scientific data

OPEN

DATA DESCRIPTOR

Comprehensive genome annotation of *Bombyx mori* p50ma strain, a newly developed standard strain

Jung Lee[1]✉, Toshiaki Fujimoto[2,4], Ken Sahara[2], Atsushi Toyoda[3] & Toru Shimada[1]

B. mori **is a model species of lepidopteran insects. The genome assembly has been successively updated since the whole genome sequences were first determined in 2004. In addition, chromosome-scale genome assemblies of not only standard strains but also practical strains have been reported. We successfully constructed a chromosome-scale female genome assembly of p50ma, a standard strain developed by Kyushu University. This assembly is now certified as a 'reference' in the NCBI datasets. To improve the usability of this strain, here we report the achievement of gene model construction based on the transcriptome information, followed by functional annotation. The assembly harbours 16,295 protein-coding genes. In addition, to improve gene knockout efficiency, we performed ATAC-seq in early embryos and comprehensively identified open chromatin regions. Finally, small RNA-seq (sRNA-seq) targeting PIWI-interacting RNA (piRNA) was performed in testes, ovaries, and early embryos to identify piRNA clusters comprehensively. These data will increase the usability of p50ma as a standard strain and facilitate NBRP users to exploit this strain.**

Background & Summary

B. mori is a model species of lepidopteran insects. In 2004, the Chinese and Japanese groups independently reported the first draft genome assembly of *B. mori*[1,2]. Since then, nine assemblies of different strains have been registered in the NCBI database. Researchers constructed these assemblies using male individuals as genomic DNA donors because the female-specific W chromosome is occupied by repetitive sequences[3,4]. Short-read sequencing or noisy long-read sequencing cannot correctly assemble the W chromosome: since repetitive sequences on the W chromosome are longer (>300 bp) than short reads for a single unit[5], assembly by short reads would be easily confused.

Neither can noisy long reads such as PacBio continuous long reads and Oxford nanopore long reads accurately assemble the W chromosome. Our preliminary attempt to assemble the female genome with nanopore long reads managed to scaffold the W-linked sequences into a single sequence, but there were gaps of 3.3 Mbp in total (Supplementary Fig. 1a, Supplementary Table 1)[6–40]. This result contrasts the later assembly with PacBio's highly accurate long high-fidelity (HiFi) reads[41], harbouring only 46 bp gaps on the W chromosome. In addition, the preliminary attempt failed in assembling some autosomal regions (supplementary Fig. 1b–d) even though the average read length and N50 read length of nanopore reads were longer than HiFi long reads used in later trials (supplementary Table 2).

Instead of exploiting noisy long reads, we determined to take advantage of the PacBio HiFi long reads[42], which led to the first successful assembly of the female genome at chromosome scale[42]. The donor strain was "p50ma", a new standard strain that Kyushu University and the National BioResource Project (NBRP) silkworm recently developed. Here, we present the annotation information of the p50ma female genome assembly to facilitate the use of this strain.

Transcriptome-based gene prediction identified 16,295 protein-coding genes in p50ma genome. The following functional annotation was also performed using EnTAP[43]. Although application examples of

[1]Gakushuin University, Faculty of Science, Department of Life Science, Mejiro 1-5-1, Toshima-ku, Tokyo, 171-8588, Japan. [2]Laboratory of Applied Entomology, Faculty of Agriculture, Iwate University, Ueda 3-18-8, Morioka, 020-8550, Japan. [3]National Institute of Genetics, Comparative Genomics Laboratory, Advanced Genomics Center, 1111 Yata, Mishima, Shizuoka, 411-8540, Japan. [4]Present address: Laboratory of Silkworm Genetic Resources, Institute of Genetic Resources, Kyushu University Graduate School of BioResources and Bioenvironmental Science, Motooka 744, Nishi-ku, Fukuoka, 819-0395, Japan. ✉e-mail: yungu.ri@gakushuin.ac.jp

CRISPR/Cas9-mediated genome editing in p50ma have not been reported, applying genome editing techniques should be a prerequisite for promoting further use of p50ma as a standard strain. Since Cas9 is known to be less efficient in heterochromatin regions[44], we performed embryonic ATAC-seq to identify open chromatin regions.

