scientific data

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of wild silkmoth, *Bombyx mandarina*

Jung Lee[1]✉, Takashi Kiuchi[2], Katsushi Yamaguchi[3], Shuji Shigenobu[3], Atsushi Toyoda[4] & Toru Shimada[1,2]

The wild silk moth, *Bombyx mandarina*, is the closest relative of the domesticated silk moth, *Bombyx mori*. National BioResource Project of Japan (NBRP) maintains a *B. mandarina* strain derived from individuals captured at Sakado (Saitama, Japan) in 1982. Now, NBRP has developed chromosome replacement strains of *B. mori*. In each strain, one autosome of *B. mori* was replaced with the corresponding chromosome of *B. mandarina*. To facilitate the use of chromosome replacement strains and *B. mandarina* itself, we constructed a chromosome-level genome assembly of *B. mandarina*. Furthermore, we performed functional annotations of the genome assembly, i.e., transcriptome-based gene prediction, Assay for Transposase-Accessible Chromatin (ATAC)-seq, and PIWI-interacting RNA (piRNA)-targeted small RNA-seq. The assembly harbors 14,859 protein-coding genes and 628 piRNA clusters across three tissues: ovaries, testis, and embryos. ATAC-seq data comprehensively detected open chromosome regions, which will benefit when CRISPR/Cas9-mediated genome editing is conducted.

Background & Summary

B. mandarina is the closest relative of the domesticated silk moth, *B. mori*[1]. After the ancestor species of *B. mandarina* and *B. mori* was initially domesticated in ancient China, it was brought to other Asian regions for raw silk production. We can find *B. mandarina* populations outside China, such as in the Korean peninsula and Japan[2,3]. The phylogenetic relationships of *B. mandarina* populations in Asia are shrouded in mystery. Chinese *B. mandarina* has 28 homologous chromosomes, while Japanese *B. mandarina* has 27 pairs[4]. In Korea, located between China and Japan, we can find n = 28 and n = 27 individuals[3,5]. Therefore, phylogenetic studies using ribosomal DNA sequences supported the close relationship between Japanese and Korean populations[2]. However, studies using mitochondrial genome sequences have proposed a different hypothesis[3]: the Japanese population is more closely related to the *B. mandarina* population in southern China than the Korean population. This discrepancy might have resulted from the limited data exploited from rDNA or mitochondrial sequence analyses. Even though the chromosome numbers of Japanese and Korean *B. mandarina* are the same, it is still being determined whether the fused chromosomes in both populations have the exact evolutionary origin because any chromosomal genome assemblies of *B. mandarina* have not been constructed.

The National BioResource Project of Japan (NBRP) is an initiative that is obliged to maintain, manage, and attribute genetic resources in Japan. The silkworm section of the NBRP maintains not only mutant strains of *B. mori* but also wild silkmoths, including *B. mandarina*; the *B. mandarina* strain maintained by the NBRP originated from an individual captured in Sakado, Saitama in 1982, and has been repeatedly inbred by sib-cross for 42 years in The University of Tokyo and Kyushu University. This strain is probably the most highly purified *B. mandarina* strain (hereafter referred to as the "Sakado" strain). Moreover, NBRP has developed a series of consomic lines in which the *B. mori* chromosomes are replaced by *B. mandarina* chromosomes[6]. *B. mandarina* and *B. mori* differ in many physiological traits, such as larval motility and adult flight, which provide good material for genetic studies.

[1]Gakushuin University, Faculty of Science, Department of Life Science, Mejiro 1-5-1, Toshima-ku, Tokyo, 171-8588, Japan. [2]The University of Tokyo, Graduate School of Agricultural and Life Sciences: Yayoi, Bunkyo-ku, Tokyo, 113-8657, Japan. [3]National Institute for Basic Biology, Trans-Omics Facility, Nishigonaka 38, Myodaiji, Okazaki, 444-8585, Japan. [4]National Institute of Genetics, Comparative Genomics Laboratory, Advanced Genomics Center, 1111 Yata, Mishima, Shizuoka, 411-8540, Japan. ✉e-mail: yungu.ri@gakushuin.ac.jp

