frontiers in Genetics

ORIGINAL RESEARCH
published: 16 September 2021
doi: 10.3389/fgene.2021.718266

High-Quality *de novo* Chromosome-Level Genome Assembly of a Single *Bombyx mori* With *BmNPV* Resistance by a Combination of PacBio Long-Read Sequencing, Illumina Short-Read Sequencing, and Hi-C Sequencing

Min Tang[1], Suqun He[1], Xun Gong[2,3], Peng Lü[1], Rehab H. Taha[4] and Keping Chen[1]*

[1] School of Life Sciences, Jiangsu University, Zhenjiang, China, [2] Institute of Clinical Pharmacology, Anhui Medical University, Hefei, China, [3] Department of Medical Rheumatology, Columbia University, New York, NY, United States, [4] Department of Sericulture, Plant Protection Research Institute, Agricultural Research Center, Giza, Egypt

OPEN ACCESS

Edited by:
Qunfeng Dong,
Loyola University Chicago,
United States

Reviewed by:
Claudia Vilo,
University of Antofagasta, Chile
Sanyuan Ma,
Southwest University, China

***Correspondence:**
Keping Chen
kpchen@ujs.edu.cn

Specialty section:
This article was submitted to
Genomic Assay Technology,
a section of the journal
Frontiers in Genetics

Received: 14 June 2021
Accepted: 05 August 2021
Published: 16 September 2021

Citation:
Tang M, He S, Gong X, Lü P, Taha RH and Chen K (2021) High-Quality de novo Chromosome-Level Genome Assembly of a Single Bombyx mori With BmNPV Resistance by a Combination of PacBio Long-Read Sequencing, Illumina Short-Read Sequencing, and Hi-C Sequencing. Front. Genet. 12:718266. doi: 10.3389/fgene.2021.718266

The reference genomes of *Bombyx mori* (*B. mori*), Silkworm Knowledge-based database (SilkDB) and SilkBase, have served as the gold standard for nearly two decades. Their use has fundamentally shaped model organisms and accelerated relevant studies on lepidoptera. However, the current reference genomes of *B. mori* do not accurately represent the full set of genes for any single strain. As new genome-wide sequencing technologies have emerged and the cost of high-throughput sequencing technology has fallen, it is now possible for standard laboratories to perform full-genome assembly for specific strains. Here we present a high-quality *de novo* chromosome-level genome assembly of a single *B. mori* with nuclear polyhedrosis virus (*BmNPV*) resistance through the integration of PacBio long-read sequencing, Illumina short-read sequencing, and Hi-C sequencing. In addition, regular bioinformatics analyses, such as gene family, phylogenetic, and divergence analyses, were performed. The sample was from our unique *B. mori* species (NB), which has strong inborn resistance to *BmNPV*. Our genome assembly showed good collinearity with SilkDB and SilkBase and particular regions. To the best of our knowledge, this is the first genome assembly with *BmNPV* resistance, which should be a more accurate insect model for resistance studies.

Keywords: ***Bombyx mori***, **PacBio sequencing, genome assembly, illumine sequencing, Hi-C technology**

INTRODUCTION

Bombyx mori (*B. mori*) or domestic silkworm is a well-known bioreactor that produces natural silk in sericulture and is frequently used as a lepidopteran model to study insect immunology and disease resistance (Maekawa et al., 1988; Arunkumar et al., 2006). In recent decades, the genetics and genomics of *B. mori* in several Asian countries have been greatly elucidated (Xia et al., 2004).

Its natural primal enemy, *B. mori* nucleopolyhedrovirus (*BmNPV*), can cause widespread death to the overwhelming majority of *B. mori* strains. Several strains bred in different laboratories display congenital *BmNPV* resistance (Xu et al., 2013). Similar to most higher organisms, the key prerequisite of most biological research is the assembly of whole-genome sequences. For *B. mori*, the first genome sequence with 3× coverage generated by shotgun sequencing technology (fosmid-end) was published in Japan in 2004, and the *B. mori* strain used was *p50T* (Koike et al., 2003; Mita et al., 2004). At almost the same time, an independent Chinese group also performed a similar whole-genome sequencing (BAC-end) with 5.9× coverage, and the *B. mori* strain used was *Dazao* (Xia et al., 2004). Both studies declared that the estimated genome coverage was more than 90%. Based on the EST database in the former project, the first version of SilkBase was released in 2006. Then, the integrated transcriptomic and genomic data were added to the project and released in the second version of SilkBase in 2015. Two years later, the *p50T* strain was sequenced again by making use of the combination of PacBio and Illumina sequencing technologies, and this genome assembly with increased accuracy was released as SilkBase v2.1 (Kawamoto et al., 2019). Furthermore, the Silkworm Knowledge-based database (SilkDB) was developed for data storage, retrieval, analysis, and visualization by Chinese groups in 2005 (Wang et al., 2005), and the assembly accuracy was later promoted by the International Silkworm Genome Consortium in 2010 (SilkDB 2.0) (International Silkworm Genome, 2008). Ten years later, cutting-edge PacBio and Hi-C biotechnologies were employed to assemble the latest version of the *Dazao* genome in 2020 (SilkDB 3.0) (Duan et al., 2010; Lu et al., 2020). To the best of our knowledge, SilkDB 3.0 and SilkBase v2.1 are the most commonly used database reference genomes. Hence, they are recognized as the basis for a number of *B. mori* studies and databases (Cao et al., 2017; Li et al., 2019; Zhu et al., 2019; Fujimoto et al., 2020).

It is a pity that both of the strains they used do not have *BmNPV*-resistance. Perhaps, as a consequence, their genome information is not perfectly accurate for the resistance studies considering the importance of the reference genome. Since countless efforts have been made for the *BmNPV* resistance without any acknowledged results, we think that it is time to optimize the reference genome group—for example, establish a *BmNPV*-resistant genome. In addition, in our opinion, the pipelines for constructing the above-mentioned two genome references could be more accurate and practical. For SilkDB 3.0, short-read sequencing data that could correct the draft genome assembly were not added to aid the assembly process (Jayakumar and Sakakibara, 2019). For SilkBase, Hi-C data should be used to help arrange and orient contigs or scaffolds (Belton et al., 2012). More importantly, neither *Dazao* nor *p50T* has resistance to *BmNPV*, which may increase genomic noise or even introduce errors in the resistance study of *B. mori*. Hence, here we introduce a high-quality *de novo* chromosome-level genome assembly of *B. mori* with *BmNPV* resistance by integrating data from PacBio long-read sequencing, Illumina short-read sequencing, and Hi-C sequencing technology.

MATERIALS AND METHODS

Sample Collection and DNA Isolation

Two five-instar *B. mori* belonging to the *NB* strain were collected from the School of Life Sciences, Jiangsu University, and downstream wet-laboratory experiments were performed at Novogene Co., Ltd. First, one of the imagoes was sufficiently ground with polyvinyl polypyrrolidone powder, and whole DNA was extracted for sequencing using sodium dodecyl sulfate (SDS) and cetyltrimethylammonium bromide (Chen et al., 2010; Kasajima, 2018). Then, 1% agarose gel was used to examine DNA degradation and contamination. Consequently, the DNA purity was checked using a NanoPhotometer® spectrophotometer (IMPLEN, CA, United States), and the DNA concentration was measured using a Qubit® DNA Assay Kit on a Qubit® 2.0 fluorometer (Life Technologies, CA, United States).

Library Construction and Illumina Sequencing

A total of 1.5 μg extracted DNA was used as input material for the DNA sample preparations and sheared with a Covaris Focused ultrasonicator. Following the recommendations of the manufacturer, a library for Next-Gen sequencing was constructed using a Truseq® Nano DNA Sample Preparation Kit (Illumina, San Diego, CA, United States) after end-repair and the addition of index code. Subsequently, the fragments were end-polished, A-tailed, and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. After fragment selection for a size of 350 bp and PCR amplification, the products were purified with the AMPure XP system (PacBio, Menlo Park, CA, United States), and the library was analyzed for size distribution by an Agilent 2100 Bioanalyzer and quantified using real-time PCR. Finally, the library constructed as detailed above was sequenced by an Illumina HiSeq X Ten instrument, and 150 bp paired-end (PE) reads, including ∼350 bp inserts, were generated. After removing the reads with adapters, the paired reads that had more than 10% N bases or over 20% low-quality bases on either end sequence were discarded. As a result, the remaining reads passed onto the downstream analysis as clean reads.

