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

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: bonohu@hiroshima-u.ac.jp; Tel.: +81 82 424 4013

Copidosoma floridanum (Encyrtidae: Hymenoptera) is a cosmopolitan species and egg-larval parasitoid of the Plusiine moth (Lepidoptera, Noctuidae, Plusiinae). This species has a unique mode of development called polyembryony, in which several thousands of genetically identical embryos are produced from a single egg. Some of the embryos develop into sterile soldier larvae, and their developmental patterns differ between US and Japanese strains. Genome sequencing can accelerate our understanding of the molecular basis underlying polyembryony, including the production of soldier castes. Until now, only the genome sequence of the US strain has been reported. In this study, we determined the whole genome sequence of the Japanese strain using the PacBio HiFi reads and succeeded in the generation of a highly contiguous assembly (552.7 Mb, N50: 17.9Mb). Gene prediction and annotation identified 13,886 transcripts derived from 10,786 gene models. Among those transcripts, 149 transcripts could not identified in the gene model of US strain. Genomic differences between the US and Japanese strains were identified as differences in the amino acid sequence of vasa, as reported in previous studies. The genome assemblies constructed in this study should facilitate our understanding of intraspecific variation in the developmental patterns of *C. floridanum.*

Keywords: *Copidosoma floridanum*, PacBio HiFi reads, genome assembly

Introduction

Copidosoma floridanum (Hymenoptera: Encyrtidae) is a cosmopolitan species and egg-larval parasitoid of the Plusiinae moth (Lepidoptera, Noctuidae, Plusiinae)[1]. This species has a unique mode of development called polyembryony, in which >2000 genetically identical individuals are produced from a single egg[2]. Although polyembryony has evolved independently in four families of Hymenoptera (Encyrtidae, Platygastridae, Braconidae, and Dryinidae), that of encyrtid wasps shows the most extreme form in terms of the number of clone size[3]. In some encyrtid wasps, some of 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 as one of the major evolutionary events that brought complexity to life on Earth[6,7].

The process and molecular basis of soldier development of *C. floridanum* have been described and investigated both in the US and Japanese strain[2,8,9]. The development and behavioral patterns of the male soldiers differ between the two strains. In the US strain, male soldiers appear in the last larval stage of the hosts[10], while in the Japanese strain, those appear at the first or second instar larvae of the hosts[11]. The aggression of male soldiers is much weaker than female soldiers in the Japanese strain whereas male soldiers in the US strain are non-aggressive[11,12]

Each strain is expected to optimize its behavior and developmental patterns in its habitat, which may be related to genomic differences. The genome or transcriptome analysis can accelerate our understanding of the molecular basis underlying polyembryony including the production of soldier caste[9]. *vasa* is one of the genes that have been compared between Japan and the U.S., and the difference is found in the 10 amino acid sequence of the N-terminal region[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 variation in soldier developmental or behavioral patterns between strains remain unknown. Although the genome of the US strain was sequenced by short-read sequencing[18], that of the Japanese strain has been not sequenced. Unlike at that time, long-read sequencing has been found to offer many advantages over short-read sequencing techniques[19]. The high quality (continuity and completeness of annotation) of insect genomes has been recently provided by long-read sequencing[20]. Single-molecule real-time (SMRT) sequencing is one of the long-read sequencing methods developed by Pacific Biosciences (PacBio)[21]. Particularly, circular consensus sequencing (CCS) generates highly accurate long high-fidelity reads (HiFi reads, >10kb, >99%)[22–24]. HiFi reads are also more suitable for identifying genes that contain repetitive sequences than other long-read or short-read methods[25,26].

In this study, we used PacBio long-read sequencing to construct a highly continuous genome assembly of Japanese *C. floridanum*. We confirmed the sequence differences of *vasa* gene between the US and Japanese strains, as previously reported[13]. We performed gene prediction and annotation, resulting in the determination of 10786 protein-coding gene model.