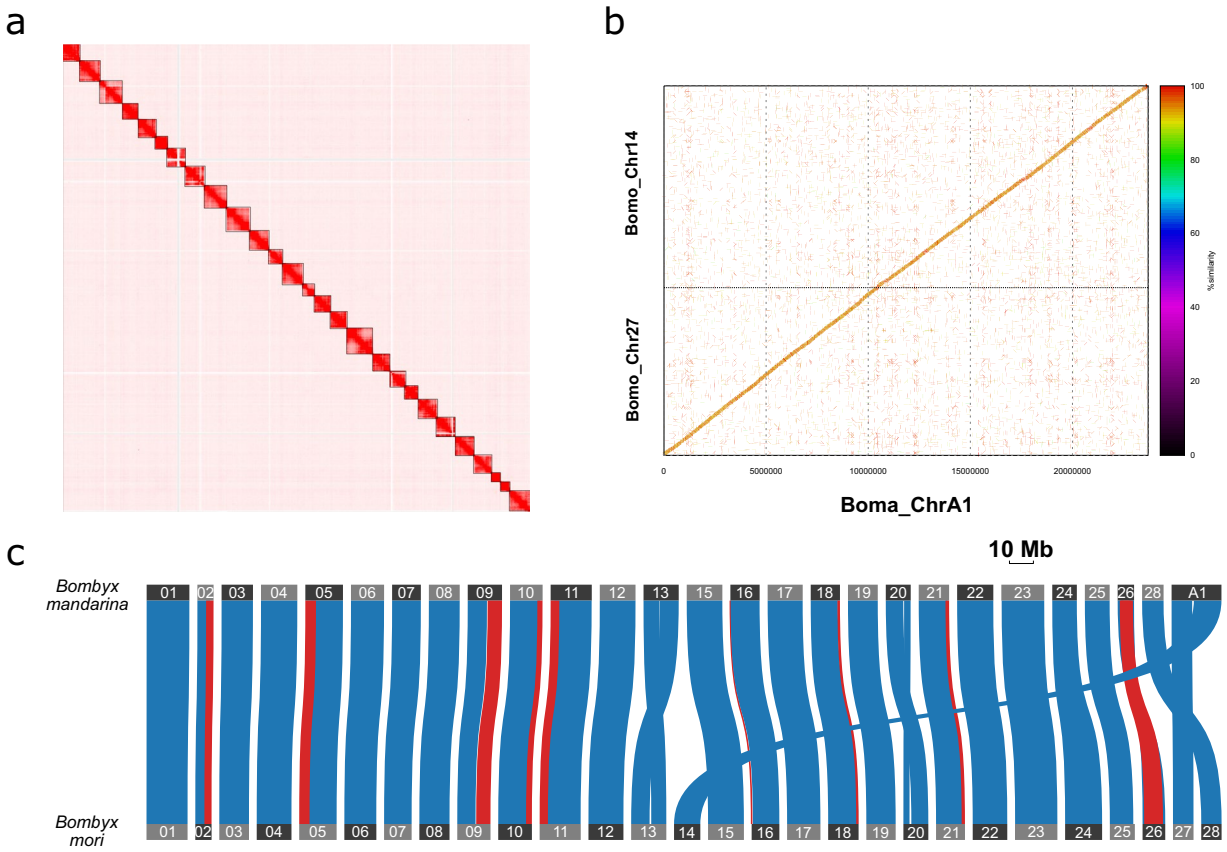

Fig. 1 A chromosome-scale genome assembly of *B. mandarina*. (**a**) Hi-C contact map of the genome assembly. Each block represents a Hi-C contact between two genomic loci. (**b**) Alignment of *B. mandarina* A1 chromosome and *B. mori* 14 and 27 chromosomes. (**c**) Single-copy orthologue (SCO) anchored synteny plot of *B. mandarina* and *B. mori* genomes. Each SCO was linked in a thin line. The lines in Inverted syntenies are shown in red while direction-conserved syntenies are in blue.

Therefore, we decided to construct the chromosome-scale genome assembly and annotate it for further phylogenetic and genetic studies. After we determined the genome sequences of the Sakado strain with long reads, the draft genome assembly was scaffolded with two scaffolding technologies: optical genome mapping and HiC-seq. As a result, we successfully constructed a chromosome-scale genome assembly of *B. mandarina*. As Banno *et al.* predicted[4], single-copy orthologue anchored genome-wide alignment of *B. mori* genome (accession No. GCF_030269925.1) and the *B. mandarina* genome (accession No. GCA_030267445.2) revealed that homologous chromosomes of chromosome 14 and 27 of *B. mori* fused into a single chromosome in *B. mandarina* (Fig. 1; hereafter we refer to this chromosome as "A1" chromosome)[4]. Transcriptome-based gene prediction and functional annotation were also performed, identifying 14,859 protein-coding genes in this assembly. For the future application of the CRISPR/Cas9, we performed embryonic ATAC-seq to associate the positions of protein-coding genes and open chromatin regions since the enzymatic activity of Cas9 is less efficient in heterochromatin regions[7]. In addition to ATAC-seq, we performed piRNA-targeted small RNA-seq to identify piRNA clusters (piCs) since piRNA is involved in the sexual differentiation of Lepidoptera[8].

Methods

Insects. *B. mandarina* (Sakado strain) was provided from NBRP silkworm (https://shigen.nig.ac.jp/silkwormbase/). The *B. mandarina* larvae were fed on fresh leaves of *Morus alba* under a long-day condition (16 h light/8 h dark) at 25 °C.

Genome assembly. Genomic DNA was extracted from silk glands of a male final instar larvae of *B. mandarina* using Genomic-tip 100/G (QIAGEN). Extracted HMW DNA was submitted for SMRTbell library preparation and the Illumina paired-end library preparation. The prepared libraries were sequenced on the PacBio RS II or illumina HiSeq2500 platform (Supplementary Table 1). The draft assembly was constructed with FALCON (v 0.7) and FALCON-Unzip (v 0.4.0)[9]. FALCON was performed with the following parameters:

genome_size = 500000000
seed_coverage = 30
length_cutoff_pr = 10000.

Assembly	Species	Sex	Karyotype	Scaffolds	Scaffold N50 (bp)	Assembly Size (bp)	Accession No.
Bma-NBRP_v1.0	*Bombyx mandarina*	Male (ref. [4])	2n = 54	27	16,433,000	419,602,541	GCA_030267445.2

Table 1. The basic metrics of the genome assembly of *B. mandarina*.

