can be considered one of the major evolutionary events that brought complexity to life on Earth[6,7].

The process and molecular basis of soldier development in *C. floridanum* have been described and investigated in the US and Japanese strains[2,8,9]; however, the development and behavioral patterns of male soldiers differed between the two strains. In the US strain, male soldiers appear in the last larval stage of the host[10], whereas in the Japanese strain, they appear in the first- or second-instar larvae[11]. Male soldiers were functional in the Japanese strain, but were not as aggressive as the female soldiers, whereas male soldiers in the US strain were nonaggressive[11,12]

Each strain optimizes its behavior and developmental patterns in its habitat, which is possibly related to genomic differences. Genome or transcriptome analyses can accelerate our understanding of the molecular basis underlying polyembryony, including the production of soldier castes[9]. *vasa* is a gene that has been compared between the Japanese and the US strains, and a 10 amino acid sequence difference in the N-terminal region was observed[13]. Although *vasa*, encoding a DEAD-box RNA helicase protein[14,15], plays a central role in soldier development[16,17], the molecular determinants that cause variations in soldier development or behavioral patterns between strains remain unknown. The genome of the US strain has been sequenced using short-read sequencing[18], while that of the Japanese strain has not. Long-read sequencing offers several advantages over short-

read sequencing[19]. High-quality (contiguity and completeness of annotation) insect genomes have recently been achieved using long-read sequencing[20]. Single-molecule real-time (SMRT) sequencing is a long-read sequencing method developed by Pacific Biosciences (PacBio)[21]. In particular, circular consensus sequencing generates highly accurate long high-fidelity reads (HiFi reads, >10 kb, >99%)[22–24]. HiFi reads are also more suitable than other long- or short-read methods for identifying genes containing repetitive sequences[25,26].

Therefore, in this study, we aimed to construct a highly contiguous genome assembly of the Japanese *C. floridanum* using PacBio long-read sequencing.

Materials and Methods

DNA extraction and genome sequencing

Genomic DNA was extracted from adult *C. floridanum* males using the Blood & Cell Culture DNA Maxi Kit (Qiagen Co. Ltd., Valencia, CA, USA). The genetic system of *C. floridanum* and other hymenopteran insects is haplodiploid; females develop from fertilized eggs (diploid), and males develop from unfertilized eggs (haploid). The sequence library was prepared using the SMRTbell Express Template Preparation Kit 2.0 or SMRTbell Prep Kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA). Libraries were sequenced using the Sequel II or IIe (Pacific Biosciences).

De novo assembly and quality assessment

Subreads and HiFi reads generated from the PacBio sequencers were assembled using Flye v 2.9[27] and Hifiasm v 0.16.1, respectively, using the '-primary' option[28]. Quast v 5.2.0 was used to assess the genome assembly statistics[29]. Gene set completeness was assessed using the Benchmark of Universal Single-Copy Orthologs (BUSCOs) v 5.2.2, with '-l insecta_odb10 (2020-09-10)'[30]. k-mer-based validation was performed using Merqury v1.3 and Meryl db (k=21)[31] for the assembly

constructed with HiFi reads. Contigs were assigned to taxonomic classifications using BlobTools v 1.1[32] with the BLAST (BLASTN) program in the National Center for Biotechnology Information BLAST software package (v 2.14.0) against the NCBI non-redundant nucleotide sequence database (nt), prepared using ncbi-blast-dbs v 10.5.0 (https://github.com/yeban/ncbi-blast-dbs, accessed on May 18, 2023). minimap2 v 2.26-r1175[33] was used to obtain mapping coverage.

Dot plots were visualized using the dotPlotly program with '-s -t -m 10000 -q 25000 -l' (https://github.com/tpoorten/dotPlotly, accessed on July 17, 2023) to compare the genomic differences between the US and Japanese strains. The 1 Mb or longer contigs were extracted using SeqKit v 2.0.0, aligned by a nucmer with default settings, and delta files were filtered using a delta-filter with '-q -r' options in MUMmer4 v4.0.0rc[34].

