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

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

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. We searched the genomic differences between US and Japanese strains. Among gene models predicted in this study, 100 gene loci in the Japanese strain had extremely different gene structure from those in the US strain. This was accomplished through the functional annotation (GGSEARCH) and long-read sequencing. Genomic differences between strains were also reflected to amino acid sequences of *vasa* that plays a central role in caste determination in this species. The genome assemblies constructed in this study will facilitate the genomic comparisons between Japanese and US strains, leading to our understanding of detail genomic regions responsible for the ecological and physiological characters of *C. floridanum.*

Introduction

Copidosoma floridanum (Hymenoptera: Encyrtidae) is a worldwide distributed egg-larval parasitoid of the Plusiinae moth (Lepidoptera, Noctuidae, Plusiinae) (Guerrieri and Noyes 2005). This species reproduces by polyembryony, producing >2,000 genetically identical individuals from a single egg (Ode and Strand 1995). 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 (Segoli et al. 2010). Embryos develop into larvae that feed on the host tissues until pupation. In some encyrtid wasps, the embryos develop into two types of larvae: reproductive larvae and the sterile soldier larvae. Sterile soldier larvae possesses large mandibles, and contribute to the defense of their clonal siblings by attacking against a competitor in the host (Cruz 1981; Strand 2003; Iwabuchi 2019). 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 (Szathmáry and Smith 1995; Whyte 2021).

The process and molecular basis of soldier development in *C. floridanum* have been described and investigated mainly in the US and Japanese strains (Ode and Strand 1995; Smith et al. 2017; Sakamoto et al. 2020). Like other hymenopteran insects (Cook 1993), *C. floridanum* is haplodiploid species in which the females develop from the fertilized eggs and the males develop from the unfertilized egg (Iwabuchi 2019). Soldier larvae develop from both male and female embryo, and the development and behavioral patterns of male soldiers differed between intraspecific strains. In the US strain, male soldiers appear in the penultimate larval stage of the host (Grbic et al. 1997), whereas in the Japanese strain, they appear in the first- or second-instar larvae of the host (Yamamoto et al. 2007). Male soldiers were functional in the Japanese strain but are not as aggressive as the female soldiers (Uka et al. 2013). Male soldiers in the US strain are nonaggressive (Giron et al. 2007). *vasa*,

72 encoding a DEAD-box RNA helicase protein (Hay et al. 1988; Lasko and Ashburner 1988), has been
73 shown to play a central role in caste determination of *C. floridanum* (Donnell et al. 2004; Vladimir
74 Zhurov et al. 2004). The *vasa* gene was originally identified in *Drosophila melanogaster* as a
75 maternal-effect factor for germ-cell specification and development of abdominal segments
76 (Schupbach and Wieschaus 1986), and they have been found in invertebrates and vertebrates (Raz
77 2000). *vasa* has been compared between the Japanese and the US strains of *C. floridanum*, and a 10
78 amino acid sequence difference in the N-terminal region was observed (Ohno et al. 2019).
79
80 Each strain should optimize its behavior and developmental patterns for its habitat. Genome
81 or transcriptome analyses should accelerate our understanding of the molecular basis underlying the
82 intraspecific differences of parasitic ability including polyembryony and caste system (Sakamoto et
83 al. 2020). The genome of the US strain has been sequenced using short-read sequencing (i5K
84 Consortium 2013), while that of the Japanese strain has not. Long-read sequencing offers several
85 advantages over short-read sequencing (Pollard et al. 2018). High-quality (high contiguity of the
86 assembly and completeness of gene sets) insect genomes have recently been achieved using long-read
87 sequencing (Hotaling et al. 2021). Single-molecule real-time (SMRT) sequencing developed by
88 Pacific Biosciences (PacBio) employ the DNA polymerase reaction from a single circular DNA
89 template (Eid et al. 2009), generating highly accurate long high-fidelity reads (HiFi reads, >10 kb,
90 >99%) (Rhoads and Au 2015; Ardui et al. 2018; Wenger et al. 2019). HiFi reads are more suitable
91 than other long- or short-read methods for identifying genes containing repetitive sequences
92 (Kawahara et al. 2022; Hotaling et al. 2023).
93
94 Comparisons among intraspecific genomes should be useful in understanding of the various
95 parasitic ability in *C. floridanum*. In this study, we aimed to construct a highly contiguous genome

96 assembly of the Japanese *C. floridanum* using PacBio long-read sequencing. We performed the quality
97 check, the gene prediction and functional annotation for the *de novo* assembly. Gene annotation and
98 amino acid sequences of *vasa* were compared to clarify the genomic differences between Japanese
99 and U.S. strains.
100

101 **Materials and Methods**

102 **DNA extraction and genome sequencing**

103 Genomic DNA was extracted from adult *C. floridanum* males using the Blood & Cell Culture DNA
104 Maxi Kit (Qiagen Co. Ltd., Valencia, CA, USA). The genetic system of *C. floridanum* and other
105 hymenopteran insects is haplodiploid; females develop from fertilized eggs (diploid), and males
106 develop from unfertilized eggs (haploid). In this study, the male adults were used. The male adults are
107 clones born from a single egg and are genetically nearly identical. The sequence library was prepared
108 using the SMRTbell Express Template Preparation Kit 2.0 or SMRTbell Prep Kit 3.0 (Pacific
109 Biosciences, Menlo Park, CA, USA). Libraries were sequenced using the Sequel II or IIe (Pacific
110 Biosciences).
111