Finally, we performed piRNA-targeted small RNA-seq to identify piRNA clusters (piCs) in early embryos, testes, and ovaries. BmN4, one of the cultured cell lines derived from *B. mori* ovaries, has been an excellent platform for piRNA research[45–47]. However, high usability of BmN4 made researchers keep a wide berth of piRNA study using living animals. In this respect, the situation in Lepidoptera is quite different from that of genus *Drosophila*, where piRNA clusters in each species have been comprehensively defined[48]. Since we considered that any studies which are conducted in the context of ignoring organisms, piCs in the p50ma genome were identified in a tissue-comprehensive manner. These datasets will increase the usability of this strain as a platform for piRNA research.

Methods

Insect strain background and rearing conditions. *B. mori* strain, p50ma was provided from NBRP silkworm (https://shigen.nig.ac.jp/silkwormbase/). Kyushu University developed p50ma strains by hybridization of p50 strain and T-series strain: F_1 individuals were repeatedly sib-crossed for at least twenty generations. 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 of p50ma with nanopore long reads. Genomic DNA was extracted by a p50ma female pupa using NucleoBond HMW DNA (Macherey-Nagel). The following adaptor ligation was conducted using 1D Ligation Sequencing Kit SQK-LSK110 (Oxford Nanopore Technologies, UK). The prepared sample was sequenced using Oxford Nanopore PromethION (Oxford Nanopore Technologies). Basic statistics of raw reads were summarized in Supplementary Table 1. Quality trimming was performed using porechop (v 0.2.3_seqan2.1.1)[49]. The draft assembly was Flye assembler (v 2.9-b1768)[50]. Optical genome mapping was conducted as previously described[6]: 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). We used two enzymes, namely, DLE-1 and Nt.BsqQI for labelling. The labelling procedure was conducted according to the Bionano Prep Direct Label and Stain Protocol. The labelled 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)[51] and Bionano Solve (v 3.8.2)[52]. In the two-enzyme pipeline, the process of creating the.cmap files was the same as in the one-enzyme pipeline. The resulting two.cmap files are submitted to runTGH.R (also bundled with Bionano Solve) with default settings. The.cmap files were deposited in DDBJ[53]. Mummer (v 4.0.0)[54] was used to compare the nanopore-reads derived assembly, and HiFi-reads derived assembly[6].

Repetitive elements annotation in the genomes. Our group[42] previously annotated repetitive elements of p50ma genome as briefly summarized here: repetitive elements in the genome assembly were identified using RepeatModeler (v 2.0.4)[55] with the "-LTRstruct" option for performing an LTR structural search. The annotated elements were specified and masked using RepeatMasker (v 4.1.2)[56] with default settings. LTR, non-LTR (LINE or SINE), DNA transposons, and rolling circles were extracted among the repetitive elements. The density information of those repetitive groups is visualized by circlize (v 0.4.16)[57] (Fig. 1). GC content and GC skew did not differ significantly among chromosomes, with GC content averaging about 38.3% (Fig. 1). However, the GC content was much higher in the W chromosome, at about 46.1%, reflecting the characteristics of W chromosomes to accumulate LTR retrotransposons[42] (Fig. 1).

RNA sample preparation for sRNA-seq and Iso-seq. All RNA samples were prepared precisely as previously described[42]. 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 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, USA). The constructed library was sequenced on the illumina Hiseq2500 platform (illumina, USA). RNA-seq library was prepared using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs) and NEBNext® Ultra™ ll Directional RNA Library Prep Kit (New England BioLabs) according to the manufacturer's protocol. The constructed library was sequenced on the illumina Novaseq6000 platform (illumina, USA). For Iso-seq, the library was constructed using Sequel Iso-seq Express Template Prep (Pacific Bioscience, USA) according to the manufacturer's protocol. The constructed library was sequenced on the PacBio Sequel platform (Pacific Bioscience, USA).