Paired-end reads were used to polish the draft assembly using Pilon (v 1.22) with default settings[10]. The polished assembly was initially scaffolded with optical genome mapping as previously described[11]: Genomic DNA was isolated from the pupa immediately after the pupation for the optical genome mapping. DNA isolation was conducted using the Bionano Prep Animal Tissue DNA Isolation Fibrous Tissue Protocol (Bionano Genomics). DLE-1 was used for the direct label stain. The DNA labeling was conducted using the Bionano Prep Direct Label and Stain Protocol. The labeled samples were scanned on the Bionano Saphyr system using Saphyr Chip G2.3. The obtained data were analyzed using Bionano Access (v 1.8.2)[12] and Bionano Solve (v 3.8.2)[13]. In the single enzyme pipeline, to create a.cmap file (describing the labeled genomic regions from the obtained data), we used a wrapper script "pipelineCL.py" bundled with the Bionano Solve with default settings. The resulting.cmap file was submitted to "hybridScaffold.pl", also bundled with the Bionano Solve, which bridges the contigs based on the location information for the labelled regions in the.cmap file. "hybridScaffold.pl" was executed with the option "-B 2 -N 2". The obtained optical genome mapping data was deposited at DDBJ[14]. The second scaffolding was conducted using Hi-C sequencing. The Hi-C library was prepared with a Dovetail Omni-C kit on pupal genomic DNA. Fastp (v 0.20.1)[15] was used for the quality trimming of Hi-C reads with '-q 25 -l 50.' The Hi-C scaffolding was conducted through juicer (v 2.0)[16] – 3D-DNA (v 180922)[17] pipeline with default settings. Hi-C contact map shows 27 distinct domains (Fig. 1a), which equals the number of chromosomes of Sakado strain[4]. The basic metrics of the assembly are summarised in Table 1.

Comparison of *B. mandarina* genome structure with the *B. mori* genome. In comparison to the *B. mori* genome (accession No. GCF_030269925.1)[18] using Mummer (v 4.0.0)[19], the so-called "M chromosome" turned out to be homologous to *B. mori* chromosomes 14 and 27 (Fig. 1b). Following the nomenclature of the chromosomes of *T. varians*[11], a bombycid species (i.e. the prefix 'A' is used if a fusion has occurred in comparison to the *B. mori* chromosome), this chromosome was termed "chromosome A1." In addition, we performed single-copy orthologue anchored genome-wide alignment of the and the *B. mandarina* genome (accession No. GCA_030267445.2)[20]. The single-copy orthologues shared with the two assemblies were extracted by BUSCO (v 5.4.6)[21]. MCScanX[22] was used to create a riparian plot (Fig. 1c).

Repetitive elements annotation in the genomes. Repetitive annotation of the *B. mandarina* genome was briefly summarised here: repetitive elements in the genome assembly were identified using RepeatModeler (v 2.0.4)[23] with the "-LTRstruct" option for performing an LTR structural search. In the following repeat annotation process using RepeatMasker (v 4.1.2)[24], the resulting "consensi.fa.classified" file was specified by "-lib" option[24]. Among the repetitive elements, LTR, non-LTR (LINE or SINE), DNA transposons, and rolling circles were extracted and the density information of those repetitive groups is visualized by circlize (v 0.4.16)[25] (Fig. 2).

RNA sample preparation for RNA-seq, sRNA-seq, and Isoform-seq (Iso-seq). RNA samples derived from 12 different tissues (Supplementary Table 2) were prepared precisely as previously described[11]. Total RNA was extracted using TRIzol reagent (Invitrogen) according to the manufacturer's protocol. Embryos were sampled 24 hours after oviposition. The two aliquots of testis and ovary-derived RNA samples were subjected to RNA-seq and sRNA-seq, respectively. The three aliquots embryo-derived RNA samples were subjected to sRNA-seq, RNA-seq, and Iso-seq, respectively.

Library preparation for RNA-seq, sRNA-seq and Iso-seq. The sRNA-seq library was prepared using TruSeq small RNA kit (illumina) according to the manufacturer's protocol with a slight modification. To target piRNA, a region of 147–158 nucleotides was extracted in the purification step of the cDNA construct using BluePippin (Sage Science). The constructed library was sequenced on the illumina HiSeq2500 platform (illumina). Except for an RNA sample from embryos, the RNA-seq libraries were prepared using TruSeq stranded mRNA kit (illumina) according to the manufacturer's protocol. The embryonic RNA-seq library was prepared using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs) and the NEBNext® Ultra™ ll Directional RNA Library Prep Kit (New England BioLabs) according to the manufacturer's protocol. The constructed RNA-seq libraries were sequenced on the illumina NovaSeq6000 platform (illumina). For Iso-seq, the library was constructed using Sequel Iso-seq Express Template Prep (Pacific Bioscience) according to the manufacturer's protocol. The constructed library was sequenced on the PacBio Sequel platform (PacBio).