112 ***De novo* assembly and quality assessment**

113 Subreads and HiFi reads generated from the PacBio sequencers were assembled using Flye v 2.9
114 (Kolmogorov et al. 2019) and Hifiasm v 0.16.1 using the '-primary' option (Cheng et al. 2021),
115 respectively. In the PacBio long-read system, polymerase reads are yielded by sequencing circularized
116 DNA. Then, the adaptors are removed from the polymerase reads, offering the subreads (refer to
117 https://www.pacb.com/technology/hifi-sequencing/). HiFi reads are defined as the consensus
118 sequence called from the subreads. Hifiasm v 0.16.1 can produce the primary contig and alternate
119 contig. The primary contig indicates the haplotype, and the alternate contig is composed of short

120 haplotype-specific contigs (haplotigs). Quast v 5.2.0 was used to assess the genome assembly statistics
121 (Gurevich et al. 2013). Gene set completeness was assessed using the Benchmark of Universal Single-
122 Copy Orthologs (BUSCOs) v 5.2.2, with '-l insecta_odb10 (2020-09-10)' (Manni et al. 2021). k-mer-
123 based validation was performed using Merqury v1.3 and Meryl db (k=21) (Rhie et al. 2020) for the
124 assembly constructed with HiFi reads. Contigs were assigned to taxonomic classifications using
125 BlobTools v 1.1(Laetsch and Blaxter 2017) with the BLAST (BLASTN) program in the National
126 Center for Biotechnology Information BLAST software package (v 2.14.0) against the NCBI non-
127 redundant nucleotide sequence database (nt), prepared using ncbi-blast-dbs v 10.5.0
128 (https://github.com/yeban/ncbi-blast-dbs, accessed on May 18, 2023). minimap2 v 2.26-r1175(Li
129 2018) was used to obtain mapping coverage.

130 Dot plots were visualized using the dotPlotly program with '-s -t -m 10000 -q 25000 -l'
131 (https://github.com/tpoorten/dotPlotly, accessed on July 17, 2023) to compare the genomic differences
132 between the US and Japanese strains. The 1 Mb or longer contigs were extracted using SeqKit v 2.0.0,
133 aligned by a nucmer with default settings, and delta files were filtered using a delta-filter with '-q -r'
134 options in MUMmer4 v4.0.0rc (Marçais et al. 2018).

135

136 **Gene prediction and functional annotation**

137 Gene prediction was performed using the BRAKER2 and BRAKER3 databases (Brůna et al. 2021;
138 Gabriel et al.). Public RNA sequencing (RNA-Seq) data (morula: DRR138914, DRR138915,
139 DRR138916, and DRR138917; male adult: SRR1864697) and OrthoDB protein sets of Arthropoda
140 (Kuznetsov et al. 2023) (https://bioinf.uni-greifswald.de/bioinf/partitioned_odb11/, accessed on June
141 17, 2023) were used as extrinsic evidence data to perform the BRAKER pipeline. The RNA-Seq reads
142 were processed using fastp v 0.23.2 with default settings (Chen 2023) and mapped to the *de novo*
143 assembled genome using HISAT2 v 2.2.1(Kim et al. 2019). The generated SAM files were converted

144 to BAM files using SAMtools v 1.12 (Danecek et al. 2021) and the BAM files were imported together
145 when the BRAKER pipeline was run. Repetitive assembly sequences were soft-masked using
146 RepeatModeler v 2.0.3 and RepeatMasker v 4.0.6 in Extensive de novo TE Annotator v 2.1.3 (Ou et
147 al. 2019). The soft-masked genome assembly was inputted into the BRAKER pipeline. GFFRead
148 v0.12.1 (Pertea and Pertea 2020) was used to convert the transcript or protein sequences from the GTF
149 file created by BRAKER. Protein sequences were assessed using BUSCO (protein mode) with '-l
150 insecta_odb10' (2020-09-10).

151 The functional annotation workflow of Fanflow4Insects (Bono et al. 2022a) was modified
152 and used for protein sequences, as described below. The protein sequences were aligned against
153 reference datasets of *Homo sapiens*, *Mus musculus*, *Caenorhabditis elegans*, *Drosophila*
154 *melanogaster*, *C. floridanum* (US strain), and UniProtKB/Swiss-Prot using GGSEARCH v 36.3.8g in
155 the FASTA package (https://fasta.bioch.virginia.edu/, accessed on July 19, 2023). Protein sequences
156 were also searched using HMMSCAN in HMMER v 3.3.2. Pfam v 35.0 was used as the protein
157 domain database for HMMSCAN (Mistry et al. 2021). Enrichment analysis was performed using
158 Metascape software (https://metascape.org/, accessed on July 19, 2023) with default settings (Zhou et
159 al. 2019).

160

161 **Results and Discussion**

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

163 This study generated HiFi reads and subreads using Japanese strains (Cflo-J). HiFi reads were
164 assembled, and the primary contig comprised 149 contigs with an N50 length of 17.9 Mb and a total
165 assembly size of 552,713,363 bp (50.8-fold coverage; Table 1 and Table S1). Hifiasm also produced
166 alternate contigs (1,162 contigs, Table S1). The genome assembly generated by subreads comprised
167 996 contigs with an N50 length of 4.6 Mb and a total assembly size of 534,638,157 bp (244.7-fold

168 coverage, Table 1 and Table S1). Cflo_2.0 (NCBI RefSeq assembly GCF_000648655.2) was used as
169 the genome assembly of US strain (Cflo-US). The assembly quality between the Japanese strain (Cflo-
170 J) and the US strain (Cflo-US) was evaluated. The N50 value of Cflo-J assemblies was much higher
171 than that of Cflo-US (1.2 M). However, the GC content of the assemblies was similar. The BUSCO
172 completeness of Cflo-J was also comparable to that of Cflo-US; however, the fragmentation score was
173 lower for the Cflo-J assemblies (Table 1). The N50 values and number of contigs were the best for the
174 HiFi assembly (Table 1 and Figure 1A). We confirmed the differences in genomic structures among
175 all the assemblies using dot-plot visualization (Figures 1B and C). Linear plots were obtained,
176 confirming the similarity of the genomic structures across the entire assembly length (Figure 1B and
177 C). The matching coordinate data for the dot-plot comparison are available in Files S1 and S2.