Transcriptome-based gene prediction. BRAKER3 (v 3.0.8) was used for gene prediction[58,59]. Thirty RNA-seq datasets[60] were downloaded from the NCBI SRA database. In addition to RNA-seq data, Iso-seq data were submitted to BRAKER3 separately[61], and Tsebra[62] finally merged their respective prediction. The detailed information on transcriptome data is summarised in Supplementary Table 3. Quality trimming for short read data was conducted using fastp (v 0.20.1)[63] with the following options: '-q 28 -l 80'. Trimmed short read data were submitted to BRAKER3 using the '--rnaseq_sets_ids' option. Then, short reads were aligned to the genome assembly by HISAT2 (v 2.2.1)[64]. The alignment rate of each dataset is summarized in Supplementary Table 1. Iso-seq data

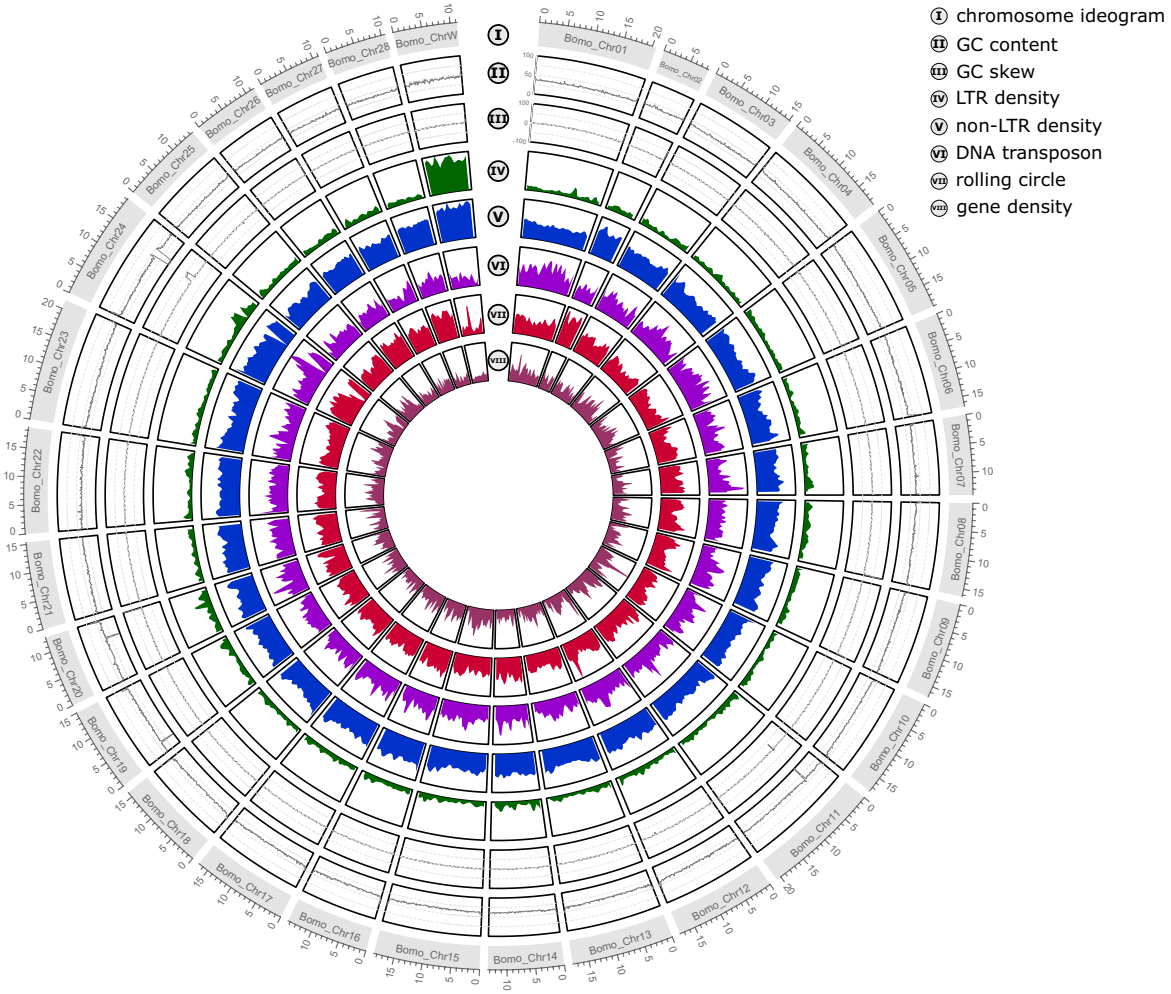

Fig. 1 General genome annotation information. Summary of p50ma genome characteristics. The outermost to the innermost circle shows I. chromosome ideograms; II. GC content; III. GC skew; IV. LTR element density; V. non-LTR retrotransposon density; VI. DNA transposon density; VII. rolling circle density; and VIII. gene model density.