Transcriptome-based gene prediction. BRAKER3 (v 3.0.8) was used for gene prediction[26,27]. The RNA-seq and Iso-seq data were submitted to BRAKER3 separately[28], and Tsebra finally merged their respective prediction[29]. The detailed information on transcriptome data is summarised in Supplementary Table 2. Quality trimming for short read data was conducted using fastp (v 0.20.1)[15] with following options: '-q 28 -l 80. ' Trimmed short read data were submitted to BRAKER3 using the '--rnaseq_sets_ids' option. The short reads were aligned to the genome assembly by hisat2 (v 2.2.1)[30]. Iso-seq data were generated consensus for each read cluster according to the following procedure[31]: Iso-seq subreads were converted to circular consensus sequences (ccs) using ccs v 6.4.0 with options '--minLength 10 --maxLength 100000 --minPasses 0 --minSnr 2.5 --minPredictedAccuracy 0.0.' lima (v 2.7.1) was used to remove primer sequences from the CCSs with options '--isoseq --peek-guess

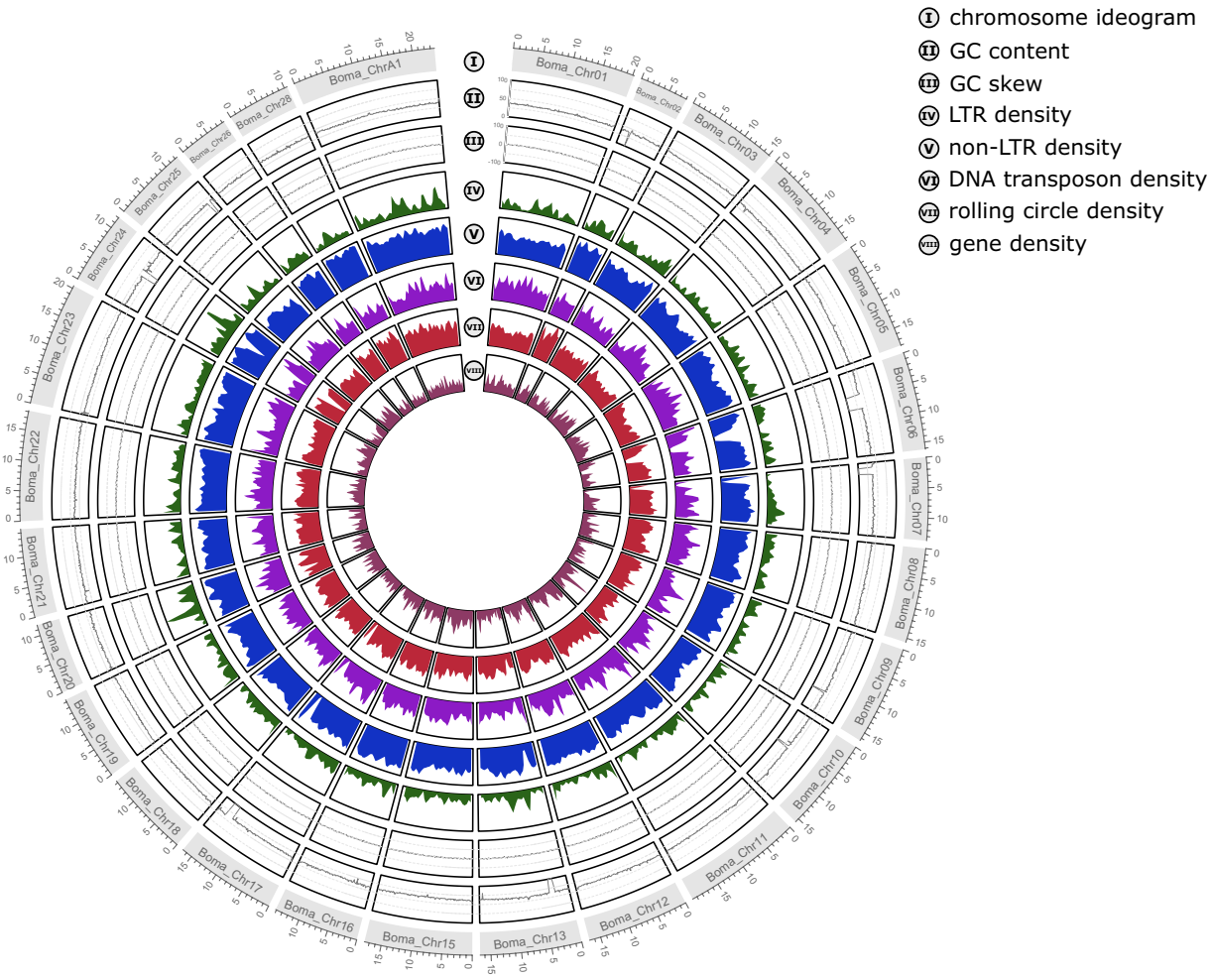

Fig. 2 General genome annotation information. Summary of *B. mandarina* genome characteristics. The outermost to the innermost circle are I. chromosome ideograms; II. GC content; III. GC skew; IV. LTR element density; V. non-LTR retrotransposon density; VI. DNA transposon density; VII. rolling circle density; and VIII. gene model density.