Library Construction and PacBio Sequencing

More than 5 μg of concentrated genomic DNA (gDNA) that passed the inspections was used for the size-select DNA fragment step. g-TUBE was utilized to shear the gDNA into ∼20-kb fragments to construct a 20-kb SMRTbell library. After shearing, AMPure PB beads were used to concentrate the sheared gDNA. Then, the fragments were treated to remove the single-strand overhangs and repair DNA damage with ExoVII enzyme. Subsequently, the ends of the double-strand fragments were polished with T4 DNA Polymerase and then ligated to single-molecule real-time (SMRT) hairpin adapters. Through EXOIII and VII enzyme treatment, the SMRTbell templates were washed out. The library was again concentrated and purified with AMPure PB beads. In this case, we employed the BluePippin system (Sage Science, Inc.) and set a 20-kb cutoff threshold for size selection. The sublibrary was also concentrated and purified with AMPure PB beads. After annealing the sequencing primer to the SMRTbell template, polymerase was bound to both ends of the SMRTbell templates using a binding kit for efficient loading into ZWMs. Finally, a total of 14 SMRT cells were run on the PacBio Sequel system with the P6-C4 sequencing reagent (Kingan et al., 2019). Briefly, a SMRTbell library was constructed with the SMRTbell Express Template Prep Kit 2.0 according to the latest protocol published by Pacific Biosciences of California, Inc.

Hi-C Experiment and Sequencing

The Hi-C experiment was performed as described previously (Belton et al., 2012; Yardimci et al., 2019). Cell suspensions were made with the other *B. mori* imago and treated with paraformaldehyde to fix the three-dimensional structures of the nuclei to retain the relationship between genomic and physical distance. Therefore, after cell lysis with RIPA lysis buffer (1 M Tris–HCl, pH 8, 1 M NaCl, 10% CA-630, and 13 units of protease inhibitor), the exposed chromatin was cross-linked *in situ* to trap sequence interactions across the entire genome and between different chromosomes. Then, the supernatant was centrifuged at 5,000 rpm at room temperature for 10 min. The obtained pellet was washed twice with 100 μl ice cold 1× NEB buffer, followed by centrifugation at 5,000 rpm for 6 min.

The chromatin was resuspended in 100 μl NEB buffer and solubilized with dilute SDS, followed by incubation at 65°C for 10 min. After quenching SDS with Triton X-100, overnight digestion was performed with the four-cutter endonuclease *Mbo*I at 37°C on a rocking platform. Subsequently, the fragmented chromatin was digested with the restriction enzymes *Hind*III and *Dpn*II at 37°C for 16 h. Next, for proximity ligation, the fragmented loci were ligated and marked with biotin to create chimeric junctions between adjacent sequences after incubation at 37°C for 45 min. The enzymatic system was inactivated with 20% SDS solution. Finally, proteinase K was added for reverse cross-linking overnight at 65°C, and the resulting sample was purified and dissolved in 90 μl of double-distilled water. The purified fragments were sheared to a size of ∼350 bp, and then their ends were repaired. The biotin-containing fragments were isolated with Dynabeads^{TM} M-280 Streptavidin purchased from Thermo Fisher Scientific Inc. After adding A-tails to the fragment ends and following ligation by Illumina PE sequencing adapters, the Hi-C library was amplified by 12–14 PCR cycles and sequenced on the Illumina NovaSeq 6000 platform with the PE set to 150 bp.

RNAseq Experiment and Transcriptome Sequencing

Transcriptome sequencing was performed as described by Conesa et al. (2016). The RNA pool was made from a whole one-instar *B. mori* using the RNeasy Maxi Kit purchased from QIAGEN, and raw sequencing data were obtained on the Illumina HiSeq X sequence platform. The subsequent quality control steps were performed with FastQC, FASTX-Toolkit, and Trimmomatic (Bolger et al., 2014).

RESULTS

De novo Assembly of the *B. mori* Genome and Assessment

The raw Illumina sequencing data first underwent quality control after image recognition, contamination, and adapter elimination. Then, the initial characterization of the *B. mori* genome was estimated through *k*-mer (*k* = 17 in this study) analysis of the clean data by jellyfish and GenomeScope (Vurture et al., 2017). As a result, the estimated genome size (475.39 and 464.90 Mbp after later correction), heterozygosity (0.23%), and repeat content (43.78.09%) were basically consistent with previous findings (Mita et al., 2004; Xia et al., 2004; International Silkworm Genome, 2008; Kawamoto et al., 2019). All the sequencing data produced in this study are listed in **Table 1**. As the table shows, long-read sequencing on the PacBio Sequel system yielded 50.08 G of data with a high coverage of 107.72× short-read sequencing on the Illumina NovaSeq 6000 system yielded 64.81 G of data with an average coverage of 139.41× and Hi-C sequencing on the HiSeq X system yielded 83.78 G of data with an average coverage of 180.21×.

During the assembly process, the long reads downloaded from the PacBio Sequel platform were subjected to self-correction by the NextCorrect module of NextDenovo.[1] Taking the corrected reads as input materials, FALCON was employed to assemble the draft framework of the *B. mori* genome (Chin et al., 2016). To correct the high error rate of the long-read sequencing technology (Rang et al., 2018), two genome sequence polishing steps were performed: the Quiver algorithm was first used to polish the genome using PacBio long reads (Chin et al., 2013), and another round of genome-wide base-level correction of the Illumina clean reads was performed using Pilon (Walker et al., 2014). For chromosome-level scaffolding (Oluwadare et al., 2019), the clean Hi-C reads were mapped to the assembled genome using BWA (Li and Durbin, 2010), and the repeats and unmapped reads were removed by SAMtools (Li et al., 2009) to obtain high-quality reads and information about the restriction sites for enzyme cutting. As a result, only uniquely mapped read pairs were considered for subsequent analysis, and the Hi-C heatmap did not show misassembly during scaffolding (**Supplementary Figure 1**). LACHESIS was finally used to cluster, order, and orient the assembled contigs (Burton et al., 2013). The assembly statistics of the completed genome are shown in **Table 2** and **Supplementary Tables 1–3**. Only scaffolds larger than 100 bp were selected to perform the assembly; the contig N50 was 3.75 Mbp, and the scaffold N50 was 17.26 Mbp.

Briefly, we assembled a high-quality chromosome-scale genome for *B. mori de novo* by making use of the 198.67 G of sequencing data (**Figure 1**). The overall length and N50 value of the contigs were 455.46 and 3.75 Mbp, respectively, while the overall length and N50 value of the scaffolds were 455.46 and 17.26 Mbp. To evaluate the quality of the newly assembled genome, its completeness was first assessed by the core eukaryotic genes mapping approach (CEGMA), whose core gene

[1] https://github.com/Nextomics/NextDenovo

TABLE 1 | Statistics of the *Bombyx mori* sequencing data for genome assembly.

Types	Approach	Sequencing platform	Insert size (bp)	Total data (G)	Read length (bp)	Sequence coverage (×)[a]
Genome	Illumina	Novaseq 6000	350	64.81	150	139.41
Genome	Pacbio	PacBio Sequel	–	50.08	∼20k	107.72
Genome	Hi-C	HiSeq X	–	83.78	150	180.21
Transcriptome Total	Illumina	HiSeq X	350	198.67	–	427.34

[a] *The coverage was calculated according to an estimated genome size of 464.90 Mbp.*

TABLE 2 | Statistics of the *B. mori* genome assembly.