were generated consensus for each read cluster according to the following procedure[65]: 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--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, IsoSeq3 conducted isoform-level clustering 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[42,66]. 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 3), we ran BRAKER3 for them separately. BUSCO analysis[66] on the constructed gene models scored 97.8% completeness (Fig. 2a). Basic statistics of the predicted gene models were summarised in Table 1.

Functional annotation of gene models. The deduced amino acid sequences of gene models were submitted to EnTAP[43] 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)[67]. A protein orthology search was also conducted against the EggNOG databases[68] to assign Gene Ontology (GO), KEGG terms and protein domains from Pfam[69] and SMART[70]. Additional family and domain search was performed against TIGRFAM[71], SFLD[72], HAMAP[73], CDD[74], SUPERFAMILY[75], PRINTS[76], PANTHER[77], and Gene3D[78] using InterProScan (v 5.68–100)[79]. The results of functional annotation are summarised in Table 2. The top 10 GOs assigned to the gene models are shown in Fig. 2b without distinguishing between molecular function, biological process, and cellular component. The top 10 GOs for each category are shown in supplementary Fig. 2.

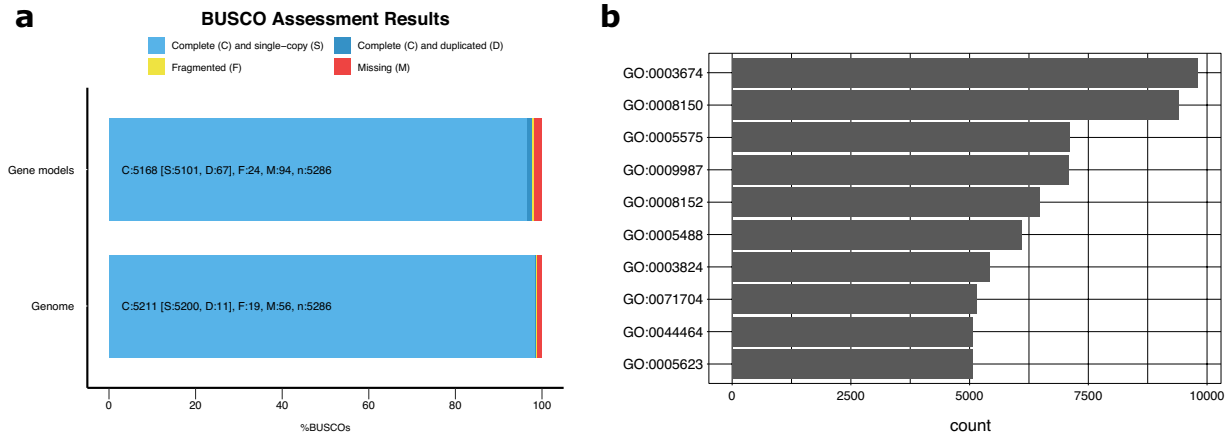

Fig. 2 BUSCO assessment and the top 10 GO assignments of transcriptome-based predicted gene models. (**a**) BUSCO scores of the gene models (top) and the genome assembly (bottom). (**b**) Overall top 10 GO assignments to gene models.

No. of protein coding gene	16,295
Average CDS length [bp]	1500.9
Average exon length [bp]	229.6
Average intron length [bp]	1705.01

Table 1. Statistical summary of the constructed gene models.

		Similarity search			Ontology search		Total	
		EggNOG	TrEMBL	Swiss-prot	EggNOG**	InterPro		
aligned	informative	12753	11925	7705	10771	12693	annotated***	14407
	uninformative*	0	2464	187	3188	—		
unaligned		3542	1906	8403	2336	3602	unannotated***	1888

Table 2. 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.