Materials and Methods

DNA extraction and genome sequencing

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

De novo assembly and quality assessment

Subreads and HiFi reads generated from the PacBio sequencers were assembled by Flye v2.9[27] and Hifiasm v0.16.1 with '-primary' option[28], respectively. Quast v5.2.0 was used to assess the genome assembly statistics[29]. The completeness in terms of gene sets was assessed through Benchmark of Universal Single-Copy Orthologs (BUSCOs) v5.2.2 with '-l insecta_odb10 (2020-09-10)'[30]. For the assembly constructed with HiFi reads, k-mer-based validation was performed with Merqury v1.3 and Meryl db (k=21)[31]. Contigs were assigned to taxonomic classification by BlobTools v1.1[32] with BLAST (BLASTN) program in the NCBI BLAST software package (v 2.14.0) against the NCBI non-redundant nucleotide sequence database (nt), which was prepared using ncbi-blast-dbs v 10.5.0 (https://github.com/yeban/ncbi-blast-dbs, accessed on 18 May 2023). To obtain mapping coverage, minimap2 v2.26-r1175[33] was used.

Dot-plot were visualized by dotPlotly program with -s -t -m 10000 -q 25000 -l (https://github.com/tpoorten/dotPlotly, accessed on 17 July 2023) to compare the genomic differences between US and Japanese strains. The 1 Mb or longer contigs were extracted by SeqKit v2.0.0, and aligned by nucmer with default settings, and then delta files were filtered with delta-filter with -'q -r' options in MUMmer4 v4.0.0rc[34].

Gene prediction and annotation

Gene prediction was performed by BRAKER2 and BRAKER3[35,36]. Public RNA sequencing (RNA-seq) data (molura: DRR138914, DRR138915, DRR138916, and DRR138917, male adult: SRR1864697), and OrthoDB 11 protein sets of Arthropoda[37] (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/, accessed on 17 June 2023) were used as extrinsic evidence data for performing BRAKER pipeline. The RNA-seq reads were processed by fastp v0.23.2 with default settings[38], and mapped to the *de novo* assembled genome by HISAT2 v2.2.1[39] The generated SAM files were converted to BAM files by SAMtools v1.12[40] and the BAM files were imported together when the BRAKER pipeline was run. Repetitive sequences of the assembly were soft-

masked by RepeatModeler v2.0.3 and RepeatMasker v4.0.6 in Extensice de novo TE Annotator (EDTA) v2.1.3[41]. The soft-masked genome assembly was input into BRAKER pipeline. GFFread v0.12.1[42] was used to convert transcript sequences or protein sequences from the GTF file created by BRAKER. Protein sequences were assessed by BUSCO (protein mode) with '-l insecta_odb10 (2020-09-10)'.

Functional annotation

The functional annotation workflow of Fanflow4Insects[43] was modified and performed 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 by GGSEARCH v36.3.8g in the FASTA package (https://fasta.bioch.virginia.edu/, accessed on 19 July 2023). The protein sequences were also searched by HMMSCAN in the HMMER v3.3.2. Pfam v35.0 was used as the protein domain database for HMMSCAN[44]. Enrichment analysis was performed using the Metascape software (https://metascape.org/, accessed on 19 July 2023)with default settings[45]

Results and Discussion

De novo genome assembly of Japanese *C. floridanum*

In this study, HiFi reads and subreads were generated using Japanese strains with PacBio sequel IIe and PacBio sequel II, respectively. The quality of assembly was then evaluated between the Japanese strain (hereafter referred to as Cflo-J) and the US strain (hereafter referred to as Cflo-US). Cflo_2.0 (NCBI RefSeq assembly GCF_000648655.2) was used as the genome assembly of Cflo-US. HiFi reads were assembled by Hifiasm, and the primary contig consisted of 149 contigs with 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 (1162 contigs, Supplementary Table S1). Subreads were assembled by Flye, and the contig consisted of 996 contigs with 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). N50 of Cflo-J assemblies were much better than that of Cflo-US (1.2M). GC content was almost the same among the assemblies. BUSCO completeness of Cflo-J also comparable to the genome of Cflo-US, but the fragmented score was lower on the assemblies of Cflo-J (Table 1). The N50 values and the number of contigs were the best for the HiFi assembly among all assemblies (Table 1 and Figure 1A). We confirmed the differences in genomic structures among all assemblies by dot-plot visualization (Figure 1B and C). Linear plots were obtained, confirming the similarity of genomic structures across the whole length of assemblies (Figure 1B and C). Match coordinate data for the dot-plot comparison is available in Supplementary File 1 and 2.

In the following sections, we continued the evaluation of the HiFi assemblies of Cflo-J that had the highest contiguity.