178 We then evaluated the HiFi assemblies of Cflo-J, which had the highest contiguity. K-mer-
179 based evaluation showed high base accuracy and completeness of primary contigs (a quality value
180 (QV) = 68.08 and completeness = 97.6%, Table S2). The Merqury spectrum plot showed a single k-
181 mer peak in the primary contigs, indicating extremely low heterozygosity (Figure S1A, red line).
182 Because *C. floridanum* males possess a haploid genome, the single k-mer peak of the primary contig
183 was reasonable. We also confirmed the presence of reads included only in the alternate contigs (Figure
184 S1A, blue line), suggesting that the genome assembly constructed in this study may contain
185 heterozygous regions.

186 Taxonomic assignments of the contigs were performed to reveal the taxonomic origin of the
187 contigs. The primary and alternate contigs were assigned to Arthropoda, and the remaining contigs
188 were either no-hit or chordata (Figures S1B and C). The mapping rate of no-hit was remarkably low
189 (primary contig, 0.27%; alternate contig, 0.51%), and that of chordata in alternate contigs was also
190 extremely low (0.05%). The assembly of Cflo-US also includes Arthropoda and no-hit contigs as
191 shown in Cflo-J (ENA https://www.ebi.ac.uk/ena/browser/view/GCA_000648655.2, accessed on

192 September 20, 2023). In conclusion, an ideal haploid representation was observed in the primary
193 contigs, and no contig contamination from non-arthropods was observed.
194
195 **Gene annotation differences between Cflo-J and Cflo-US**
196 As repetitive sequences may affect gene prediction accuracy (Hoff et al. 2019), these sequences in the
197 HiFi primary contig were investigated using RepeatModeler and RepeatMasker. The repeat sequence
198 was 285.4 Mb, corresponding to 51.65% of the HiFi primary contig, and "Unclassified" accounted for
199 a substantial percentage of the total (34.97%) (Table S3). Gene prediction was performed using several
200 RNA-Seq data and protein dataset combinations. The BRAKER3 using RNA-Seq data and the
201 Arthropoda protein dataset showed the best BUSCO completeness (96.4%, Table 2). The numbers of
202 predicted genes and transcripts are listed in Table 2. BRAKER2 predicted more genes compared to
203 BRAKER3 (Table 2). The gene model with the highest BUSCO completeness (96.4%) ("best gene
204 model") predicted 10,786 protein-coding genes and 13,886 transcripts. The annotated GFF files are
205 shown in File S3. The best gene model was annotated using Fanflow4Insects (Bono et al. 2022b).
206 Table 3 shows the results of the searches using Fanflow4Insects for each reference dataset. Of the
207 10786 gene models, 10585 were annotated using GGSEARCH or HMMSCAN. The full annotation
208 results for GGSEARCH and HMMSCAN are presented in Table S4.
209 We evaluated whether annotation differences existed between the Cflo-US assembly and
210 Cflo-J HiFi assembly. We identified 149 transcripts in 115 gene loci with no corresponding proteins
211 to Cflo-US and with corresponding proteins to human, mouse, *C.elegans*, and fly (Table S5). The
212 human gene IDs corresponding to 149 transcripts were used for enrichment analysis (Figure 2A).
213 "Metabolism of carbohydrates" was the most significantly enriched Gene Ontology (GO) term. In
214 addition, GO terms for amino acid metabolism and translation were enriched.

215 Excluding gene loci with at least one isoform corresponding to Cflo-US genes, the number
216 of genes as mentioned above was limited from 115 to 100. We investigated why these 100 genes did
217 not hit the Cflo-US gene model. To verify whether corresponding regions of the 100 genes are absent
218 in the Cflo-US, TBLASTN of 100 genes was performed for the genome of Cflo-US. We manually
219 examined hit regions of TBLASTN using NCBI Genome Data viewer. TBLASTN results showed that
220 all 100 genes hit in the Cflo-US genome. Thus, the difference in genome annotation between Cflo-US
221 and Cflo-J is not due to differences in the extent of the sequenced genomic regions. Among 100 genes,
222 21 genes had no corresponding gene models in Cflo-US assembly, while 79 genes had corresponding
223 gene models in Cflo-US (Table S6). The gene model g524 was representative example without
224 corresponding gene model of Cflo-US (Figure 2B). On both sides of g524 are g523 and g525. Amino
225 acid sequences of g523 and g525 have sequence similarity to XP_014209029.1 and XP_014209034.1
226 in Cflo-US genome, respectively (Table S4). Although these genomic position and sequence similarity
227 indicate that the genomic position around g524 is conserved in both genome assembly, the gene model
228 corresponding to g524 was absent in Cflo-US.
229 Among 79 genes with corresponding gene model in Cflo-US, TBLASTN results were
230 divided into two groups: genes that hit the single gene model or genes that hit more than two gene
231 model (Table S6). In addition, differences of the amino acid length between Cflo-J and Cflo-US were
232 also revealed by the TBLASTN (Table S6). g490 was representative example as the gene that hit the
233 single gene model of Cflo-US. TBLASTN showed that the amino acid sequence of g490.t1 had
234 similarity to XP_014209022.1 of Cflo-US gene model (Table S6). The exon structures of
235 XP_014209022.1 were different from g490 of Cflo-J (Figure 2C). Multiple sequence alignment
236 showed that the amino acid length of XP_014209022.1 of Cflo-US was extremely longer at N-terminal
237 region than the corresponding protein of other species (Figure S2). When BLASTP was adopted,
238 g490.t1 showed high sequence similarity to XP_014209022.1 (E-value: 0.0), despite the gene