ATAC library preparation and data processing. Another batch of early embryo samples subjected to Iso-seq and sRNA-seq were subjected to ATAC-seq. Fragmentation and amplification of the ATAC-seq libraries were conducted according to Buenrostro *et al.*[80]. 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)[81]. Alignments containing mismatches were then removed using bamutils (v 0.5.9)[82]. Next, we removed duplicated reads using GATK MarkDuplicates (v 4.1.7)[83]. The resulting bam files were converted to bigwig files using deepTools bamCoverage (v 3.5.1)[84]. Heatmap was created using deepTools computeMatrix, and the starting point of the gene model was set to the reference point (Fig. 3).

Small RNA mapping. The small RNA reads were trimmed using Trim Galore (v 0.6.6)[85] in small RNA mode. The trimmed small RNA reads were mapped to the assembled transcriptome, allowing up to 3 nucleotide mismatches using HISAT2 (v 2.1.0)[64] and ngsutils (v 0.5.9)[82]. The information for each library is summarised in Supplementary Table 3.

piRNA cluster detection. The piC detection was performed as previously described[5]. proTRAC (v 2.4.4)[86] 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. 4). The two nearest gene models define the identity of piC. 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)[87] (Fig. 4a). The aggregation relationship of those piCs was visualized by ComplexUpset (v 1.3.3)[88] (Fig. 4b).

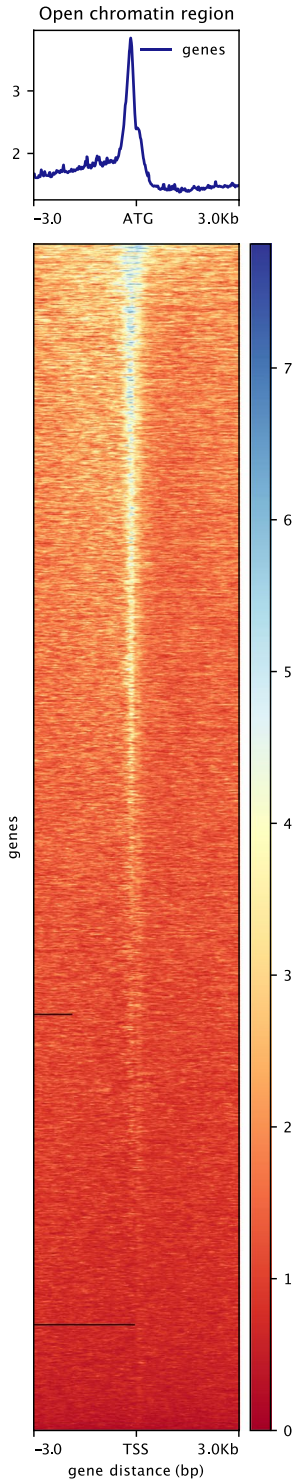

Fig. 3 Heatmap around gene bodies of ATAC-seq on early embryos.

Data Records

Newly sequenced transcriptome data for this study have been deposited in DDBJ. Iso-seq data derived from larval nerve system and pupal ovary are available under accession numbers DRS302553[89] and DRS302552[90], respectively. piRNA-targeted small RNA-seq data derived from pupal ovary and pupal testis are available under accession numbers DRR396195[91] and DRR396196[92], respectively. Embryonic ATAC-seq reads were registered under the accession number DRR515038[93]. Nanopore long read data were available under accession number DRR583866[94]. The accession numbers AP039547–AP039581[6–40] were assigned to the sequences of the nanopore-derived genome assembly. For previously released data, see Supplementary Table 3. Nanopore-derived genome assembly is also available at the FigShare repository[95]. In addition to the nanopore-derived genome