Gene prediction and annotation

Gene prediction was performed using the BRAKER2 and BRAKER3 databases[35,36]. Public RNA sequencing (RNA-Seq) data (molura: DRR138914, DRR138915, DRR138916, and DRR138917; male adult: SRR1864697) and 11 OrthoDB protein sets of Arthropoda[37] (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/, accessed on June 17, 2023) were used as extrinsic evidence data to perform the BRAKER pipeline. The RNA-Seq reads were processed using fastp v 0.23.2 with default settings[38] and mapped to the *de novo* assembled genome using HISAT2 v 2.2.1.[39]. The generated SAM files were converted to BAM files using SAMtools v 1.12,[40] and the BAM files were imported together when the BRAKER pipeline was run. Repetitive assembly sequences were soft-masked using RepeatModeler v 2.0.3 and RepeatMasker v 4.0.6 in Extensive de novo TE Annotator v 2.1.3[41]. The soft-masked genome assembly was inputted into the BRAKER pipeline. GFFread v0.12.1[42] was used to convert the transcript or protein sequences from the GTF file created by

BRAKER. Protein sequences were assessed using BUSCO (protein mode) with '-l insecta_odb10' (2020-09-10).

Functional annotation

The functional annotation workflow of Fanflow4Insects[43] was modified and used for protein sequences, as described below. The protein sequences were aligned against reference datasets of *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *C. floridanum* (US strain), and UniProtKB/Swiss-Prot using GGSEARCH v 36.3.8g in the FASTA package (https://fasta.bioch.virginia.edu/, accessed on July 19 2023). Protein sequences were also searched using HMMSCAN in HMMER v 3.3.2. Pfam v 35.0 was used as the protein domain database for HMMSCAN[44]. Enrichment analysis was performed using Metascape software (https://metascape.org/, accessed on July 19, 2023) with default settings[45].

Results and Discussion

***De novo* genome assembly of Japanese *C. floridanum* and the quality assessment**

This study generated HiFi reads and subreads using Japanese strains. The assembly quality between the Japanese strain (Cflo-J) and the US strain (Cflo-US) was evaluated. Cflo_2.0 (NCBI RefSeq assembly GCF_000648655.2) was used for the genome assembly of Cflo-US. HiFi reads were assembled, and the primary contig comprised 149 contigs with an N50 length of 17.9 Mb and a total assembly size of 552,713,363 bp (50.8-fold coverage; Table 1 and Supplementary Table S1). Hifiasm also produced alternate contigs (1,162 contigs, Supplementary Table S1). The subread contig comprised 996 contigs with an N50 length of 4.6 Mb and a total assembly size of 534,638,157 bp (244.7-fold coverage, Table 1 and Supplementary Table S1). The N50 value of Cflo-J assemblies was much higher than that of Cflo-US (1.2 M). However, the GC content of the assemblies was

similar. The BUSCO completeness of Cflo-J was also comparable to that of Cflo-US; however, the fragmentation score was lower for the Cflo-J assemblies (Table 1). The N50 values and number of contigs were the best for the HiFi assembly (Table 1 and Figure 1A). We confirmed the differences in genomic structures among all the assemblies using dot-plot visualization (Figures 1B and C). Linear plots were obtained, confirming the similarity of the genomic structures across the entire assembly length (Figure 1B and C). The matching coordinate data for the dot-plot comparison are available in Supplementary Files 1 and 2. We then evaluated the HiFi assemblies of Cflo-J, which had the highest contiguity.

Hifiasm generated primary and alternate contigs (Supplementary Table S1). K-mer-based evaluation showed high base accuracy and completeness of primary contigs (a quality value (QV) = 68.08 and completeness = 97.6%, Supplementary Table S2). The Merqury spectrum plot showed a single k-mer peak in the primary contigs, indicating extremely low heterozygosity (Supplementary Figure S1A, red line). Because *C. floridanum* males possess a haploid genome, the single k-mer peak of the primary contig was reasonable. We also confirmed the presence of reads included only in the alternate contigs (Supplementary Figure S1A, blue line), suggesting that the genome assembly constructed in this study may contain heterozygous regions.

Taxonomic assignments of the contigs were performed to reveal the taxonomic origin of the contigs. The primary and alternate contigs were assigned to Arthropoda, and the remaining contigs were either no-hit or chordata (Supplementary Figures S1B and C). The mapping rate of no-hit was remarkably low (primary contig, 0.27%; alternate contig, 0.51%), and that of chordata in alternate contigs was also extremely low (0.05%). The assembly of Cflo-US also comprises Arthropoda and no-hit contigs (ENA https://www.ebi.ac.uk/ena/browser/view/GCA_000648655.2, accessed on September 20, 2023). An ideal haploid representation was observed in the primary contigs, and no contig contamination from non-arthropods was observed.