Sample ID	Length		Number	
	Contig[a] (bp)	Scaffold (bp)	Contig[a]	Scaffold
Total	455,458,164	455,459,137	241	70
Max	12,311,683	21,717,039	–	–
Number ≥2,000	–	–	241	70
N50	3,751,516	17,263,692	42	12
N60	2,941,979	16,129,549	57	15
N70	2,487,328	15,544,771	73	18
N80	1,798,075	14,680,113	95	21
N90	1,077,501	12,534,506	126	24

[a] *Contig after scaffolding.*

database includes 248 genes derived from six model eukaryotes (Parra et al., 2007). **Supplementary Table 4** shows that 223 core eukaryotic genes were retrieved, which was up to 89.92% of the total number of genes. In addition to CEGMA, Benchmarking Universal Single-Copy Ortholog (BUSCO) software, invoking TBLASTN (Gertz et al., 2006), AUGUSTUS (Stanke and Morgenstern, 2005), and HMMER (Wheeler and Eddy, 2013), was employed (Seppey et al., 2019). Of the 978 orthologous genes, 98.1% complete BUSCOs (C + D) were detected in its report (see **Supplementary Tables 4**, **5**). CEGMA and BUSCO revealed that we obtained a high-quality and complete *B. mori* genome.

Moreover, the reference genomes from SilkDB and SilkBase were used to check the collinearity with our genome assembly with JCVI (Tang et al., 2008). As expected, great collinearity existed among the three chromosome-level genomes (**Figure 2**). The distortions on several chromosomes, such as Chr1, Chr10, and Chr23, were attributable to scaffold perversions, which is quite reasonable, as large chromosomal structure changes may occur among closely related strains. In addition, optical mapping methods and approaches for the ordering and orientation of scaffolds can also cause a systematic deviation (Ekblom and Wolf, 2014; Belser et al., 2018; Deschamps et al., 2018). Nevertheless, crevices with different sizes indicate small INDELs that may be the origins of BmNPV resistance; this will be analyzed in detail in future studies in our laboratory.

Genome Annotation

Supplementary Figure 2 shows three routes in the road map of the genome annotation phase. From the left, an integration of homology-based and *de novo* approaches was used to identify the repeat sequences. First, LTR_FINDER[2] (Xu and Wang, 2007), RepeatScout,[3] and RepeatModeler[4] were employed to create a *de novo* repeat sequence dataset. Then, based on the RepBase database, RepeatMasker (see text footnote 3) and RepeatProteinMask[5] were used to recognize the sequences that had a high similarity in RepBase. Moreover, Tandem Repeats Finder (Benson, 1999) and RepeatModeler2 (Flynn et al., 2020) were also utilized to detect tandem repeats and transposable element (TE) families in the assembled genome with default settings. Consequently, all the above-mentioned repetitive elements were combined to annotate the genome assembly using RepeatMasker (**Table 3** and **Supplementary Figure 3**; **Supplementary Table 6**).

For the prediction of gene structures, *de novo*, homology-based, and transcriptome-based strategies were integrated to predict genes in the *B. mori* genome (**Table 4** and **Supplementary Figure 2**). *De novo* prediction was performed with AUGUSTUS (Stanke and Morgenstern, 2005), GlimmerHMM (Majoros et al., 2004), SNAP (Johnson et al., 2008), Geneid,[6] and Genscan (Burge and Karlin, 1997), while homology-based prediction was conducted with BLAT, which aligns the protein sequences of *Ame*, *Pra*, *Bmo*, *Pxy*, *Tca*, and *Dme* to the above-mentioned genome assembly (Kent, 2002). Then, the transcript reads were aligned to the genome assembly using the TopHat package (Kim et al., 2013), followed by gene structure prediction with Cufflinks (Ghosh and Chan, 2016). Finally, all gene models were merged by EVidenceModeler (EVM)[7] (Haas et al., 2008) after removing redundancy. To enrich the contents, the RNAseq sequence data were assembled *de novo* with Trinity (Grabherr et al., 2011; Haas et al., 2013) and applied to insert untranslated regions and alternative splicing information to the EVM dataset by PASA (Haas et al., 2003). In conclusion, 9,113 gene structures were synchronously supported by *de novo*, homology-based, and transcriptome-based prediction strategies (**Supplementary Figure 5**). The same prediction pipeline was also applied to several proximal species (*Bmj*, *Ame*, *Bmo*, *Dme*, *Pra*, *Pxy*, and *Tca*; see **Supplementary Table 10**). Notably, the number of predicted genes for *Bmo was* 13,850, which was very close to the 13,103 predicted genes for *Bmj*.

[2] http://tlife.fudan.edu.cn/tlife/ltr_finder
[3] http://www.repeatmasker.org
[4] http://www.repeatmasker.org/RepeatModeler.html
[5] https://github.com/rmhubley/RepeatMasker/blob/master/RepeatProteinMask
[6] https://genome.crg.es/geneid.html
[7] http://evidencemodeler.sourceforge.net/

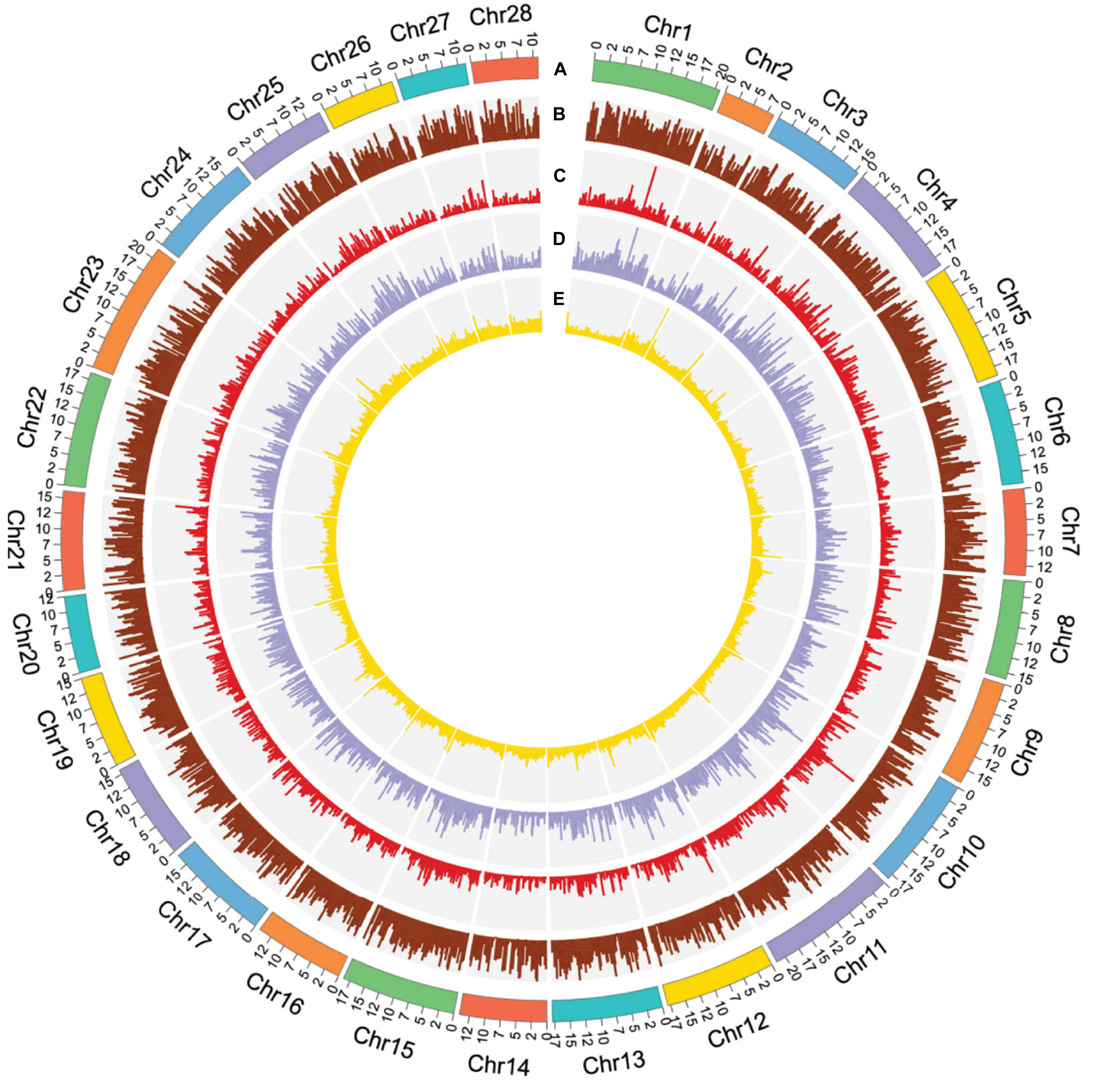

FIGURE 1 | Summary of gene distribution and genetic diversity across the 28 chromosomes. **(A)** Chromosome position. **(B)** Gene density in 200-kb windows. **(C)** CDS density in 200-kb windows. **(D)** Exon density in 200-kb windows. **(E)** GC content in 200-kb windows.