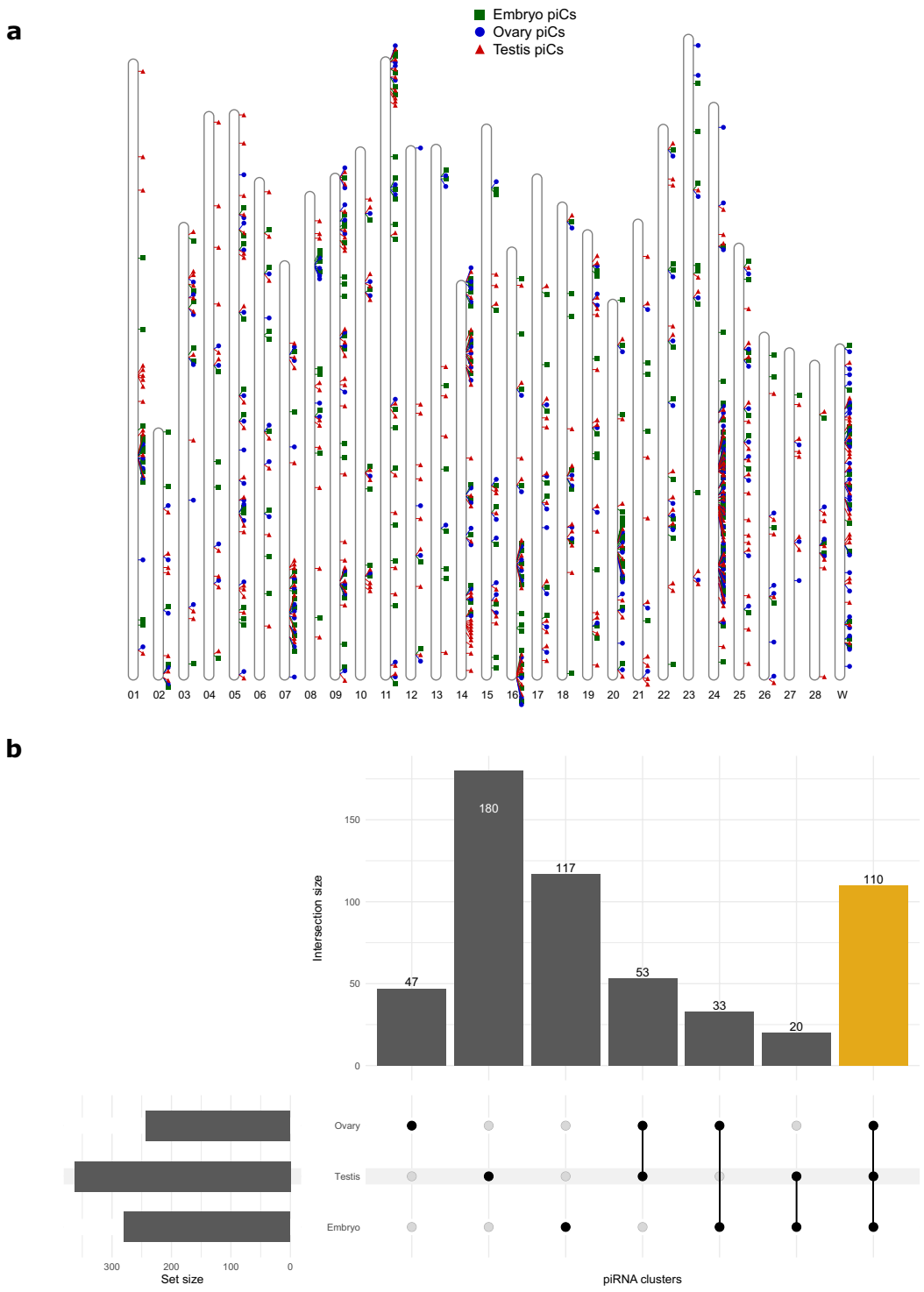

Fig. 4 piRNA clusters on p50ma genome. (**a**) piCs distribution detected in early embryos (box), pupal ovary (circle), and pupal testis (triangle). (**b**) UpSet plot visualizing piCs assigned to each tissue. The nearest two gene models defined the identity of piCs: When comparing piCs identified in different tissues, if the nearest upstream and downstream gene models are the same, those piCs were treated as the same piC.

assembly, we deposited the following files at the FigShare repository[95]: 20240716_Bmo_Tsebra_merged.filtered.renamed.cds/aa.fasta, 20240716_Bmo_Tsebra_merged.filtered.renamed.gtf/gff3, 20240716_entap_results.tsv, p50ma_Testis_piCs.gtf, p50ma_Ovary_piCs.gtf and p50ma_Early-Embryo_piCs.gtf. Two fasta files, 20240716_Bmo_Tsebra_merged.filtered.renamed.cds/aa.fasta, contain nucleotide and amino acid sequences of the predicted gene models. 20240716_Bmo_Tsebra_merged.filtered.renamed.gtf/gff3 describe the genomic coordinate information of the gene models. The result of functional annotation was summarised in 20240716_entap_results.

tsv. The coordinate information for piRNA clusters identified in three tissues was summarised in p50ma_Testis_piCs.gtf, p50ma_Ovary_piCs.gtf, and p50ma_Early-Embryo_piCs.gtf.