239 structures were largely different. This indicates that GGSEARCH adopted in this study can detect the
240 genes with extremely different structure of genes. g909 was the representative example as the gene
241 that hit several gene model of Cflo-US (Table S6). TBLASTN showed that g909 had the similarity
242 XP_023247179.1 and XP_014214160.1, and these two genes of Cflo-US were distributed in different
243 scaffolds (Figure 2D). Multiple sequence alignment indicated that XP_023247179.1 and
244 XP_014214160.1 correspond to the first and second half of the amino acid sequence of g909.t1,
245 respectively (Figure S3). The example of g909 suggested that long-read sequencing contributed to the
246 accuracy of genome annotation. TBLASTN also detected the genes that hit more than two gene model
247 on the same scaffold (Table S6, e.g. g1061.t1).

248 In this study, we found genes with extremely different gene annotations between Cflo-US
249 and Cflo-J. This success is attributed to the method used in this study. GGSEARCH, which was
250 adopted in this study, searches for sequence similarity with the subject within 80-120% of the query
251 length. This contributed to discovery of protein structure differences between Cflo-J and Cflo-US. In
252 addition, highly contiguous genome assembly obtained by long-read sequencing suggested the
253 fragmentation of gene model in Cflo-US genome. All the differences in genome annotation revealed
254 in this study are noteworthy as the discoveries of genomic differences between Cflo-J and Cflo-US.

255

256 **Common features and differences of *vasa* between Cflo-J and Cflo-US**

257 Ohno et al. (2019) showed that the amino acid sequence of *vasa* differed between Cflo-J and
258 Cflo-US. The amino acid sequences of Cflo-US (XP_014219851.1) and Cflo-J *vasa* (g8014.t1),
259 including the amino acid sequence (BBI30140.1) determined by Ohno et al. (2019), were compared.
260 We confirmed the differences in amino acid sequences at the N-terminal region between Cflo-US *vasa*
261 (XP_014219851.1) and Cflo-J *vasa* (g8014.t1), as previously reported[13] (Figure 3A, red arrows).

262 Within the Japanese strain (BBI30140.1 vs. g8014.t1), the amino acid sequences were almost identical
263 (Figure 3A).
264 We also investigated the genomic position of *vasa* (Figure 3B). *vasa* was located in an intron
265 between the seventh and eighth exons of *cap'n'collar* (*cnc*), and the directions of transcription were
266 opposite (Figure 2C, blue and green arrows). This positional relationship was also confirmed in Cflo-
267 US (NCBI Genome Browser: LOC106647823, Location: scaffold6|NW_019379589.1:2,626,762 –
268 2,630,760.). Generally, overlapping genes such as *vasa* and *cnc* are negatively selected evolutionarily,
269 whereas overlapping genes are sometimes retained in a species-specific manner (Craig et al. 1997). In
270 addition, overlapping genes are likely to be co-expressed. Furthermore, the *vasa* genomic positions in
271 other insects were preliminarily examined (Table S7), and the positional relationship between *vasa*
272 and *cnc* may be conserved in Hymenoptera. Genomic positional relationship of *vasa* and *cnc* may be
273 related to the oogenesis or embryogenesis in hymenopteran insects. *cnc* is related to the nuclei
274 arrangement of oocyte in *Drosophila melanogaster* (Guichet et al. 2001), suggesting its involvement
275 in egg development in *C. floridanum*. The evolution to control the caste development by *vasa* in
276 *Copidosoma* may have required a conserved positional relationship between *vasa* and *cnc* in the
277 Hymenoptera. However, further verifications for relationship between *vasa* and *cnc* are needed.
278

279 **Conclusion**

280 We obtained a highly contiguous genome assembly of Japanese *C. floridanum*. We showed that long-
281 read sequencing and functional annotation method adopted in this study can highlight the differences
282 of genome annotation between Cflo-J and Cflo-US. These regions may be due to the influence of the
283 annotation and assembly accuracy, but in any case, they are remarkable genes that reflect differences
284 among the strains. Parasitic wasps can control the physiological state of host insects, and their use
285 for pest control is expected (Soldà et al. 2008). Comparison of the genome sequences of Cflo-US and

286 Cflo-J will contribute to the identification of detailed genomic regions relevant to the ecological and
287 physiological characteristics of *C. flolidanum*, leading to a detail understanding of its parasitic ability
288 against pests.
289
290 **Data availability**
291 All sequencing data (assembled sequences and raw sequence reads) were deposited in the DDBJ under
292 the umbrella BioProject accession number PRJDB16642. Genome assembly from primary contigs was
293 deposited in the DDBJ under the accession numbers BTPR01000001−BTPR01000149. Alternate
294 contigs were deposited in DDBJ under the accession numbers BTPQ01000001−BTPQ01001162. Raw
295 sequence reads were deposited in the DDBJ under accession number DRA017002 (PacBio HiFi reads)
296 and DRA017125 (PacBio Subreads). Gene annotation and functional annotation of protein-coding
297 genes are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v5). Supplementary
298 Files are available at figshare (https://doi.org/10.6084/m9.figshare.24145182.v5). Scripts of
299 genome assembly, quality check, and gene annotation were also available
300 (https://doi.org/10.6084/m9.figshare.24145182.v5)
301
302 **Acknowledgments**
303 We would like to thank all laboratory members at Hiroshima University for their valuable comments.
304 Computations were performed on the computers at the Hiroshima University Genome Editing
305 Innovation Center. Computations were partially performed on the NIG supercomputer at ROIS
306 National Institute of Genetics.
307
308 **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.