Annotation of the primary contig

As repetitive sequences may affect gene prediction accuracy[46], these sequences in the HiFi primary contig were investigated using RepeatModeler and RepeatMasker. The repeat sequence was 285.4 Mb, corresponding to 51.65% of the HiFi primary contig, and "Unclassified" accounted for a substantial percentage of the total (34.97%) (Supplementary Table S3). Gene prediction was performed using several RNA-Seq data and protein dataset combinations. The BRAKER3 using RNA-Seq data and the Arthropoda protein dataset showed the best BUSCO completeness (96.4%, Table 2). The numbers of predicted genes and transcripts are listed in Table 2. BRAKER2 predicted more genes compared to BRAKER3 (Table 2). The gene model with the highest BUSCO completeness (96.4%) ("best gene model") predicted 10,786 protein-coding genes and 13,886 transcripts. The annotated GFF files are shown in Supplementary File 3. The best gene model was annotated using Fanflow4Insects[43]. Table 3 shows the results of the searches using Fanflow4Insects for each reference dataset. Of the 10786 gene models, 10585 were annotated using GGSEARCH or HMMSCAN. The full annotation results for GGSEARCH and HMMSCAN are presented in Supplementary Table S4.

We evaluated whether annotation differences existed between the Cflo-US assembly and Cflo-J HiFi primary contigs. In the best gene model, 149 transcripts with no corresponding proteins were observed in the Cflo-US dataset; however, they were similar to other reference datasets (Supplemental Table S5). Human gene IDs corresponding to 149 transcripts were used for enrichment analysis (Figure 2A). "Metabolism of carbohydrates" was the most significantly enriched Gene Ontology (GO) term. In addition, GO terms for amino acid metabolism and translation were enriched. We successfully complemented the annotation of genes necessary for maintaining life activities.

Ohno et al. (2019)[13] showed that the amino acid sequence of *vasa* differed between Cflo-J and Cflo-US. The amino acid sequences of Cflo-US (XP_014219851.1) and Cflo-J *vasa* (g8014.t1), including the amino acid sequence (BBI30140.1) determined by Ohno et al. (2019), were compared. We confirmed the differences in amino acid sequences at the N-terminal region between Cflo-US *vasa* (XP_014219851.1) and Cflo-J *vasa* (g8014.t1), as previously reported[13] (Figure 2B, red arrows). Within the Japanese strain (BBI30140.1 vs. g8014.t1), the amino acid sequences were almost identical (Figure 2B).

Polyembryony has mostly been reported in Hymenoptera insects[47]. Soldier development, regulated by *vasa* expression, has been reported in *C. floridanum*. We also investigated the genomic position of *vasa* (Figure 2C). *vasa* was located in an intron between the seventh and eighth exons of *cap'n'collar* (*cnc*), and the directions of transcription were opposite (Figure 2C, blue and green arrows). This positional relationship was also confirmed in Cflo-US (NCBI Genome Browser: Gene: LOC106647823, Location:scaffold6|NW_019379589.1:2,626,762 – 2,630,760.). Furthermore, the *vasa* genomic positions in other insects were preliminarily examined (Supplementary Table S6), and the positional relationship between *vasa* and *cnc* may be conserved in Hymenoptera. Soldier development observed in *C. floridanum* may be related to the conserved positional relationship between *vasa* and *cnc*, but the actual regulatory relationship is unknown. *cnc* is related to the normal arrangement of oocyte nuclei in *Drosophila melanogaster*[48] and is thus a gene involved in embryonic development. *cnc* may also be involved in soldier development associated with *vasa* expression.

In conclusion, we obtained a highly contiguous Japanese *C. floridanum* genome assembly. This genome assembly contributes to understanding the molecular basis of intraspecific variation in *C. floridanum* polyembryony including soldier development. Furthermore, parasitic wasps can control the physiological state of host insects, and their use for insect control is expected[49].

Therefore, the genome assembly and annotation generated in this study contribute to our understanding of the molecular basis for pest-insect control.

Funding

This study was supported by the Center of Innovation for Bio-Digital Transformation (BioDX), an open innovation platform for industry-academia co-creation of JST (COI-NEXT, grant number JPMJPF2010 to H.T. and H.B.), and JSPS KAKENHI (grant number 23K17418 to H.T. and H.B.).