In this study, Hifiasm generated the primary and alternate contigs (Supplementary Table S1). K-mer-based evaluation by Merqury showed high base accuracy and completeness of primary contig (QV = 68.08 and Completeness = 97.6%, Supplementary Table S2). Merqury spectrum plot showed a single k-mer peak in the primary contigs, revealing extremely low heterozygosity (Supplementary Figure S1A, red line). Since *C. floridanum* male possesses haploid genome, the single k-mer peak of 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. To reveal the taxonomic origin of contigs, taxonomic assignments of contigs were investigated by Blobtools. Primary and alternate contigs were almost assigned to Arthropoda, and the remaining were either no-hit or chordata (Supplementary Figure 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 altenate contig was also extremely low (0.05 %). The assembly of the Cflo-US also consists of Arthropoda and no-hit contigs (shown in ENA https://www.ebi.ac.uk/ena/browser/view/GCA_000648655.2, accessed on 20 Sep 2023). Anyway, an ideal haploid representation was observed in the primary contigs, and no contamination of contigs from non-arthropods was confirmed.

Annotation of primary contig

Since repetitive sequences may affect the accuracy of gene prediction[46], repetitive sequences in the HiFi primary contig were investigated by RepeatModeler and RepeatMasker. The repeat sequence was 285.4 Mb, which corresponds to 51.65% of the HiFi primary contig, and "Unclassified" accounted for a substantial percentage of the total (34.97%) (Supplementary Table S3). The HiFi primary contig soft-masked by RepeatModelr/RepeatMasker was transferred to the BRKAER pipeline. Gene prediction was performed in the BRAKER pipeline using several different combinations of RNA-seq data and protein datasets. The BRAKER3 using RNA-seq data and Arthropoda protein dataset showed the best BUSCO completeness (96.4%, Table 2). The numbers of predicted genes and transcripts are also shown in Table 2. BRAKER2 tended to predict a larger number of genes compared to BRAKER3 (Table 2). The gene model with the highest BUSCO completeness (96.4%) (hereafter referred to as the “best gene model”) predicted 10786 protein-coding genes and 13886 transcripts. Annotation GFF file are available as Supplementary File 3. We annotated the best gene model by Fanflow4Insects[43]. Table 3 shows the results of searches by Fanflow4Insects (GGSEARCH and HMMSCAN) for each reference dataset. Of the 10786 gene models, 10585 gene model was annotated by GGSEARCH or HMMSCAN. The full annotation results of GGSEARCH and HMMSCAN are available in Supplementary Table S4.

We evaluated whether annotation differences existed between the Cflo-US assembly and the Cflo-J HiFi primary contig. In the best gene model, there were 149 transcripts with no corresponding protein in the Cflo-US dataset but with similarities to the other reference datasets (Supplemental Table S5). The human gene IDs corresponding to the 149 transcripts were used for enrichment analysis using Metascape (Figure 2A). “Metabolism of carbohydrates” was the most significantly enriched GO term. In addition, the GO terms for amino acid metabolism and translation were enriched. We have succeeded in complementing the annotation of genes that are fundamentally necessary for the maintenance of life activities.

Ohno et al (2019)[13] showed that the amino acid sequence of *vasa* was different between Cflo-J and Cflo-US. The amino acid sequences of Cflo-US (XP_014219851.1) and Cflo-J *vasa* (g8014.t1) were compared, including the amino acid sequence (BBI30140.1) determined in Ohno et al (2019). 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 reported previously (Figure 2B, red arrows). Within the Japanese strain (BBI30140.1 vs g8014.t1), amino acid sequences were almost identical compared to the US strain (Figure 2B). We also investigated the genomic position of *vasa* (Figure 2C). *vasa* was located in intron between the 7th and 8th exons of *cap'n'collar* (*cnc*), and the direction of transcription was opposite to each other. 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 of other insects were examined preliminarily (Supplementary Table S6). As long as we have examined, the positional relationship between *vasa* and *cnc* may be conserved in Hymenoptera. In insects, polyembryony has been reported mostly in Hymenoptera[47]. Soldier development regulated by vasa expression has been reported in partly polyembryonic wasps of the Hymenoptera. In conclusion, the soldier differentiation observed in hymenopterans may have something to do with the conserved positional relationship between vasa and cnc, but the actual regulatory relationship is unknown. A simple hypothesis is that cnc may also be involved in the development of soldiers through the regulation of vasa expression. In fact, cnc is related to the normal arrangement of oocytes in Drosophila melanogaster, and thus is a gene involved in embryonic development. However, since vasa and cnc show a conserved positional relationship in Hymenoptera, it is possible that the basis of this positional relationship was acquired during the evolution of Hymenoptera.