Technical Validation

BUSCO (v 5.4.6)[66] with lepidoptera_odb10 lineage dataset was used to assess the quality of gene models. For comparison, the results are summarized in Fig. 2, together with the results of BUSCO analysis for the genome assembly. 97.77% of the complete and single-copy BUSCO sequences were present in the gene models, while 98.58% 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. The mapping rates of RNA-seq data to genome assembly were summarized in Supplementary Table 4. The mapping rates ranged between 82.5–96.0% for all samples. The mapping rates and the above-mentioned BUSCO completeness scores ensure the RNA-seq data quality and the genome assembly quality.

Code availability

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

Received: 15 August 2024; Accepted: 19 February 2025;
Published online: 28 February 2025

References

1. Mita, K. *et al*. The genome sequence of silkworm, Bombyx mori. *DNA Res.* **11**, 27–35 (2004).
2. Xia, Q. *et al*. A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science (80-.).* **306**, 1937–1940 (2004).
3. Abe, H., Mita, K., Yasukochi, Y., Oshiki, T. & Shimada, T. Retrotransposable elements on the W chromosome of the silkworm, Bombyx mori. *Cytogenet. Genome Res.* **110**, 144–151 (2005).
4. Abe, H. *et al*. Partial deletions of the W chromosome due to reciprocal translocation in the silkworm Bombyx mori. *Insect Mol. Biol.* **14**, 339–352 (2005).
5. Abe, H., Fujii, T., Shimada, T. & Mita, K. Novel non-autonomous transposable elements on W chromosome of the silkworm, Bombyx mori. *J. Genet.* **89**, 375–387 (2010).
6. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039547 (2025).
7. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039548 (2025).
8. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039549 (2025).
9. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039550 (2025).
10. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039551 (2025).
11. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039552 (2025).
12. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039553 (2025).
13. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039554 (2025).
14. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039555 (2025).
15. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039556 (2025).
16. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039557 (2025).
17. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039558 (2025).
18. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039559 (2025).
19. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039560 (2025).
20. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039561 (2025).
21. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039562 (2025).
22. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039563 (2025).
23. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039564 (2025).
24. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039565 (2025).
25. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039566 (2025).
26. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039567 (2025).
27. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039568 (2025).
28. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039569 (2025).
29. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039570 (2025).
30. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039571 (2025).
31. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039572 (2025).
32. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039573 (2025).
33. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039574 (2025).
34. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039575 (2025).
35. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039576 (2025).
36. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039577 (2025).
37. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039578 (2025).
38. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039579 (2025).
39. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039580 (2025).
40. *NCBI nucleotide* https://identifiers.org/ncbi/nucleotide:AP039581 (2025).
41. *NCBI Genome assembly database* https://identifiers.org/insdc.gca:GCA_030269925.2 (2023).
42. 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).
43. Hart, A. J. *et al*. EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol. Ecol. Resour.* **20**, 591–604 (2020).
44. Jain, S. *et al*. TALEN outperforms Cas9 in editing heterochromatin target sites. *Nat. Commun.* **12**, 4–13 (2021).
45. Kawaoka, S. *et al*. The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *Rna* **15**, 1258–1264 (2009).
46. Nishida, K. M. *et al*. Respective functions of two distinct siwi complexes assembled during PIWI-interacting RNA biogenesis in bombyx germ cells. *Cell Rep.* **10**, 193–203 (2015).
47. Nishida, K. M. *et al*. Hierarchical roles of mitochondrial Papi and Zucchini in Bombyx germline piRNA biogenesis. *Nature* **555**, 260–264 (2018).