Acknowledgments

We would like to thank all laboratory members at Hiroshima University for their valuable comments. Computations were performed on the computers at the Hiroshima University Genome Editing Innovation Center. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Authors' contributions

Conceptualization: H.T. and H.B.; methodology: Kouhei T. and H.B.; software: Kouhei T. Formal analysis, data curation, and visualization: Kouhei T., Keita T., and T.S. Validation: Kouhei T., and H.B. Investigation: Kouhei T., and H.B. Resources: Kouhei T., H.B., and H.T. Writing—original draft: Kouhei T. Writing—review and editing: T.S., M.K, Keita T., K.O., H.T., and H.B. Supervision: H.T., and H.B. Project administration: H.B. Funding acquisition: H.T., and H.B.

Conflict of interest

The authors declare no conflict of interest.

Data availability

All sequencing data (assembled sequences and raw sequence reads) were deposited in the DDBJ under the umbrella BioProject accession number PRJDB16642. Genome assembly from primary contigs was deposited in the DDBJ under the accession numbers BTPR01000001−BTPR01000149. Alternate contigs were deposited in DDBJ under the accession numbers BTPQ01000001−BTPQ01001162. Raw sequence reads were deposited in the DDBJ under accession number DRA017002 (PacBio HiFi reads) and DRA017125 (PacBio Subreads). Gene annotation and functional annotation of protein-coding genes are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v2). Supplementary Files are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v2).

References

1. Guerrieri, E., and Noyes, J. 2005, Revision of the European species of *Copidosoma* Ratzeburg (Hymenoptera: Encyrtidae), parasitoids of caterpillars (Lepidoptera). Syst Entomol, 30, 97–174.

2. Ode, P. J., and Strand, M. R. 1995, Progeny and Sex Allocation Decisions of the Polyembryonic Wasp *Copidosoma floridanum*. Source: Journal of Animal Ecology.

3. Segoli, M., Harari, A. R., Rosenheim, J. A., Bouskila, A., and Keasar, T. 2010, September, The evolution of polyembryony in parasitoid wasps. J Evol Biol, pp. 1807–19.

4. Strand, M. R. 2003, Polyembryony Carde R., and Resch V., (eds.),. Academic Press, San Diego.

5. Cruz, Y. P. 1979, 0028-0836 I 81 I 490446-02 so 1.oo Nature Vol. Bos, R. J. Air Pollut. Control Ass. Springer.

6. Szathmáry, E., and Smith, J. M. 1995, The major evolutionary transitions. Nature, 374, 227–32.

7. Whyte, B. A. 2021, The weird eusociality of polyembryonic parasites. Biol Lett, 17.

8. Smith, M. S., Shirley, A., and Strand, M. R. 2017, *Copidosoma floridanum* (Hymenoptera: Encyrtidae) Rapidly Alters Production of Soldier Embryos in Response to Competition. Ann Entomol Soc Am, 110, 501–5.

9. Sakamoto, T., Nishiko, M., Bono, H., et al. 2020, Analysis of molecular mechanism for acceleration of polyembryony using gene functional annotation pipeline in *Copidosoma floridanum*. BMC Genomics, 21, 152.

10. Grbic, M., Rivers, D., and Strand, M. R. 1997, Caste formation in the polyembryonic wasp *Copidosoma floridanum* (Hymenoptera: Encyrtidae): in vivo and in vitro analysis. J Insect Physiol, 43, 553–65.

11. Uka, D., Takahashi-Nakaguchi, A., Yoshimura, J., and Iwabuchi, K. 2013, Male soldiers are functional in the Japanese strain of a polyembryonic wasp. Sci Rep, 3, 2312.

12. Giron, D., Harvey, J. A., Johnson, J. A., and Strand, M. R. 2007, Male soldier caste larvae are non-aggressive in the polyembryonic wasp Copidosoma floridanum. Biol Lett, 3, 431–4.

13. Ohno, H., Sakamoto, T., Okochi, R., et al. 2019, Apoptosis-mediated vasa down-regulation controls developmental transformation in Japanese *Copidosoma floridanum* female soldiers. Dev Biol, 456, 226–33.

14. Hay, B., Yeh Jan, L., and Nung Jan, Y. 1988, A Protein Component of Drosophila Polar Granules Is Encoded by vasa and Has Extensive Sequence Similarity to ATP-Dependent Helicases. Cell.