In this study, we obtained highly contiguous genome assembly of the Japanese *C. floridanum*. This genome assembly is expected to contribute to our understanding of the molecular basis of intraspecific variation in polyembryony of *C. floridanum*. Furthermore, parasitic bees have mechanisms to control the physiological state of host insects, and their use for insect control is expected (Beckage and Gelman 2004). In conclusion, the genome assembly and annotation

generated in this study also should contribute to our understanding of the molecular basis for the pest insect control.

Funding

This work 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

Authors declare no conflict of interest.

Data availability

All sequencing data (assembled sequences and raw sequence reads) have been deposited in DDBJ under umbrella BioProject accession number PRJDB16642. The genome assembly from primary contigs has been deposited in DDBJ under the accession numbers BTPR01000001−BTPR01000149. Alternate contigs has been deposited in DDBJ under the accession numbers BTPQ01000001−BTPQ01001162. The raw sequence reads have been deposited in DDBJ under the accession numbers DRA017002 (PacBio HiFi reads). Gene annotation and functional annotation of the protein-coding genes are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v1). Supplementary Files are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v1).

Supplementary data

Supplementary Table S1 Statistics of the alternate contigs generated by Hifiasm

Supplementary Table S2 Statistics summary generated by k-mer assembly evaluation by Merqury
Supplementary Table S3 Summary of repetitive elements in Cflo-J HiFi assembly (primary)
Supplementary Table S4 All results of functional annotation by Fanflow4Insects
Supplementary Table S5 Transcripts with no corresponding protein in the Cflo-US dataset
SupplementaryTable S6 Genomic location of *vasa* gene
Supplementary File 1: Match coordinate data between Cflo-US assembly and Cflo-J HiFi assembly
Supplementary File 2: Match coordinate data between Cflo-J subread assembly and Cflo-J HiFi assembly
Supplementary File 3: Annotation GTF file generated by BRAKER

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*.

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 Atrhtropoda protein	BRAKER2 with Arthropoda protein
Complete (single, duplicated)	96.4 %(79.5%, 16.9%)	94%(73.9%, 20.1%)	93.1% (78.8%, 14.3%)	92.5% (82.8%, 9.7%)
Fragmented	0.5%	2.40%	0.70%	1.80%
Missing	3.1%	3.60%	6.20%	5.70%
No. of genes	10786	18531	9913	22781
No. of transcripts	13886	19986	12370	24487