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.).

References

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 46(5):2159–2168. doi:10.1093/nar/gky066.

Bono H, Sakamoto T, Kasukawa T, Tabunoki H. 2022a. Systematic Functional Annotation Workflow for Insects. Insects. 13(7):586. doi:10.3390/insects13070586.

Bono H, Sakamoto T, Kasukawa T, Tabunoki H. 2022b. Systematic Functional Annotation Workflow for Insects. Insects. 13(7):586. doi:10.3390/insects13070586.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3(1). doi:10.1093/nargab/lqaa108.

Chen S. 2023. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta. 2(2). doi:10.1002/imt2.107.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18(2):170–175. doi:10.1038/s41592-020-01056-5.

Cook JM. 1993. Sex determination in the Hymenoptera: a review of models and evidence.

Craig SF, Slobodkin LB, Wray GA, Biermann CH. 1997. The 'paradox' of polyembryony: A review of the cases and a hypothesis for its evolution. Evol Ecol. 11(2):127–143. doi:10.1023/A:1018443714917.

Cruz YP. 1981. A sterile defender morph in a polyembryonic hymenopterous parasite. Nature. 294(5840):446–447. doi:10.1038/294446a0.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, et al. 2021. Twelve years of SAMtools and BCFtools. Gigascience. 10(2). doi:10.1093/gigascience/giab008.

Donnell DM, Corley LS, Chen G, Strand MR. 2004. Caste determination in a polyembryonic wasp involves inheritance of germ cells. Proc Natl Acad Sci U S A. 101(27):10095–10100. doi:10.1073/pnas.0403625101.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, et al. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. Science (1979). 323(5910):133–138. doi:10.1126/science.1162986.

Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, Stanke M. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. doi:10.1101/2023.06.10.544449. https://doi.org/10.1101/2023.06.10.544449.

Giron D, Harvey JA, Johnson JA, Strand MR. 2007. Male soldier caste larvae are non-aggressive in the polyembryonic wasp *Copidosoma floridanum*. Biol Lett. 3(4):431–434. doi:10.1098/rsbl.2007.0199.

Grbic M, Rivers D, Strand MR. 1997. Caste formation in the polyembryonic wasp Copidosoma floridanum (Hymenoptera: Encyrtidae): in vivo and in vitro analysis. J Insect Physiol. 43(6):553–565. doi:10.1016/S0022-1910(97)00004-8.

Guerrieri E, Noyes J. 2005. Revision of the European species of Copidosoma Ratzeburg (Hymenoptera: Encyrtidae), parasitoids of caterpillars (Lepidoptera). Syst Entomol. 30(1):97–174. doi:10.1111/j.1365-3113.2005.00271.x.

Guichet A, Peri F, Roth S. 2001. Stable anterior anchoring of the oocyte nucleus is required to establish dorsoventral polarity of the drosophila egg. Dev Biol. 237(1):93–106. doi:10.1006/dbio.2001.0354.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: Quality assessment tool for genome assemblies. Bioinformatics. 29(8):1072–1075. doi:10.1093/bioinformatics/btt086.

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

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. Methods Mol Biol. 1962:65–95. doi:10.1007/978-1-4939-9173-0_5.

Hotaling S, Sproul JS, Heckenhauer J, Powell A, Larracuente AM, Pauls SU, Kelley JL, Frandsen PB. 2021. Long Reads Are Revolutionizing 20 Years of Insect Genome Sequencing. Genome Biol Evol. 13(8). doi:10.1093/gbe/evab138.

Hotaling S, Wilcox ER, Heckenhauer J, Stewart RJ, Frandsen PB. 2023. Highly accurate long reads are crucial for realizing the potential of biodiversity genomics. BMC Genomics. 24(1):117. doi:10.1186/s12864-023-09193-9.

i5K Consortium. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered. 104(5):595–600. doi:10.1093/jhered/est050.

Iwabuchi K. 2019. Polyembryonic Insects. Singapore: Springer Singapore.

Kawahara AY, Storer CG, Markee A, Heckenhauer J, Powell A, Plotkin D, Hotaling S, Cleland TP, Dikow RB, Dikow T, et al. 2022. Long-read HiFi sequencing correctly assembles repetitive heavy

fibroin silk genes in new moth and caddisfly genomes. GigaByte. 2022:1–14. doi:10.46471/gigabyte.64.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8):907–915. doi:10.1038/s41587-019-0201-4.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):540–546. doi:10.1038/s41587-019-0072-8.

Kuznetsov D, Tegenfeldt F, Manni M, Seppey M, Berkeley M, Kriventseva EV, Zdobnov EM. 2023. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. Nucleic Acids Res. 51(D1): D445–D451. doi:10.1093/nar/gkac998.

Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. F1000Res. 6:1287. doi:10.12688/f1000research.12232.1.