For gene function annotation, the final protein sequences obtained above were used as inputs for the SwissProt,[8] Nr,[9] Kyoto Encyclopedia of Genes and Genomes (KEGG),[10] InterPro,[11] Gene Ontology,[12] and Pfam[13] databases. After alignment to these published protein databases, 99.6% of the genes were annotated with functional entities (**Supplementary Table 8**). In addition, non-coding RNAs were also annotated (**Supplementary Table 9**). tRNAs were detected with tRNAscan-SE[14] (Lowe and Chan, 2016; Chan and Lowe, 2019), while miRNAs and snRNAs were detected with the INFERNAL module of Rfam[15] (Kalvari et al., 2018, 2021).

Genomic Comparison

Genomic comparison in bioinformatics analysis, including gene family clustering and phylogenetic and divergence analysis, was performed by comparing our in-house *B. mori* genome assembly with the assemblies of a set of 17 other representative species.

[8] https://www.uniprot.org/

[9] https://www.ncbi.nlm.nih.gov/refseq/about/non-redundantproteins/

[10] https://www.genome.jp/kegg/

[11] https://www.ebi.ac.uk/interpro/

[12] http://geneontology.org/docs/go-annotations/

[13] 13 http://pfam.xfam.org/

[14] http://lowelab.ucsc.edu/tRNAscan-SE/

[15] http://infernal.janelia.org/

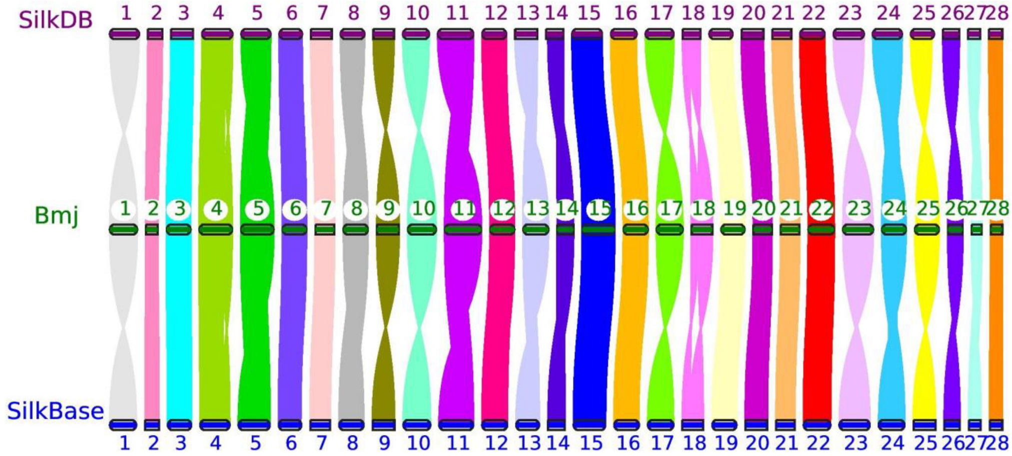

FIGURE 2 | Genomic collinearity of three chromosome-level genomes—*Bmj*, Silkworm Knowledge-based database (SilkDB), and SilkBase—by JCVI. The number in the circle represents the chromosome identity for each strain.

TABLE 3 | Statistics for the classification of repetitive elements.

	De novo + Repbase length (bp)	% in genome	TE protein length (bp)	% in genome	Combined TE length (bp)	% in genome
DNA	43,928,578	9.64	8,097,056	1.78	48,033,677	10.55
LINE	151,419,612	33.25	40,869,238	8.97	162,254,924	35.62
SINE	36,753,778	8.07	0	0	36,753,778	8.07
LTR	46,826,667	10.28	8,338,880	1.83	47,903,023	10.52
Other	0	0	0	0	0	0
Unknown	2,722,614	0.6	0	0	2,722,614	0.6
Total	253,233,852	55.6	57,220,911	12.56	257,512,897	56.54

De novo + Repbase is a transposable element (TE) pool generated by the integration of RepBase and the combination of RepeatModeler, RepeatScout, and LTR_FINDER based on the 80-80-80 principle of Uclust. TE proteins are also a TE pool derived from the application of RepeatProteinMask to the RepBase protein database. Combined TEs are the integration of De novo + Repbase and TE proteins after removing overlaps. Unknown refers to repetitive elements that cannot be classified by RepeatMasker.

TABLE 4 | Statistics of gene structure annotation.

	Gene set	Number	Average transcript length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
De novo	Augustus	11,666	9,995.37	1,376.92	6.04	227.91	1,709.45
	GlimmerHMM	48,042	8,378.78	538.88	3.37	160.04	3,312.04
	SNAP	22,743	11,183.17	645.71	4.06	158.86	3,438.46
	Geneid	9,281	33,434.53	779.78	6.32	123.39	6,138.85
	Genscan	13,002	23,083.74	1,343.72	6.02	223.36	4,334.14
Homolog	*Ame*	9,438	6,629.92	1,047.42	4.46	234.63	1,611.49
	Bmo	32,454	4,125.35	908.63	3.04	298.64	1,574.88
	Dme	6,852	7,308.34	1,179.68	4.99	236.56	1,537.21
	Pra	27,430	4,470.74	1,002.10	3.23	310.71	1,558.78
	Pxy	20,272	4,816.18	1,392.41	3.28	424.03	1,499.16
	Tca	20,637	3,647.42	1,130.13	2.7	418.63	1,481.13
RNAseq	PASA	36,779	16,791.09	1,516.73	7.27	208.72	2,437.34
	Cufflinks	35,742	26,014.39	4,582.43	8.24	556.26	2,961.05
	EVM	14,449	12,572.50	1,297.05	5.9	219.83	2,301.05
	Pasa-update[a]	13,773	15,181.10	1,452.26	6.59	220.32	2,455.20
	Final set[b]	13,103	14,748.03	1,444.25	6.53	221.28	2,407.13

[a] *Includes UTRs.*

[b] *Includes UTRs; generated from the longest transcript after alternative splicing and redundant single-exon elimination by PASA2 update.*