15. Lasko, P. F., and Ashburner, M. 1988, The product of the Drosophila gene vasa is very similar to eukaryotic initiation factor-4A. 51. Southern, E. J. molec. Biol. Oxford University Press.

16. Vladimir Zhurov, Tomislav Terzin, and Miodrag Grbic. 2004, Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature, 432, 761–4.

17. Donnell, D. M., Corley, L. S., Chen, G., and Strand, M. R. 2004, Caste determination in a polyembryonic wasp involves inheritance of germ cells. Proc Natl Acad Sci U S A, 101, 10095–100.

18. i5K Consortium. 2013, The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered, 104, 595–600.

19. Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., and Sandhu, M. S. 2018, Long reads: their purpose and place. Hum Mol Genet, 27, R234–41.

20. Hotaling, S., Sproul, J. S., Heckenhauer, J., et al. 2021, Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. Genome Biol Evol, 13.

21. Eid, J., Fehr, A., Gray, J., et al. 2009, Real-Time DNA Sequencing from Single Polymerase Molecules. Science (1979), 323, 133–8.

22. Rhoads, A., and Au, K. F. 2015, PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics, 13, 278–89.

23. Ardui, S., Ameur, A., Vermeesch, J. R., and Hestand, M. S. 2018, Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res, 46, 2159–68.

24. Wenger, A. M., Peluso, P., Rowell, W. J., et al. 2019, Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol, 37, 1155–62.

25. Kawahara, A. Y., Storer, C. G., Markee, A., et al. 2022, Long-read HiFi sequencing correctly assembles repetitive heavy fibroin silk genes in new moth and caddisfly genomes. GigaByte, 2022, 1–14.

26. Hotaling, S., Wilcox, E. R., Heckenhauer, J., Stewart, R. J., and Frandsen, P. B. 2023, Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genomics, 24, 117.

27. Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019, Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol, 37, 540–6.

28. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., and Li, H. 2021, Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods, 18, 170–5.

29. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013, QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29, 1072–5.

30. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., and Zdobnov, E. M. 2021, BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol, 38, 4647–54.

31. Rhie, A., Walenz, B. P., Koren, S., and Phillippy, A. M. 2020, Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol, 21, 245.

32. Laetsch, D. R., and Blaxter, M. L. 2017, BlobTools: Interrogation of genome assemblies. F1000Res, 6, 1287.

33. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34, 3094–100.

34. Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. 2018, MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol, 14, e1005944.

35. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., and Borodovsky, M. 2021, BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform, 3.

36. Gabriel, L., Brůna, T., Hoff, K. J., et al. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA.

37. Kuznetsov, D., Tegenfeldt, F., Manni, M., et al. 2023, OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res, 51, D445–51.

38. Chen, S. 2023, Ultrafast one□pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta, 2.

39. Kim, D., Paggi, J. M., Park, C., Bennett, C., and Salzberg, S. L. 2019, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol, 37, 907–15.

40. Danecek, P., Bonfield, J. K., Liddle, J., et al. 2021, Twelve years of SAMtools and BCFtools. Gigascience, 10.

41. Ou, S., Su, W., Liao, Y., et al. 2019, Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol, 20, 275.

42. Pertea, G., and Pertea, M. 2020, GFF Utilities: GffRead and GffCompare. F1000Res, 9, 304.

43. Bono, H., Sakamoto, T., Kasukawa, T., and Tabunoki, H. 2022, Systematic Functional Annotation Workflow for Insects. Insects, 13, 586.

44. Mistry, J., Chuguransky, S., Williams, L., et al. 2021, Pfam: The protein families database in 2021. Nucleic Acids Res, 49, D412–9.

45. Zhou, Y., Zhou, B., Pache, L., et al. 2019, Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun, 10, 1523.

46. Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. 2019, Whole-Genome Annotation with BRAKER. Methods Mol Biol, 1962, 65–95.

47. Craig, S. F., Slobodkin, L. B., Wray, G. A., and Biermann, C. H. Thèparadox' of polyembryony: A review of the cases and a hypothesis for its evolution.

48. Guichet, A., Peri, F., and Roth, S. 2001, Stable anterior anchoring of the oocyte nucleus is required to establish dorsoventral polarity of the drosophila egg. Dev Biol, 237, 93–106.

49. Beckage, N. E., and Gelman, D. B. 2004, Wasp Parasitoid Disruption of Host Development: Implications for New Biologically Based Strategies for Insect Control. Annu Rev Entomol, pp. 299–330.