Lasko PF, Ashburner M. 1988. The product of the Drosophila gene vasa is very similar to eukaryotic initiation factor-4A. Oxford University Press.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 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(10):4647–4654. doi:10.1093/molbev/msab199.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 14(1): e1005944. doi: 10.1371/journal.pcbi.1005944.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49(D1): D412–D419. doi:10.1093/nar/gkaa913.

Ode PJ, Strand MR. 1995. Progeny and Sex Allocation Decisions of the Polyembryonic Wasp Copidosoma floridanum.

Ohno H, Sakamoto T, Okochi R, Nishiko M, Sasaki S, Bono H, Tabunoki H, Iwabuchi K. 2019. Apoptosis-mediated vasa down-regulation controls developmental transformation in Japanese Copidosoma floridanum female soldiers. Dev Biol. 456(2):226–233. doi: 10.1016/j.ydbio.2019.09.005.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20(1):275. doi:10.1186/s13059-019-1905-y.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare. F1000Res. 9:304. doi:10.12688/f1000research.23297.2.

Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. 2018. Long reads: their purpose and place. Hum Mol Genet. 27(R2): R234–R241. doi:10.1093/hmg/ddy177.

Raz E. 2000. The function and regulation of vasa-like genes in germ-cell development. Genome Biol. 1(3): REVIEWS1017. doi:10.1186/gb-2000-1-3-reviews1017.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):245. doi:10.1186/s13059-020-02134-9.

Rhoads A, Au KF. 2015. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 13(5):278–289. doi: 10.1016/j.gpb.2015.08.002.

Sakamoto T, Nishiko M, Bono H, Nakazato T, Yoshimura J, Tabunoki H, Iwabuchi K. 2020. Analysis of molecular mechanism for acceleration of polyembryony using gene functional annotation pipeline in Copidosoma floridanum. BMC Genomics. 21(1):152. doi:10.1186/s12864-020-6559-3. https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-020-6559-3.

Schupbach T, Wieschaus E. 1986. Maternal-effect mutations altering the anterior-posterior pattern of the Drosophila embryo. Roux's Archives of Developmental Biology. 195(5):302–317. doi:10.1007/BF00376063.

Segoli M, Harari AR, Rosenheim JA, Bouskila A, Keasar T. 2010. The evolution of polyembryony in parasitoid wasps. J Evol Biol. 23(9):1807–1819. doi:10.1111/j.1420-9101.2010.02049.x.

Smith MS, Shirley A, Strand MR. 2017. Copidosoma floridanum (Hymenoptera: Encyrtidae) Rapidly Alters Production of Soldier Embryos in Response to Competition. Ann Entomol Soc Am. 110(5):501–505. doi:10.1093/aesa/sax056.

Soldà G, Suyama M, Pelucchi P, Boi S, Guffanti A, Rizzi E, Bork P, Tenchini ML, Ciccarelli FD. 2008. Non-random retention of protein-coding overlapping genes in Metazoa. BMC Genomics. 9. doi:10.1186/1471-2164-9-174.

Strand MR. 2003. Polyembryony. Carde R., Resch V., editors. San Diego: Academic Press.

Szathmáry E, Smith JM. 1995. The major evolutionary transitions. Nature. 374(6519):227–232. doi:10.1038/374227a0.

Uka D, Takahashi-Nakaguchi A, Yoshimura J, Iwabuchi K. 2013. Male soldiers are functional in the Japanese strain of a polyembryonic wasp. Sci Rep. 3(1):2312. doi:10.1038/srep02312.

Vladimir Zhurov, Tomislav Terzin, Miodrag Grbic. 2004. Early blastomere determines embryo proliferation and caste fate in a polyembryonic wasp. Nature. 432(7018):761–764. doi:10.1038/nature03030.

Whyte BA. 2021. The weird eusociality of polyembryonic parasites. Biol Lett. 17(4). doi:10.1098/rsbl.2021.0026.

Yamamoto D, Henderson R, Corley LS, Iwabuchi K. 2007. Intrinsic, inter-specific competition between egg, egg–larval, and larval parasitoids of plusiine loopers. Ecol Entomol. 32(2):221–228. doi:10.1111/j.1365-2311.2006.00857.x.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK. 2019. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 10(1):1523. doi:10.1038/s41467-019-09234-6. http://www.nature.com/articles/s41467-019-09234-6.

518

519

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

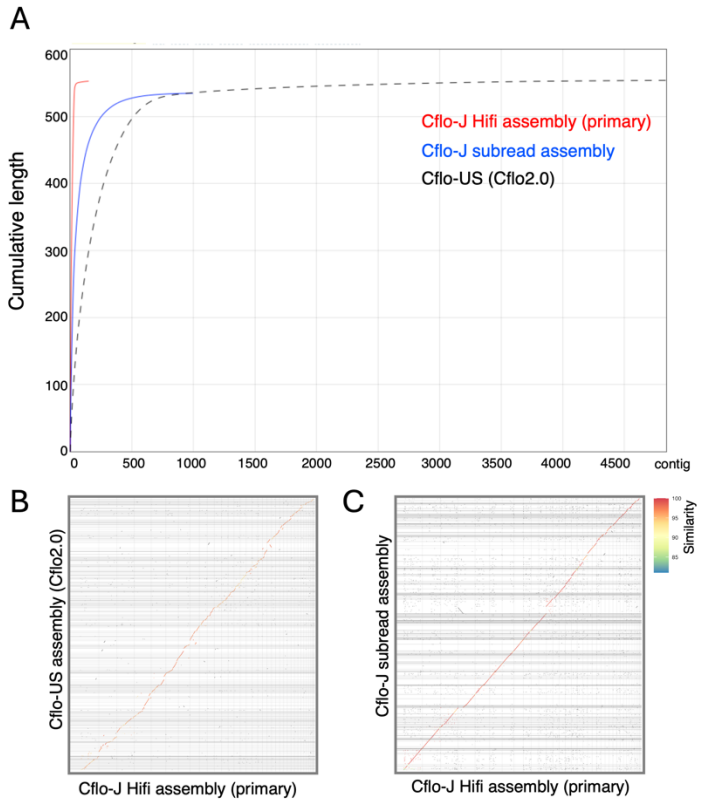
Table 3 Summary of functional annotation results