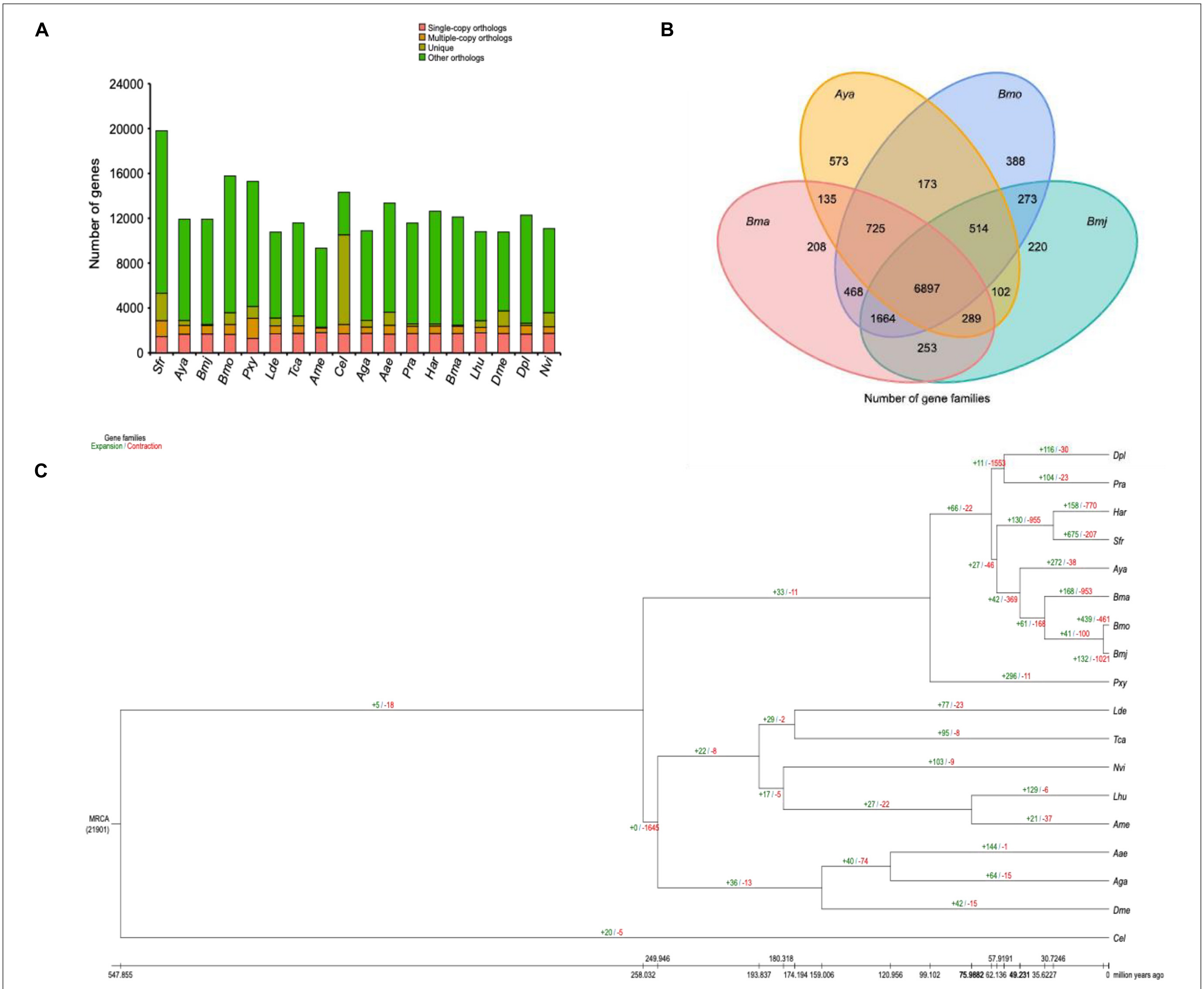

FIGURE 3 | Gene family and phylogenetic and divergence analyses of *Bombyx mori* and other representative insect genomes. **(A)** Clusters of orthologous and paralogous gene families in *B. mori* and 17 other sequenced insect genomes. The *y*-axis indicates the number of gene families, while the *x*-axis indicates the 18 species. The total number of gene families was 21,951, of which the number of single-copy orthologs was 620. **(B)** Venn diagram representing the distribution of shared gene families among *B. mori* and three other silkworms (*Bma*, *Aya*, *Bmo*, and *Bmj*). **(C)** Gene family expansions and contractions estimated by CAFÉ v3.0 and CAFE_fig. The green and red numbers indicate expanded and contracted gene families, respectively.

In addition to three closely related species (*Bmo*, *Aya*, and *Bma*) and five other lepidopteran species (*Sfr*, *Pxy*, *Pra*, *Har*, and *Dpl*), three dipteran species (*Aya*, *Aae*, and *Dme*), three hymenopteran species (*Ame*, *Lhu*, and *Nvi*), and two coleopteran species (*Lde* and *Tca*) were also included in the comparison list. Moreover, *Caenorhabditis elegans*, belonging to Rhabditia, was also included. For convenience, their abbreviations are summarized in **Supplementary Table 10**.

First, OrthoMCL[16] was used for gene family clustering analysis (Li et al., 2003; Chen et al., 2006) using the following parameters: mode, 3; inflation, 1.5 blast_file, gg_file. As a result, a total of 21,904 gene families were identified, and 650 strict single-copy orthologs were recovered in the 18 genomes (**Figure 3A**). In addition, a core set of 6,897 gene families was shared by *Bmj* (for *shipshape*, use *Bmj* instead of *B. mori*), *Bmo*, *Bma*, and *Aya* (**Figure 3B**), and only 220 gene families uniquely belonged to *Bmj*. A consequent large-scale analysis found 84 gene families specific to *Bmj* but not the 17 other selected species. Most of these specific genes were involved in receptor activity, transmembrane signaling receptor activity, and glycosphingolipid biosynthesis (**Supplementary Figure 7**).

Second, CAFÉ v3.0[17] and CAFE_fig[18] were used to analyze changes in the gene family size using the following

[16] http://orthomcl.org/orthomcl/

[17] https://hahnlab.github.io/CAFE/index.html

[18] https://github.com/LKremer/CAFE_fig

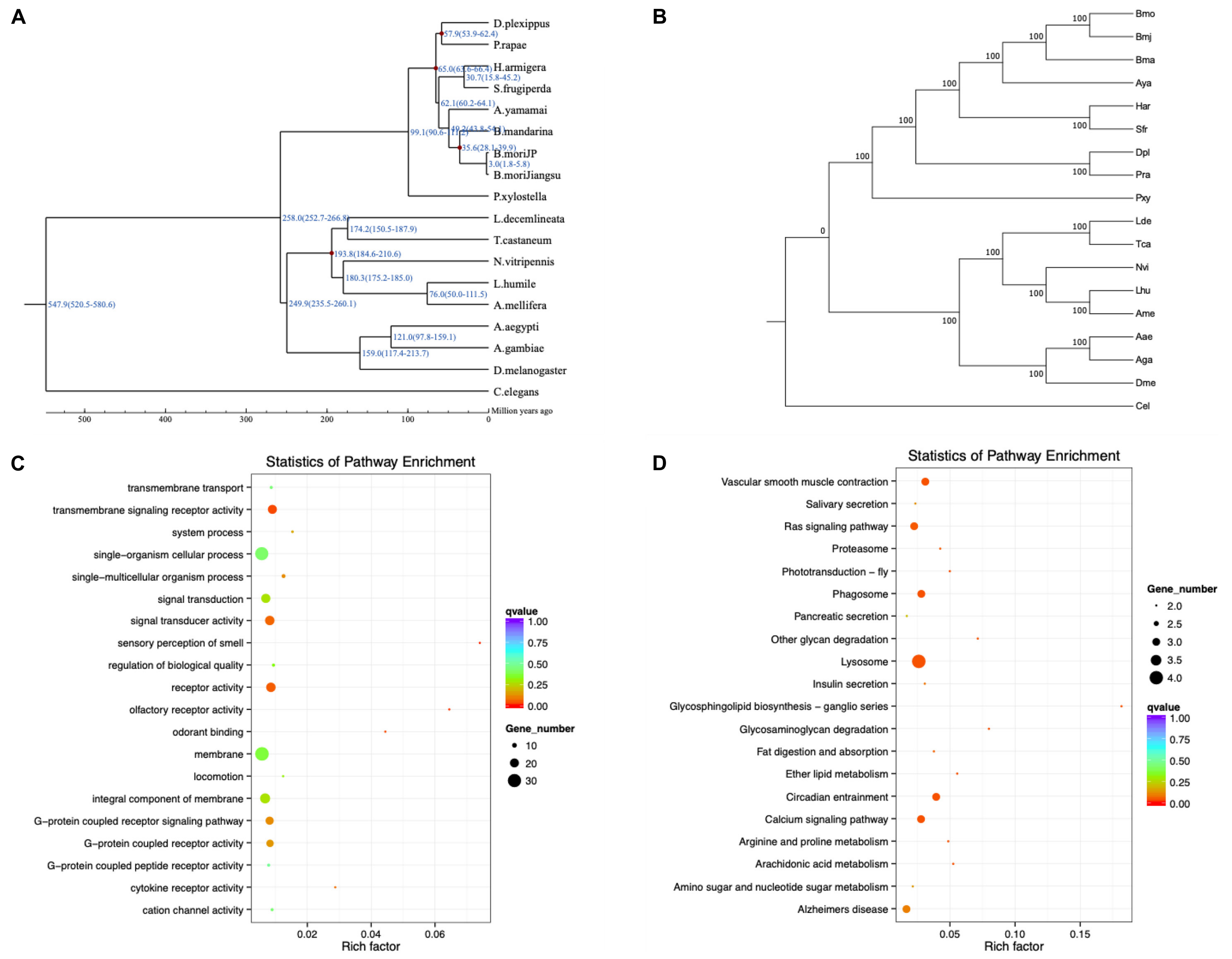

FIGURE 4 | Evolutionary analysis of *B. mori* by Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis. **(A)** Divergence time estimation of *B. mori* and closely related species. The divergence time is shown at the branches of the phylogenetic tree, along with the confidence intervals in parentheses. **(B)** Phylogenetic tree of *Bmj* species. **(C,D)** GO and KEGG pathway enrichment analysis of positively selected genes related to *BmNPV* resistance.