Tables

Table 1 Comparison of assembly statistics with the current reference genomes

	parameter	Cflo-J HiFi (primary)	Cflo-J subread	Cflo-US (Cflo2.0)
assembly statistics	Length	552,713,363	534,638,157	503,815,983
	N50	17,879,143	4,591,336	1,210,516
	L50	9	29	130
	No. of contigs	149	996	4,840
	Largest contig	42,954,611	26,880,362	8,149,722
	GC %	35.62	35.78	35.50
BUSCO results (OrthoDB v10 Insecta database)	Complete (single, duplicated)	97.9 % (96.4%, 1.5%)	98.1% (96.9%,1.2%)	96.7% (95.0%,1.7%)
	Fragmented	0.6%	0.7%	1.5%
	Missing	1.5%	1.2%	1.8%

Table 2 BUSCO analysis for the protein set predicted by BRAKER

	BRAKER3 with RNA seq + Arthropoda protein	BRAKER2 with RNA seq	BRAKER3 with Arthropoda protein	BRAKER2 with Arthropoda protein
Complete (single, duplicated)	96.4 % (79.5%, 16.9%)	94.0% (73.9%, 20.1%)	93.1% (78.8%, 14.3%)	92.5% (82.8%, 9.7%)
Fragmented	0.5%	2.4%	0.7%	1.8%
Missing	3.1%	3.6%	6.2%	5.7%
No. of genes	10786	18531	9913	22781
No. of transcripts	13886	19986	12370	24487

Table 3 Summary of functional annotation results

Annotation category	Reference	The number of gene count
Gene annotation by top hit (GGSEARCH)	Human	8467
	Mouse	8291

	C. elegans	7061
	fly	8403
	Copidosoma floridanum	9850
	UniProtKB/Swiss-Prot	9038
	At least one of the above	10537
Genes annotated only by HMMSCAN	Pfam domain database	48
All genes annotated at protein level	At least one of the above	10585
Hypothetical protein	No hit	201
Total		10786