--ignore-biosamples.' After the trimming of adaptors, PolyA tail trimming and concatemer removal were performed by isoseq3 (v 3.8.2) in 'refine' mode with option '--require-polya.' Finally, isoform-level clustering was conducted by isoseq3 in 'cluster' mode with option '--use-qvs.' The resulting clustered.bam file was submitted to BRAKER3. Before gene prediction with Iso-seq data, BUSCO analysis on the genome assembly was conducted to obtain complete and single-copy BUSCO sequences[30]. Complete and single-copy BUSCO sequences were submitted to BRAKER3 with an Iso-seq-derived bam file. Since we had two Iso-seq datasets (Supplementary Table 2), we ran BRAKER3 for them separately. The resulting gene models were also submitted to BUSCO analysis[30], scoring 94.5% completeness (Fig. 3a). The basic metrics of gene models were summarised in Table 2.

Functional annotation of gene models. The deduced amino acid sequences of gene models were submitted to EnTAP[6] for functional annotation. A protein similarity search was conducted against the latest complete UniProtKB/TrEMBL protein data set and complete UniProtKB/Swiss-Prot data set using diamond (v 0.9.14)[32]. A protein orthology search was also conducted against the Evolutionary genealogy of genes: Non-supervised Orthologous Groups (EggNOG) databases30 to assign Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) terms, and protein domains from pfam[33] and smart[34]. Additional family and domain search was performed against tigrfam[35], sfld[36], hamap[37], cdd[38], superfamily[39], prints[40], panther[41], and gene3d[42] using InterProScan (v 5.68-100)[43]. The results of functional annotation are summarized in Table 3. The top 10 GOs assigned to the gene models are shown in Fig. 3b without distinguishing between molecular function, biological process, and cellular component. The top 10 GOs for each category are shown in Supplementary Fig. 1.

ATAC library preparation and data processing. Another batch of early embryo samples was subjected to Isoform-seq (Iso-seq), and small RNA-seq was subjected to ATAC-seq. Fragmentation and amplification of the ATAC-seq libraries were conducted according to Buenrostro *et al.*[44]. The constructed libraries were sequenced on the Illumina HiSeq. ATAC-seq reads were pretreated with fastp and mapped to the genome with bwa-mem2 (v 2.2.1)[45]. Alignments containing mismatches were then removed using bamutils (v 0.5.9)[46]. Next, we removed

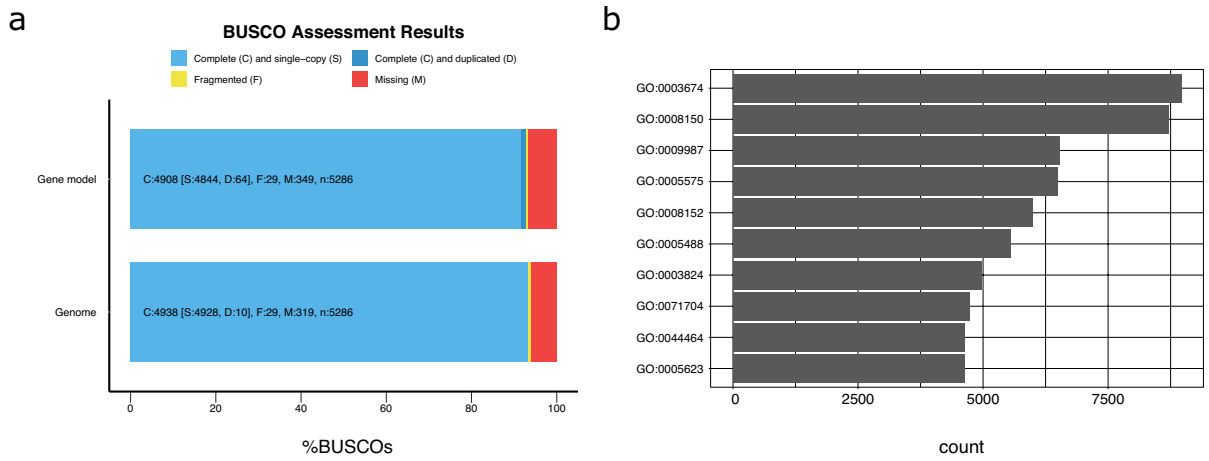

Fig. 3 BUSCO assessment and the top 10 GO assignments of the predicted gene models. (**a**) BUSCO scores of the gene models (top) and the genome assembly (bottom). (**b**) Overall top 10 GO assignments to gene models.

No. of protein coding gene	14,859
Average CDS length [bp]	1526.6
Average exon length [bp]	230.50
Average intron length [bp]	1791.4

Table 2. Statistical summary of the constructed gene models.

		Similarity search			Ontology search		total	
		EggNOG	TrEMBL	Swiss-prot	EggNOG**	InterPro		
aligned	informative	11,717	11,054	7,135	9,839	12,153	annotated***	13,111
	uninformative*	0	2,039	176	2,895	—		
unaligned		3,142	1,766	7,548	2,125	2,706	unannotated***	1,748

Table 3. Brief summary of functional annotation. *When the query sequences were aligned to sequences whose description contains any of conserved/predicted/unnamed/hypothetical/putative/unidentified/uncharacterized/unknown/uncultured/uninformative, such alignment was categorized as "uninformative," and the query sequence was treated as an unannotated sequence. **In this column, queries with at least one GO term were treated as "Informative," while queries without GO terms were treated as "Uninformative" "Unaligned" in this column means queries without protein family assignment. *** "Annotated" means at least one match yielded from any of the databases. "Unannotated" means no match yielded from all databases.

duplicated reads using GATK MarkDuplicates (v 4.1.7)[47]. The resulting bam files were converted to bigwig files using deepTools bamCoverage (v 3.5.1)[48]. Heatmap was created using deepTools computeMatrix, and the starting point of the gene model was set to the reference point (Fig. 4).