parameters (-i -t 8 -l -p 0.05). As **Figure 3C** shows, 132 gene families were expanded, while 1,021 gene families were contracted from the *Bmj* genome. In comparison with the closely interlinked *Bmo*, which gained 439 gene families and missed 461 gene families, *Bmj* lost many more gene families. These missed genes were rich in diverse functions, such as catalytic activity, fatty acid biosynthesis, fatty acid metabolism, and AMPK signaling pathway (**Supplementary Figure 8**). Corresponding to the phylogenetic analysis performed with the mcmctree program of the PAML package[19] (parameters: rootage, 1,000; clock, 3; alpha, 0.603820) (Yang, 1997), **Figure 3A** shows that *Bmj* has a very close relationship to *Bmo* and phylogenetically diverged from the common ancestor ∼3 million years ago. However, the history of human breeding is up to 5,000 years long (Tabunoki et al., 2016; Meng et al., 2017), and the best explanation is manual intervention in sericulture, which increases their divergence. By using the phylogenetic analysis as a supplement, the above-mentioned 650 single-copy gene families were subjected to internal alignment with MUSCLE,[20] and then the aligned results were merged to form a super alignment matrix (Edgar, 2004a,b). After that, the ML TREE algorithm in RAxML[21] was used with default parameters to create a phylogenetic tree (**Figure 3C**; Stamatakis, 2014).

To detect positively selected genes related to different phenotypes of *Bmj*, a positive selection analysis was executed in three diverse combinations with the other 17 species. MUSLCE was again employed to align the protein sequences

[19] http://abacus.gene.ucl.ac.uk/software/paml.html

[20] http://www.drive5.com/muscle/

[21] http://sco.h-its.org/exelixis/web/software/raxml/index.html

of the single-copy gene families between foreground branches and background branches. Then, the aligned results were cleaned by removing the low-quality regions using Gblocks and then reversed to CDS[22] (Castresana, 2000). For each gene family, the branch-site model in the codeml program of the PAML package was applied to check whether it was positively selected in the *Bmj* branch (Yang, 1997). Instead of simply seeking genes with ka/ks > 1, PAML makes use of two hypothesized likelihood ratios to double-check the positive selection (Yang and Nielsen, 2002; Zhang et al., 2005) (**Figures 4A,B**). For the eating pattern of mulberry (Ge et al., 2018a,b), the foreground branches were *Bmj*, *Bmo*, and *Bma*, and the background branches were *Aya*, *Har*, *Sfr*, *Pra*, *Dpl*, and *Pxy* in the first group. Sixty-two targeted genes that were involved in metabolic processes, catalytic activity, and Huntington's disease were obtained (**Supplementary Figure 9**). For the capability of silk production, the foreground branches were *Bmj*, *Bmo*, *Bma*, and *Aya*, and the background branches were *Lhu*, *Ame*, *Nvi*, *Lde*, *Tca*, *Aae*, *Aga*, and *Dme* in the second group. A total of 119 targeted genes were enriched in metal ion binding, zinc ion binding, and transmembrane transport (**Supplementary Figure 10**). In the third group, for resistance to *BmNPV*, the foreground branch was *Bmj*, while the background branches were *Bmo*, *Har*, *Sfr*, *Pra*, *Dpl*, *Pxy*, *Bma*, and *Aya*. The downstream analysis indicated that the *BmNPV* resistance of our NB species may be related to pathways of transmembrane signaling receptor activity and lysosome and signal transducer activity (**Figures 4C,D** and **Supplementary Figure 11**).

CONCLUSION AND DISCUSSION

Since approximately 5,000 years ago, *B. mori* has been one of the most important agricultural economic insects for silk production in Asia (Xiang et al., 2018). As it is completely domesticated in sericulture, its survival and reproduction totally depend on human beings (Jiang and Xia, 2014). With advances of biotechnology, *B. mori* has been treated as an important bioreactor for producing recombinant protein (Tomita et al., 2000, 2003). Moreover, because of their intermediate genome size, short life cycle, affordability, and ability to be easily used for drug screening, *B. mori* is a perfect model organism for scientific discovery, especially for lepidopterans, and is useful for economic research (Kaito et al., 2002; International Silkworm Genome, 2008; Tabunoki et al., 2016; Meng et al., 2017). To maintain the consistency of biological background among different laboratories, the *B. mori* strains of *Dazao* and *p50T* were first separately sequenced and annotated in China and Japan by whole-genome shotgun sequencing in 2004, and the genome sequence has been updated every few years, which is considered the gold standard (Mita et al., 2004; Xia et al., 2004; Meng et al., 2017; Kawamoto et al., 2019).

According to a search of NCBI PubMed with the keywords "(*Bombyx mori*) OR (silkworm)," 7,024 relevant works have been published since 2004. Among these, transcriptomics studies have inevitably made use of the above-mentioned reference genomes to identify transcripts and quantify gene expression. To obtain the brain transcriptome profiles of *BmNPV*-infected and non-infected silkworm larvae, Wang et al. (2015) mapped their clean reads to the SilkDB reference genome. Moreover, Li et al. (2016) comprehensively investigated the transcriptomic changes between susceptible and resistant *B. mori* strains after *BmNPV* infection. Three years later, they (Li et al., 2019) updated the analysis of the different alternative splicing events based on the same reference genome. Recently, Sun Q. et al. (2020) also downloaded genome sequences and annotation files from the SilkDB website for the transcriptome analysis of the immune response of *B. mori* after bidensovirus infection at the early stage. Based on the SilkBase genome sequence, Shoji et al. (2013) characterized a novel chromodomain-containing gene and later (Shoji et al., 2014) identified a number of genes whose expression can be enhanced by heterochromatin protein 1. Therefore, it follows that the two reference genomes with non-resistance to *BmNPV* are being widely used as the *B. mori* "Bible." However, with the increasing demand for precise reference genomes, a personalized edition is needed for *B. mori* research. With our more refined reference genome, one can find precise bioinformatics information such as SNPs and INDELs related to *BmNPV* resistance.

Thanks to the rapid advances in sequencing technology and bioinformatic tools in the past decade, transcriptome sequencing (∼$130/sample) has become a cost-effective technology for obtaining biological information. However, the emerging third-generation sequencing technologies, such as Nanopore/PacBio long-read sequencing and Hi-C sequencing, require a larger budget than that of standard laboratories (Makałowski and Shabardina, 2019; Wang et al., 2019). Even the use of whole-genome sequencing on the popular Illumina platform to generate short reads for draft assembly correction is expensive, let alone optical genome mapping (i.e., Bionano maps), which can correct chimeric contigs and scaffolding errors caused by Hi-C sequencing (Chan et al., 2018; Bocklandt et al., 2019; Choi et al., 2020; Ning et al., 2020; Sun L. et al., 2020). In addition, the entire pipeline of genome assembly is a tedious and time-consuming process, as ideal computing modules should be constructed for each individual species. As such, methods and algorithms have been developed or improved by scientists with different professional backgrounds (Shin et al., 2019; Nakabayashi and Morishita, 2020). Thus, there is a need for many users to find a single universal pipeline for datasets from different species for more accurate results (Minervini et al., 2020). In short, for the further analysis of species, an updated reference genome with a high resolution is needed to enhance the robustness of the final results.

[22] http://molevol.cmima.csic.es/castresana/Gblocks_server.html

DATA AVAILABILITY STATEMENT

The final chromosome assembly was submitted to NCBI and deposited in the BioProject under accession number PRJNA721561.