Table3 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

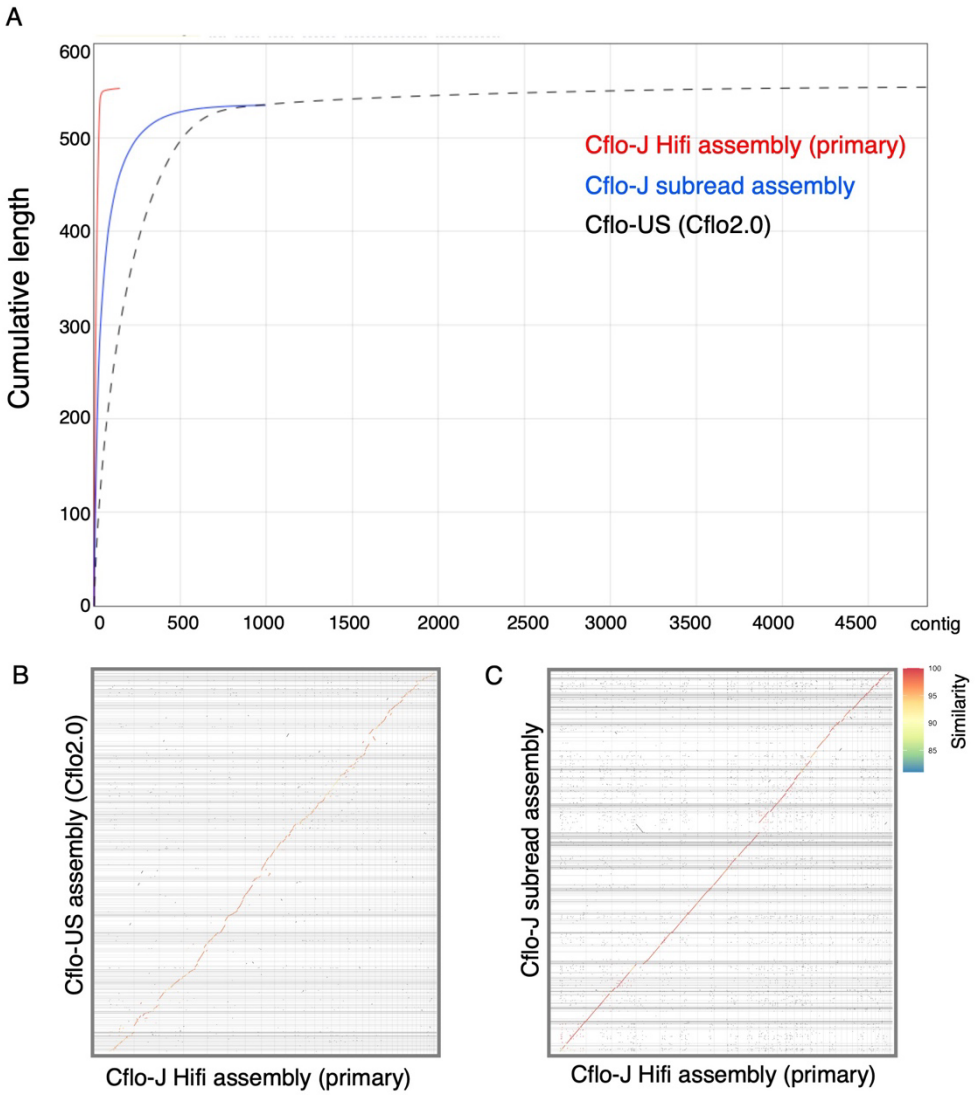

Figure 1 Contiguity and sequence similarity of each assembly. (A) The cumulative base pair length and number of contigs for each assembly are shown by each color. The assemblies generated from HiFi reads contain the fewest number of contigs. (B, C) Dot-plot between Cflo-J HiFi assembly and Cflo-US assembly,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

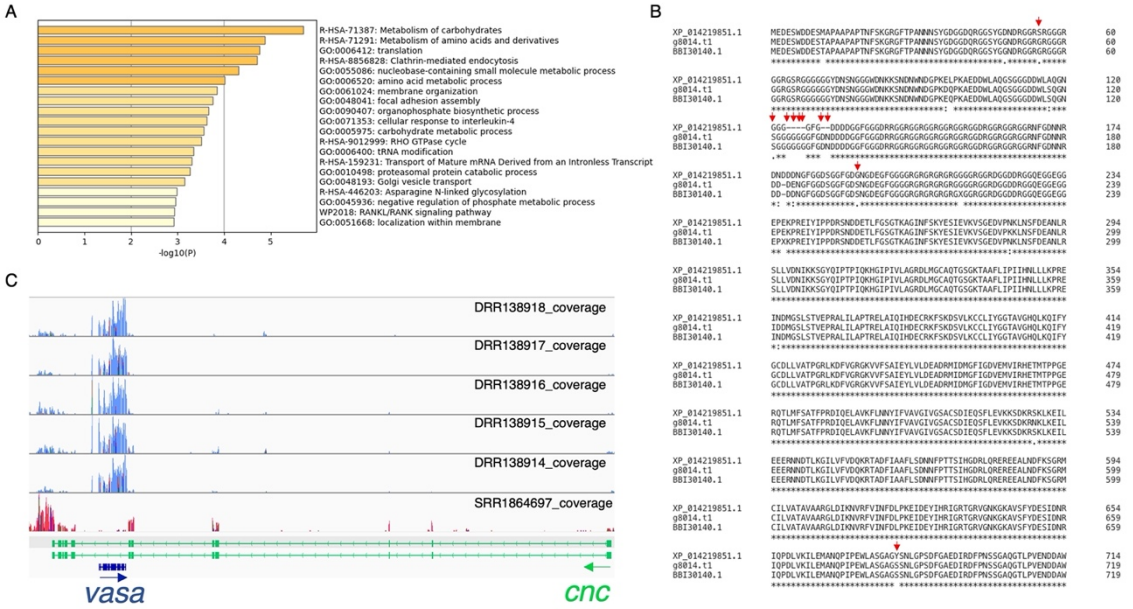

Figure 2 Comparison of gene annotation between Cflo-US assembly and Cflo-J HiFi assembly. (A) Results of enrichment analysis using gene IDs of *H. sapiens*. Bars are colored differently according to the value of -log10(P value). (B) Comparison of amino acid sequences of *vasa* gene in US (XP_014219851.1)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