Small RNA mapping. The small RNA reads were trimmed using Trim Galore (v 0.6.6)[49] in small RNA mode. The trimmed small RNA reads were mapped to the genome assembly, allowing up to 3 nucleotide mismatches using Hisat2 (v 2.1.0)[30] and ngsutils (v 0.5.9)[46]. The information for each library is summarized in Supplementary Table 2.

piRNA cluster detection. The piC detection was performed as previously described[11,50]. proTRAC (v 2.4.4)[50] was used with options '-clsize 5000 -pimin 23 -pimax 29 -1Tor10A 0.3 -1Tand10A 0.3 -clstrand 0.0 -clsplit 1.0 -distr 1.0-99.0 -spike 90-1000 -nomotif -pdens 0.05.' As a result, we successfully identified 560 piRNA clusters in the three tissues (Fig. 5). The identity of piC is defined by the two nearest gene models. If multiple piCs were predicted between such two genes, such piCs were treated as a single piC. The genomic positions of piCs identified in testes, ovaries, and early embryos were visualized by RIdeogram (v 0.2.2)[51] (Fig. 5a). The aggregation relationship of those piCs was visualized by ComplexUpset (v 1.3.3)[52] (Fig. 5b).

Data Records

The raw sequence data reported in this paper have been deposited in DDBJ. Genomic data for the draft assembly were deposited under the accession code PRJDB5778[53] while genomic data for scaffolding were deposited under the accession code PRJDB13954[14]. The genome assembly is available under the accession code GCA_030267445.2[20]. Except for embryonic transcriptome data, all transcriptome data. i.e. RNA-seq and Iso-seq

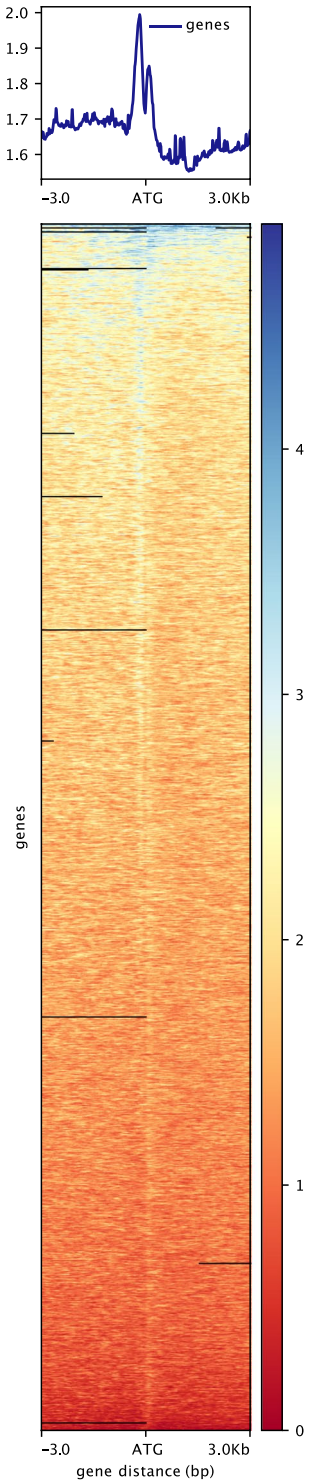

Fig. 4 Heatmap around gene bodies of ATAC-seq on early embryos. The ATAC-seq peaks of 3 kb upstream and downstream from the start codon of protein-coding genes are shown. The peaks of ATAC-seq are concentrated at the 5′ end of the gene body, which is a typical result of ATAC-seq.

data were registered under the accession code PRJDB13954[14]. Embryonic RNA-seq, ATAC-seq and Iso-seq data are available under the accession code PRJDB13955[54]. The annotated gene models have been deposited to the figshare repository[55].