AUTHOR CONTRIBUTIONS

KC designed the study. MT and SH performed the data analysis and wrote the manuscript. PL and XG revised the manuscript. All the authors made a direct and intellectual contribution to this topic and approved the article for publication.

FUNDING

This work was financially supported by grants from the National Natural Science Foundation of China (32002235, 31602008, 31861143051, and 31872425) and the Senior Talent Foundation of Jiangsu University (19JDG039).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.718266/full#supplementary-material

REFERENCES

Arunkumar, K. P., Metta, M., and Nagaraju, J. (2006). Molecular phylogeny of silkmoths reveals the origin of domesticated silkmoth, *Bombyx mori* from Chinese *Bombyx mandarina* and paternal inheritance of Antheraea proylei mitochondrial DNA. *Mol. Phylogenet. Evol.* 40, 419–427. doi: 10.1016/j.ympev.2006.02.023

Belser, C., Istace, B., Denis, E., Dubarry, M., Baurens, F. C., Falentin, C., et al. (2018). Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* 4, 879–887. doi: 10.1038/s41477-018-0289-4

Belton, J. M., McCord, R. P., Gibcus, J. H., Naumova, N., Zhan, Y., and Dekker, J. (2012). Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* 58, 268–276. doi: 10.1016/j.ymeth.2012.05.001

Benson, G. (1999). Tandem repeats finder- a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Bocklandt, S., Hastie, A., and Cao, H. (2019). Bionano genome mapping: high-throughput, ultra-long molecule genome analysis system for precision genome assembly and haploid-resolved structural variation discovery. *Adv. Exp. Med. Biol.* 1129, 97–118. doi: 10.1007/978-981-13-6037-4_7

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi: 10.1093/bioinformatics/btu170

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119–1125. doi: 10.1038/nbt.2727

Cao, X., Huang, Y., Xia, D., Qiu, Z., Shen, X., Guo, X., et al. (2017). BmNPV-miR-415 up-regulates the expression of TOR2 via Bmo-miR-5738. *Saudi J. Biol. Sci.* 24, 1614–1619. doi: 10.1016/j.sjbs.2015.09.020

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi: 10.1093/oxfordjournals.molbev.a026334

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.* 1962, 1–14. doi: 10.1007/978-1-4939-9173-0_1

Chan, S., Lam, E., Saghbini, M., Bocklandt, S., Hastie, A., Cao, H., et al. (2018). Structural variation detection and analysis using bionano optical mapping. *Methods Mol. Biol.* 1833, 193–203. doi: 10.1007/978-1-4939-8666-8_16

Chen, F., Mackey, A. J., Stoeckert, C. J. Jr., and Roos, D. S. (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* 34, D363–D368. doi: 10.1093/nar/gkj123

Chen, H., Rangasamy, M., Tan, S. Y., Wang, H., and Siegfried, B. D. (2010). Evaluation of five methods for total DNA extraction from western corn rootworm beetles. *PLoS One* 5:e11963. doi: 10.1371/journal.pone.0011963

Chin, C. S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* 10, 563–569. doi: 10.1038/nmeth.2474

Chin, C. S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi: 10.1038/nmeth.4035

Choi, J. H., Lee, J. H., and Choi, J. W. (2020). Applications of bionano sensor for extracellular vesicles analysis. *Materials* 13:3677. doi: 10.3390/ma13173677

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

Deschamps, S., Zhang, Y., Llaca, V., Ye, L., Sanyal, A., King, M., et al. (2018). A chromosome-scale assembly of the sorghum genome using nanopore sequencing and optical mapping. *Nat. Commun.* 9:4844. doi: 10.1038/s41467-018-07271-1

Duan, J., Li, R., Cheng, D., Fan, W., Zha, X., Cheng, T., et al. (2010). SilkDB v2.0: a platform for silkworm (*Bombyx mori*) genome biology. *Nucleic Acids Res.* 38, D453–D456. doi: 10.1093/nar/gkp801

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113. doi: 10.1186/1471-2105-5-113

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Ekblom, R., and Wolf, J. B. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evol. Appl.* 7, 1026–1042. doi: 10.1111/eva.12178

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., et al. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U.S.A.* 117, 9451–9457. doi: 10.1073/pnas.1921046117

Fujimoto, S., Kawamoto, M., Shoji, K., Suzuki, Y., Katsuma, S., and Iwanaga, M. (2020). Whole-genome sequencing and comparative transcriptome analysis of *Bombyx mori* nucleopolyhedrovirus La strain. *Virus Genes* 56, 249–259. doi: 10.1007/s11262-019-01727-2

Ge, Q., Chen, L., Tang, M., Zhang, S., Liu, L., Gao, L., et al. (2018a). Analysis of mulberry leaf components in the treatment of diabetes using network pharmacology. *Eur. J. Pharmacol.* 833, 50–62. doi: 10.1016/j.ejphar.2018.05.021

Ge, Q., Zhang, S., Chen, L., Tang, M., Liu, L., Kang, M., et al. (2018b). Mulberry leaf regulates differentially expressed genes in diabetic mice liver based on RNA-Seq analysis. *Front. Physiol.* 9:1051. doi: 10.3389/fphys.2018.01051

Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A., and Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol.* 4:41. doi: 10.1186/1741-7007-4-41

Ghosh, S., and Chan, C. K. (2016). Analysis of RNA-seq data using tophat and cufflinks. *Methods Mol. Biol.* 1374, 339–361. doi: 10.1007/978-1-4939-3167-5_18

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K. Jr., Hannick, L. I., et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666. doi: 10.1093/nar/gkg770

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* 8, 1494–1512. doi: 10.1038/nprot.2013.084

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7. doi: 10.1186/gb-2008-9-1-r7

International Silkworm Genome. (2008). The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 38, 1036–1045. doi: 10.1016/j.ibmb.2008.11.004

Jayakumar, V., and Sakakibara, Y. (2019). Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. *Brief Bioinform.* 20, 866–876. doi: 10.1093/bib/bbx147

Jiang, L., and Xia, Q. (2014). The progress and future of enhancing antiviral capacity by transgenic technology in the silkworm *Bombyx mori*. *Insect Biochem. Mol. Biol.* 48, 1–7. doi: 10.1016/j.ibmb.2014.02.003

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and de Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24, 2938–2939. doi: 10.1093/bioinformatics/btn564

Kaito, C., Akimitsu, N., Watanabe, H., and Sekimizu, K. (2002). Silkworm larvae as an animal model of bacterial infection pathogenic to humans. *Microb. Pathog.* 32, 183–190. doi: 10.1006/mpat.2002.0494

Kalvari, I., Nawrocki, E. P., Argasinska, J., Quinones-Olvera, N., Finn, R. D., Bateman, A., et al. (2018). Non-coding RNA analysis using the Rfam database. *Curr. Protoc. Bioinformatics* 62:e51. doi: 10.1002/cpbi.51

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49, D192–D200. doi: 10.1093/nar/gkaa1047

Kasajima, I. (2018). Successful tips of DNA extraction and PCR of plants for beginners. *Trends Res.* 1, 1–2. doi: 10.15761/tr.1000115

Kawamoto, M., Jouraku, A., Toyoda, A., Yokoi, K., Minakuchi, Y., Katsuma, S., et al. (2019). High-quality genome assembly of the silkworm, *Bombyx mori*. *Insect Biochem. Mol. Biol.* 107, 53–62. doi: 10.1016/j.ibmb.2019.02.002

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Res.* 12, 656–664. doi: 10.1101/gr.229202

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2- accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.