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

524

525 **Figures**



526

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

534

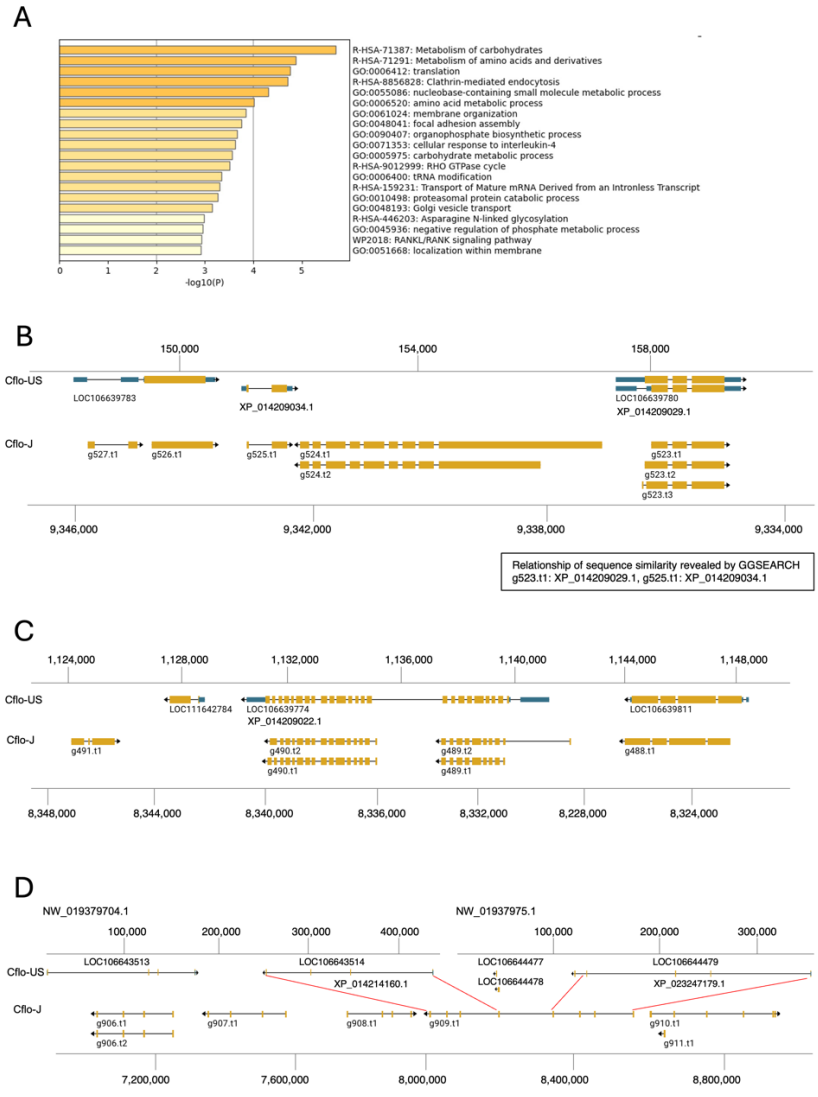


Figure 2

Features of genes with no corresponding proteins to Cflo-US. (A) The results of enrichment analysis (Metascape) of 149 transcripts with no corresponding proteins to Cflo-US and with corresponding proteins to human, mouse, *C. elegans*, and fly. Bars are colored differently according to the value of -log10 (P-value). (B-D) Representative gene model illustrating the genome annotation differences. Yellow and blue boxes indicate the ORF and UTR, respectively. Red lines indicate the region with similar exon structures.