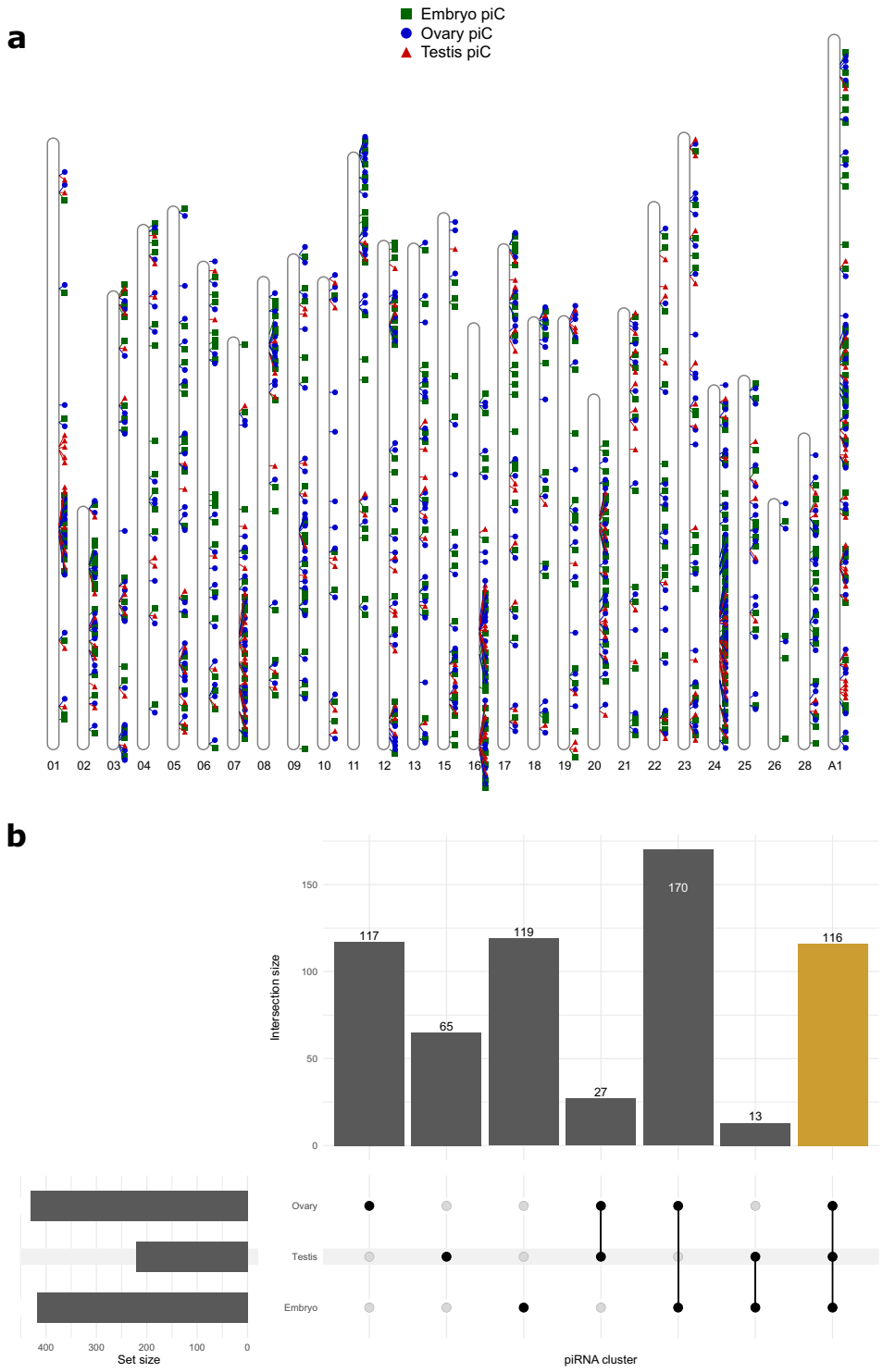

Fig. 5 piRNA clusters on *B. mandarina* genome. (**a**) piCs distribution detected in early embryos (box), pupal ovary (circle), and pupal testis (triangle). (**b**) UpSet plot visualising piCs that are each assigned to each tissue. The vertical bars correspond to the intersections. When the circles corresponding to tissues are connected by a line, the bar above circles represents the number of piCs commonly identified in concerning tissues. For example, the yellow bar indicates the number of piCs identified in all tissues. The identity of piCs was defined by the nearest two gene models: When comparing piCs identified in different tissues, if the nearest upstream and downstream gene models are the same, those piCs were treated as the same piC.

Technical Validation

BUSCO (v 5.4.6)[21] with lepidoptera_odb10 lineage dataset was used to assess the quality of gene models. For comparison, the results are summarised in Fig. 3, together with the results of BUSCO analysis for the genome assembly. 92.8% of the complete and single-copy BUSCO sequences were present in the gene models, while 93.4% of the complete and single-copy BUSCO sequences were in the genome assembly. BUSCO completeness scores were almost the same between the genome assembly and the gene model, suggesting that the gene prediction process is highly accurate across all genome regions.

Code availability

Programs exploited in this study were executed with the default parameters except where otherwise specified in the Methods section.

Received: 18 September 2024; Accepted: 2 January 2025;
Published online: 07 January 2025