48. Gebert, D. *et al*. Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. *Mol. Cell* **81**, 3965–3978.e5 (2021).
49. Wick, R. *Porechop*. https://github.com/rrwick/Porechop (2017).
50. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).
51. Bionano Genomics, I. *Bionano Access Software*. https://bionano.com/access-software/ (2024).
52. Bionano Genomics, I. *Bionano Access Software*. https://bionano.com/software-downloads/ (2024).
53. *DNA Data Bank of Japan* https://ddbj.nig.ac.jp/resource/sra-submission/DRA015869.
54. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, 1–14 (2018).
55. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
56. Smit, A., Hubley, R. & Green, P. *RepeatMasker* http://www.repeatmasker.org (2021).
57. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30**, 2811–2812 (2014).
58. 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).
59. 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).
60. Yokoi, K., Tsubota, T., Jouraku, A., Sezutsu, H. & Bono, H. Reference transcriptome data in silkworm Bombyx mori. *Insects* **12** (2021).
61. 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).
62. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 1–12 (2021).
63. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
64. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
65. Brůna, T., Gabriel, L. & Hoff, K. J. Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA (2024).
66. 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).
67. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).
68. Huerta-Cepas, J. *et al*. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
69. Mistry, J. *et al*. Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
70. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* **46**, D493–D496 (2018).
71. Haft, D. H. *et al*. TIGRFAMs: A protein family resource for the functional identification of proteins. *Nucleic Acids Res.* **29**, 41–43 (2001).
72. Akiva, E. *et al*. The Structure-Function Linkage Database. *Nucleic Acids Res.* **42**, 521–530 (2014).
73. Pedruzzi, I. *et al*. HAMAP in 2015: Updates to the protein family classification and annotation system. *Nucleic Acids Res.* **43**, D1064–D1070 (2015).
74. Wang, J. *et al*. The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384–D388 (2023).
75. 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).
76. 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).
77. 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).
78. Lewis, T. E. *et al*. Gene3D: Extensive prediction of globular domains in proteins. *Nucleic Acids Res.* **46**, D435–D439 (2018).
79. Jones, P. *et al*. InterProScan 5: Genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
80. Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. ATAC-seq method. *Curr. Protoc. Mol. Biol.* **2015**, 1–10 (2016).
81. Vasimuddin, M., Misra, S., Li, H. & Aluru, S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. *2019 IEEE Int. Parallel Distrib. Process. Symp*. 314–324 (2019).
82. Breese, M. R. & Liu, Y. NGSUtils: A software suite for analyzing and manipulating next-generation sequencing datasets. *Bioinformatics* **29**, 494–496 (2013).
83. van der Auwera, G. & O'Connor, B. D. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. (O'Reilly Media, Incorporated, 2020).
84. Ramírez, F. *et al*. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
85. Krueger, F. *Trim Galore*. https://github.com/FelixKrueger/TrimGalore (2020).
86. Rosenkranz, D. & Zischler, H. proTRAC - a software for probabilistic piRNA cluster detection, visualization and analysis. *BMC Bioinformatics* **13**, 5 (2012).
87. 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).
88. Krassowski, M., Arts, M., Lagger, C. & Max *krassowski/complex-upset: v1.3.5*. https://doi.org/10.5281/zenodo.7314197 (2022).
89. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRS302553 (2023).
90. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRS302552 (2023).
91. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRR396195 (2023).
92. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRR396196 (2023).
93. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRR583866 (2023).
94. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:DRR515038 (2023).
95. Lee, J. *Bombyx mori* (strain: p50ma) genome annotation files including gene models, its functional annotation result, and piRNA cluster information. *FigShare* https://doi.org/10.6084/m9.figshare.26303761 (2024).

Acknowledgements

Insects were donated from Kyushu University and Shinshu 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 study was supported by the 2022 PacBio APAC Plant & Animal Sciences HiFi-SeqXP Grant co-hosted by NovogeneAIT to J.L. and the 2022 Gakushuin University Computer Centre Collaborative Research Program to J.L.

Author contributions

J.L. designed the research plan, performed RNA extraction, analyzed the obtained data, and wrote the manuscript. T.S. also designed this research plan and performed the data analysis. J.L. and T.A. developed strategies for genome assembly in p50ma. T.F. and K.S. assessed the accuracy of genome assembly and checked the manuscript.

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/s41597-025-04679-5.

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.