364

Figures

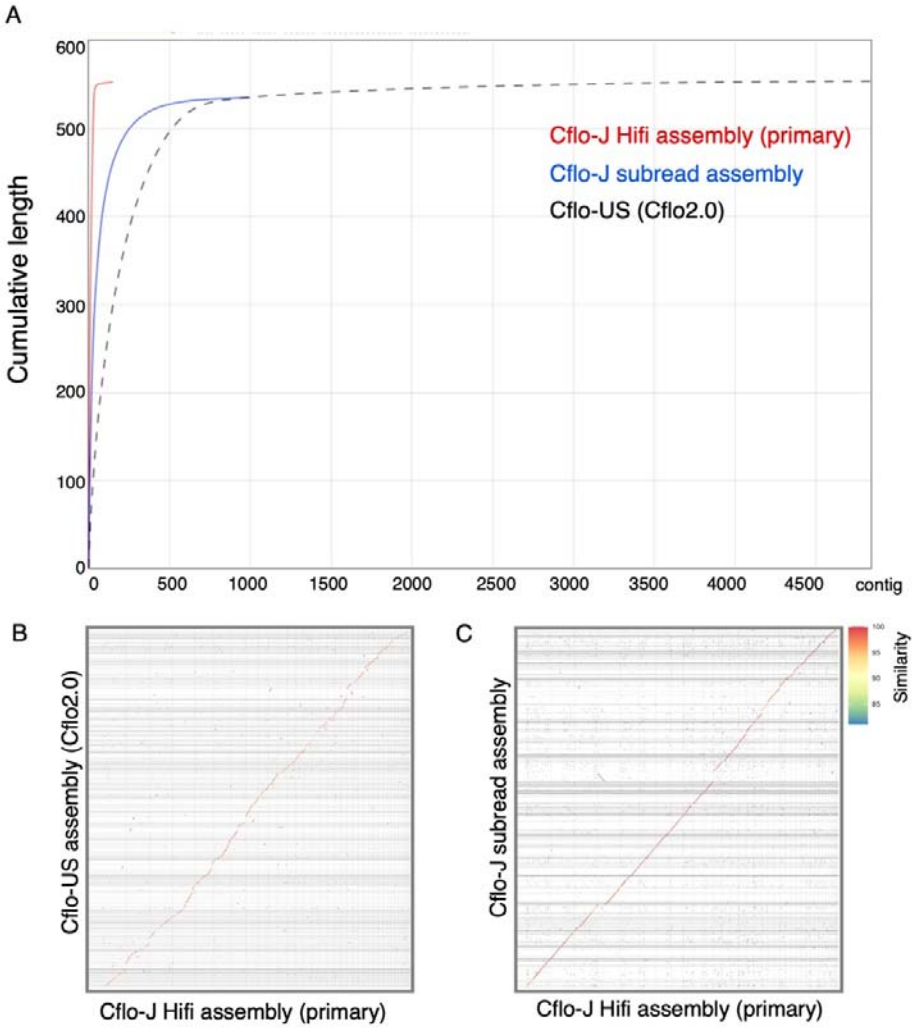

Figure 1 Contiguity and sequence similarity of each assembly. (A) Each color shows the cumulative base pair length and number of contigs for each assembly. The assemblies generated from HiFi reads contain the fewest contigs. (B, C) Dot-plot between the Japanese *C. floridanum* strain (Cflo-J) and US *C. floridanum* (Cflo-US) high-fidelity (HiFi) assemblies and between Cflo-J HiFi and Cflo-J subread assemblies. The vertical axis indicates Cflo-US or Cflo-J subread assemblies. The horizontal axis indicates Cflo-J HiFi assembly. The average identity percentages (per query) are displayed in different colors.

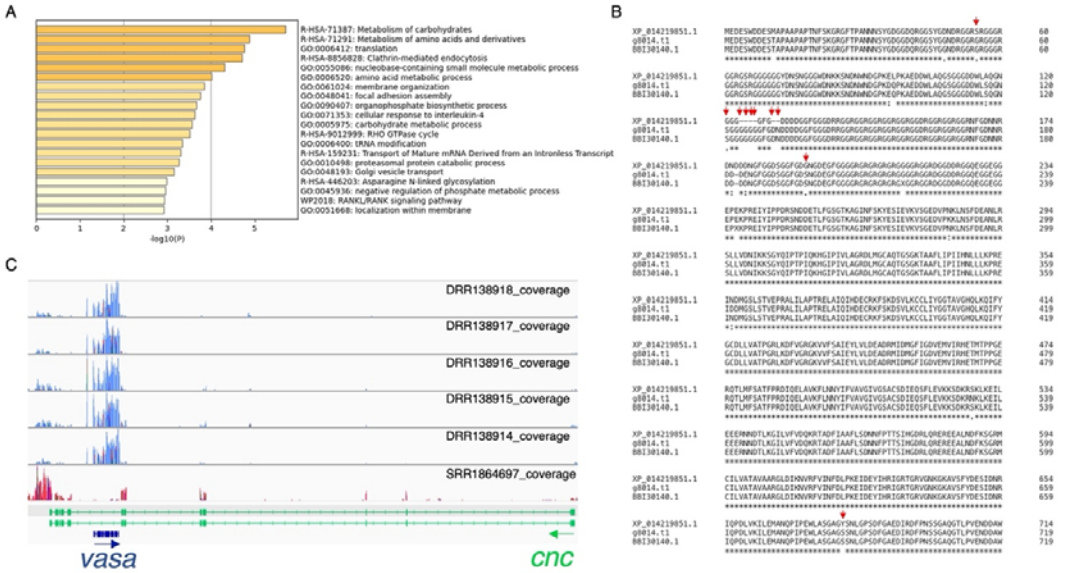

Figure 2 Comparison of gene annotation between the Cflo-US and Cflo-J HiFi assemblies. (A) Enrichment analysis using *Homo sapiens* gene IDs. Bars are colored differently according to the value of -log10 (P-value). (B) Comparison of amino acid sequences of the *vasa* gene in the US (XP_014219851.1) and Japanese strains (g8014.t1 and BBI30140.1). g8014.t1 is the transcript ID of *vasa* predicted through HiFi read assembly, and BBI30140.1 is the protein ID of *vasa* in the Japanese lineage reported by Ohno et al. (2019). Red arrows indicate amino acids that differ between the US and Japanese strains, as Ohno et al. 2019 revealed. (C) Genomic position of *vasa* with RNA-Seq read coverages. The CDS regions are shown below the RNA-Seq coverage data, including the *vasa* and *cap'n'collar* (*cnc*) exon structures. Blue and green arrows indicate the direction of transcription.