References

1. Xiang, H. *et al.* The evolutionary road from wild moth to domestic silkworm. *Nat Ecol Evol* **2**, 1268–1279 (2018).
2. Kawanishi, Y. *et al.* Sequence Comparison of *Mariner*-like Elements among the Populations of *Bombyx mandarina* Inhabiting China, Korea and Japan. *J Insect Biotechnol Sericology* **76**, 2_79–2_87 (2007).
3. Kim, M.-J. *et al.* Phylogeographic Relationships among Bombyx mandarina (Lepidoptera: Bombycidae) Populations and Their Relationships to B. mori Inferred from Mitochondrial Genomes. *Biology (Basel)* **11**, 68 (2022).
4. Banno, Y., Nakamura, T., Nagashima, E., Fujii, H. & Doira, H. M chromosome of the wild silkworm, Bombyx mandarina (n = 27), corresponds to two chromosomes in the domesticated silkworm, Bombyx mori (n = 28). *Genome* **47**, 96–101 (2004).
5. Kawanishi, Y. *et al.* Method for rapid distinction of *Bombyx mandarina* (Japan) from *B. mandarina* (China) based on rDNA sequence differences. *J Insect Biotechnol Sericology* **77**, 2_79–2_85 (2008).
6. Fujii, T. *et al.* Development of interspecific semiconsomic strains between the domesticated silkworm, *Bombyx mori* and the wild silkworm, *B. mandarina*. *J Insect Biotechnol Sericology* **90**, 2_033–2_040 (2021).
7. Jain, S. *et al.* TALEN outperforms Cas9 in editing heterochromatin target sites. *Nat Commun* **12**, 4–13 (2021).
8. Kiuchi, T. *et al.* A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* **509**, 633–636 (2014).
9. PACIFIC BIOSCIENCES. FALCON and FALCON-Unzip. https://github.com/PacificBiosciences/FALCON/.
10. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, (2014).
11. Lee, J. *et al.* W chromosome sequences of two bombycid moths provide an insight into the origin of Fem. *Mol Ecol* **33**, 1–12 (2024).
12. Bionano Genomics, I. Bionano Access Software. https://bionano.com/access-software/.
13. Bionano Genomics, I. Bionano Solve software. https://bionano.com/software-downloads/.
14. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRP009939 (2023).
15. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
16. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98 (2016).
17. Dudchenko, O. *et al. De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds*. https://www.science.org.
18. Lee, J. & Shimada, T. Comparative W chromosome sequences between two bombycid moths, Bombyx mori and Trilocha varians. *Assembly* https://identifiers.org/assembly:GCF_030269925.1 (2023).
19. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput Biol* **14**, 1–14 (2018).
20. Lee, J. & Shimada, T. Japanese Bombyx mandarina genome project. *Assembly* https://identifiers.org/assembly:GCA_030267445.2 (2023).
21. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. 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**, 4647–4654 (2021).
22. Wang, Y. *et al.* MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* **40**, 1–14 (2012).
23. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457 (2020).
24. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013–2015.
25. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in. *R. Bioinformatics* **30**, 2811–2812 (2014).
26. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
27. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
28. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* **42**, 1–8 (2014).
29. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 1–12 (2021).
30. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357–360 (2015).
31. Brůna, T., Gabriel, L. & Hoff, K. J. Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA (2024).
32. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2014).
33. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
34. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res* **46**, D493–D496 (2018).
35. Haft, D. H. *et al.* TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res* **29**, 41–43 (2001).
36. Akiva, E. *et al.* The Structure-Function Linkage Database. *Nucleic Acids Res* **42**, 521–530 (2014).
37. Pedruzzi, I. *et al.* HAMAP in 2015: Updates to the protein family classification and annotation system. *Nucleic Acids Res* **43**, D1064–D1070 (2015).
38. Wang, J. *et al.* The conserved domain database in 2023. *Nucleic Acids Res* **51**, D384–D388 (2023).
39. Pandurangan, A. P., Stahlhacke, J., Oates, M. E., Smithers, B. & Gough, J. The SUPERFAMILY 2.0 database: A significant proteome update and a new webserver. *Nucleic Acids Res* **47**, D490–D494 (2019).

40. Attwood, T. K. *et al*. The PRINTS database: A fine-grained protein sequence annotation and analysis resource-its status in 2012. *Database* **2012**, 1–9 (2012).
41. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res* **47**, D419–D426 (2019).
42. Lewis, T. E. *et al*. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res* **46**, D435–D439 (2018).
43. Jones, P. *et al*. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
44. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* **109**, (2015).
45. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. in *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* 314–324. https://doi.org/10.1109/IPDPS.2019.00041 (IEEE, 2019).
46. Breese, M. R. & Liu, Y. NGSUtils: A software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494–496 (2013).
47. van der Auwera, G. & O'Connor, B. D. *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. (O'Reilly Media, Incorporated, 2020).
48. Ramírez, F. *et al*. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160–W165 (2016).
49. Krueger, F. *Trim Galore*. https://github.com/FelixKrueger/TrimGalore (2020).
50. Rosenkranz, D. & Zischler, H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**, 5 (2012).
51. Hao, Z. *et al*. RIdeogram: Drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput Sci* **6**, 1–11 (2020).
52. Michał, K. ComplexUpset.
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRP004537 (2018).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRS302564 (2023).
55. Lee, J. & Shimada, T. Bombyx mandarina (Sakado strain) gene models. *figshare* https://doi.org/10.6084/m9.figshare.26940925 (2024).

Acknowledgements

Insects were donated from Kyushu University according to a Grant-in Aid "National BioResource Project (NBRP, RR2002), Silkworm Genetic Resources" for Scientific Research from the Ministry of Education, Science, Sports and Culture of Japan. This study was supported by JSPS KAKENHI Grant Numbers 20K15535 and 24K17900 to J.L and JSPS KAKENHI Grant Number J18H03949 to T.S. This project was also supported by the NBRP Genome Information Upgrading Program in 2016.

Author contributions

J.L. designed the research plan, performed RNA extraction, analyzed the obtained data, and wrote the manuscript. T.S. and T.K. also designed this research plan. T.S. also performed the data analysis. T.K. reared *B. mandarina* larvae and collected their silk glands for genome sequencing. A.T. extracted the genomic DNA from the sample and was responsible for library preparation and sequencing operation. K.Y. and S.S. prepared the Hi-C seq library.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04395-0.

Correspondence and requests for materials should be addressed to J.L.