A

```
XP_014219851.1      MEDESWDDESMAPAAPAPTNFSKGRGFTPANNNSYGDGGDQRGGSYGDNDRGGRSRGGGR      60
g8014.t1            MEDESWDDESTAPAAPAPTNFSKGRGFTPANNNSYGDGGDQRGGSYGGNDRGGRGRGGGR      60
BBI30140.1          MEDESWDDESTAPAAPAPTNFSKGRGFTPANNNSYGDGGDQRGGSYGGNDRGGRGRGGGR      60
                    ********** ************************************.******.*****

XP_014219851.1      GGRGSRGGGGGGYDNSNGGGWDNKKSNDNWNDGPKELPKAEDDWLAQGSGGGDDWLAQGN      120
g8014.t1            GGRGSRGGGGGGYDNSNGGGWDNKKSNDNWNDGPKDQPKAEDDWLAQGSGGGDDWLSQGN      120
BBI30140.1          GGRGSRGGGGGGYDNSNGGGWDNKKSNDNWNDGPKEQPKAEDDWLAQGSGGGDDWLSQGN      120
                    ************************************: *******************:***

XP_014219851.1      GGG----GFG--DDDDGGFGGGDRRGGRGGRGGRGGRGGRGGDRGGRGGRGGRNFGDNNR      174
g8014.t1            SGGGGGGGGFGDNDDDDGGFGGGDRRGGRGGRGGRGGRGGRGGDRGGRGGRGGRNFGDNNR      180
BBI30140.1          SGGGGGGGGFGDNDDDDGGFGGGDRRGGRGGRGGRGGRGGRGGDRGGRGGRGGRNFGDNNR      180
                    .**    ***  ************************************************

XP_014219851.1      DNDDDNGFGGDSGGFGDGNGDEGFGGGGRGRGRGRGGGGRGGRDGGDDRGGQEGGEGG      234
g8014.t1            DD-DENGFGGDSGGFGDSNGDEGFGGGGRGRGRGRGGGGRGGRDGGDDRGGQEGGEGG      239
BBI30140.1          DD-DDNGFGGDSGGFGDSNGDEGFGGGGRGRGRGRGXGGRGGRDGGDDRGGQEGGEGG      239
                    *: *:************.******************** **********************

XP_014219851.1      EPEKPREIYIPPDRSNDDETLFGSGTKAGINFSKYESIEVKVSGEDVPNKLNSFDEANLR      294
g8014.t1            EPEKPREIYIPPDRSNDDETLFGSGTKAGINFSKYESIEVKVSGEDVPKKLNSFDEANLR      299
BBI30140.1          EPXKPREIYIPPDRSNDDETLFGSGTKAGINFSKYESIEVKVSGEDVPNKLNSFDEANLR      299
                    ** ************************************************:************

XP_014219851.1      SLLVDNIKKSGYQIPTPIQKHGIPIVLAGRDLMGCAQTGSGKTAAFLIPIIHNLLLKPRE      354
g8014.t1            SLLVDNIKKSGYQIPTPIQKHGIPIVLAGRDLMGCAQTGSGKTAAFLIPIIHNLLLKPRE      359
BBI30140.1          SLLVDNIKKSGYQIPTPIQKHGIPIVLAGRDLMGCAQTGSGKTAAFLIPIIHNLLLKPRE      359
                    ************************************************************

XP_014219851.1      INDMGSLSTVEPRALILAPTRELAIQIHDECRKFSKDSVLKCCLIYGGTAVGHQLKQIFY      414
g8014.t1            IDDMGSLSTVEPRALILAPTRELAIQIHDECRKFSKDSVLKCCLIYGGTAVGHQLKQIFY      419
BBI30140.1          INDMGSLSTVEPRALILAPTRELAIQIHDECRKFSKDSVLKCCLIYGGTAVGHQLKQIFY      419
                    *:**********************************************************

XP_014219851.1      GCDLLVATPGRLKDFVGRGKVVFSAIEYLVLDEADRMIDMGFIGDVEMVIRHETMTPPGE      474
g8014.t1            GCDLLVATPGRLKDFVGRGKVVFSAIEYLVLDEADRMIDMGFIGDVEMVIRHETMTPPGE      479
BBI30140.1          GCDLLVATPGRLKDFVGRGKVVFSAIEYLVLDEADRMIDMGFIGDVEMVIRHETMTPPGE      479
                    ************************************************************

XP_014219851.1      RQTLMFSATFPRDIQELAVKFLNNYIFVAVGIVGSACSDIEQSFLEVKKSDKRSKLKEIL      534
g8014.t1            RQTLMFSATFPRDIQELAVKFLNNYIFVAVGIVGSACSDIEQSFLEVKKSDKRNKLKEIL      539
BBI30140.1          RQTLMFSATFPRDIQELAVKFLNNYIFVAVGIVGSACSDIEQSFLEVKKSDKRSKLKEIL      539
                    ************************************************************.******

XP_014219851.1      EEERNNDTLKGILVFVDQKRTADFIAAFLSDNNFPTTSIHGDRLQREREEALNDFKSGRM      594
g8014.t1            EEERNNDTLKGILVFVDQKRTADFIAAFLSDNNFPTTSIHGDRLQREREEALNDFKSGRM      599
BBI30140.1          EEERNNDTLKGILVFVDQKRTADFIAAFLSDNNFPTTSIHGDRLQREREEALNDFKSGRM      599
                    ************************************************************

XP_014219851.1      CILVATAVAARGLDIKNVRFVINFDLPKEIDEYIHRIGRTGRVGNKGKAVSFYDESIDNR      654
g8014.t1            CILVATAVAARGLDIKNVRFVINFDLPKEIDEYIHRIGRTGRVGNKGKAVSFYDESIDNR      659
BBI30140.1          CILVATAVAARGLDIKNVRFVINFDLPKEIDEYIHRIGRTGRVGNKGKAVSFYDESIDNR      659
                    ************************************************************

XP_014219851.1      IQPDLVKILEMANQPIPEWLASGAGYSNLGPSDFGAEDIRDFPNSSGAQGTLPVENDDAW      714
g8014.t1            IQPDLVKILEMANQPIPEWLASGAGSSNLGPSDFGAEDIRDFPNSSGAQGTLPVENDDAW      719
BBI30140.1          IQPDLVKILEMANQPIPEWLASGAGSSNLGPSDFGAEDIRDFPNSSGAQGTLPVENDDAW      719
                    ************************* ***********************************
```

B

Figure 3 Amino acid sequences genomic position and of *vasa* gene. Comparison of gene annotation between the Cflo-US and Cflo-J HiFi assemblies. (A) 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 in Cflo-J 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. (B) Genomic position of *vasa* with RNA-

551 Seq read coverages. The CDS regions are shown at the bottom, including the *vasa* and *cap'n'collar*

552 (*cnc*) exon structures. Blue and green arrows indicate the direction of transcription.

553