Kingan, S. B., Heaton, H., Cudini, J., Lambert, C. C., Baybayan, P., Galvin, B. D., et al. (2019). A high-quality de novo genome assembly from a single mosquito using PacBio sequencing. *Genes* 10:62. doi: 10.3390/genes10010062

Koike, Y., Mita, K., Suzuki, M. G., Maeda, S., Abe, H., Osoegawa, K., et al. (2003). Genomic sequence of a 320-kb segment of the Z chromosome of *Bombyx mori* containing a kettin ortholog. *Mol. Genet. Genomics* 269, 137–149. doi: 10.1007/s00438-003-0822-6

Li, G., Qian, H., Luo, X., Xu, P., Yang, J., Liu, M., et al. (2016). Transcriptomic analysis of resistant and susceptible *Bombyx mori* strains following BmNPV infection provides insights into the antiviral mechanisms. *Int. J. Genomics* 2016:2086346. doi: 10.1155/2016/2086346

Li, G., Zhou, K., Zhao, G., Qian, H., and Xu, A. (2019). Transcriptome-wide analysis of the difference of alternative splicing in susceptible and resistant silkworm strains after BmNPV infection. *3 Biotech* 9:152. doi: 10.1007/s13205-019-1669-9

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi: 10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Lowe, T. M., and Chan, P. P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44, W54–W57. doi: 10.1093/nar/gkw413

Lu, F., Wei, Z., Luo, Y., Guo, H., Zhang, G., Xia, Q., et al. (2020). SilkDB 3.0: visualizing and exploring multiple levels of data for silkworm. *Nucleic Acids Res.* 48, D749–D755. doi: 10.1093/nar/gkz919

Maekawa, H., Takada, N., Mikitani, K., Ogura, T., Miyajima, N., Fujiwara, H., et al. (1988). Nucleolus organizers in the wild silkworm *Bombyx mandarina* and the domesticated silkworm *B. mori*. *Chromosoma* 96, 263–269. doi: 10.1007/bf00286912

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Makałowski, W., and Shabardina, V. (2019). Bioinformatics of nanopore sequencing. *J. Hum. Genet.* 65, 61–67. doi: 10.1038/s10038-019-0659-4

Meng, X., Zhu, F., and Chen, K. (2017). Silkworm: a promising model organism in life science. *J. Insect Sci.* 17:97. doi: 10.1093/jisesa/iex064

Minervini, C. F., Cumbo, C., Orsini, P., Anelli, L., Zagaria, A., Specchia, G., et al. (2020). Nanopore sequencing in blood diseases: a wide range of opportunities. *Front. Genet.* 11:76. doi: 10.3389/fgene.2020.00076

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., et al. (2004). The genome sequence of silkworm, *Bombyx mori*. *DNA Res.* 11, 27–35. doi: 10.1093/dnares/11.1.27

Nakabayashi, R., and Morishita, S. (2020). HiC-Hiker: a probabilistic model to determine contig orientation in chromosome-length scaffolds with Hi-C. *Bioinformatics* 36, 3966–3974. doi: 10.1093/bioinformatics/btaa288

Ning, D. L., Wu, T., Xiao, L. J., Ma, T., Fang, W. L., Dong, R. Q., et al. (2020). Chromosomal-level assembly of *Juglans sigillata* genome using Nanopore, BioNano, and Hi-C analysis. *Gigascience* 9:giaa006. doi: 10.1093/gigascience/giaa006

Oluwadare, O., Highsmith, M., and Cheng, J. (2019). An overview of methods for reconstructing 3-D chromosome and genome structures from Hi-C data. *Biol. Proced. Online* 21:7. doi: 10.1186/s12575-019-0094-0

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Rang, F. J., Kloosterman, W. P., and de Ridder, J. (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biol.* 19:90. doi: 10.1186/s13059-018-1462-9

Seppey, M., Manni, M., and Zdobnov, E. M. (2019). BUSCO: assessing genome assembly and annotation completeness. *Methods Mol. Biol.* 1962, 227–245. doi: 10.1007/978-1-4939-9173-0_14

Shin, S. C., Kim, H., Lee, J. H., Kim, H. W., Park, J., Choi, B. S., et al. (2019). Nanopore sequencing reads improve assembly and gene annotation of the *Parochlus steinenii* genome. *Sci. Rep.* 9:5095. doi: 10.1038/s41598-019-41549-8

Shoji, K., Hara, K., Kawamoto, M., Kiuchi, T., Kawaoka, S., Sugano, S., et al. (2014). Silkworm HP1a transcriptionally enhances highly expressed euchromatic genes via association with their transcription start sites. *Nucleic Acids Res.* 42, 11462–11471. doi: 10.1093/nar/gku862

Shoji, K., Kiuchi, T., Hara, K., Kawamoto, M., Kawaoka, S., Arimura, S., et al. (2013). Characterization of a novel chromodomain-containing gene from the silkworm, *Bombyx mori*. *Gene* 527, 649–654. doi: 10.1016/j.gene.2013.06.071

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33, W465–W467. doi: 10.1093/nar/gki458

Sun, L., Gao, T., Wang, F., Qin, Z., Yan, L., Tao, W., et al. (2020). Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology. *Mol. Ecol. Resour.* 20, 1361–1371. doi: 10.1111/1755-0998.13190

Sun, Q., Guo, H., Xia, Q., Jiang, L., and Zhao, P. (2020). Transcriptome analysis of the immune response of silkworm at the early stage of *Bombyx mori* bidensovirus infection. *Dev. Comp. Immunol.* 106:103601. doi: 10.1016/j.dci.2019.103601

Tabunoki, H., Bono, H., Ito, K., and Yokoyama, T. (2016). Can the silkworm (*Bombyx mori*) be used as a human disease model? *Drug Discov. Ther.* 10, 3–8. doi: 10.5582/ddt.2016.01011

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. (2008). Synteny and collinearity in plant genomes. *Science* 320, 486–488. doi: 10.1126/science.1153917

Tomita, M., Munetsuna, H., Sato, T., Adachi, T., Hino, R., Hayashi, M., et al. (2000). Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. *Nat. Biotechnol.* 18, 81–84. doi: 10.1038/71978

Tomita, M., Munetsuna, H., Sato, T., Adachi, T., Hino, R., Hayashi, M., et al. (2003). Transgenic silkworms produce recombinant human type III procollagen in cocoons. *Nat. Biotechnol.* 21, 52–56. doi: 10.1038/nbt771

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., et al. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* 33, 2202–2204. doi: 10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. doi: 10.1371/journal.pone.0112963

Wang, G., Zhang, J., Shen, Y., Zheng, Q., Feng, M., Xiang, X., et al. (2015). Transcriptome analysis of the brain of the silkworm *Bombyx mori* infected with *Bombyx mori* nucleopolyhedrovirus: a new insight into the molecular mechanism of enhanced locomotor activity induced by viral infection. *J. Invertebr. Pathol.* 128, 37–43. doi: 10.1016/j.jip.2015.04.001

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., et al. (2005). SilkDB: a knowledgebase for silkworm biology and genomics. *Nucleic Acids Res.* 33, D399–D402. doi: 10.1093/nar/gki116

Wang, J., Yang, J., Ying, Y. L., and Long, Y. T. (2019). Nanopore-based confined spaces for single-molecular analysis. *Chem. Asian J.* 14, 389–397. doi: 10.1002/asia.201801648

Wheeler, T. J., and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29, 2487–2489. doi: 10.1093/bioinformatics/btt403

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., et al. (2004). A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). *Science* 306, 1937–1940. doi: 10.1126/science.1102210

Xiang, H., Liu, X., Li, M., Zhu, Y., Wang, L., Cui, Y., et al. (2018). The evolutionary road from wild moth to domestic silkworm. *Nat. Ecol. Evol.* 2, 1268–1279. doi: 10.1038/s41559-018-0593-4

Xu, Y. P., Cheng, R. L., Xi, Y., and Zhang, C. X. (2013). Genomic diversity of *Bombyx mori* nucleopolyhedrovirus strains. *Genomics* 102, 63–71. doi: 10.1016/j.ygeno.2013.04.015

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555–556. doi: 10.1093/bioinformatics/13.5.555

Yang, Z., and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19, 908–917. doi: 10.1093/oxfordjournals.molbev.a004148

Yardimci, G. G., Ozadam, H., Sauria, M. E. G., Ursu, O., Yan, K. K., Yang, T., et al. (2019). Measuring the reproducibility and quality of Hi-C data. *Genome Biol.* 20:57. doi: 10.1186/s13059-019-1658-7

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. doi: 10.1093/molbev/msi237

Zhu, Z., Guan, Z., Liu, G., Wang, Y., and Zhang, Z. (2019). SGID: a comprehensive and interactive database of the silkworm. *Database* 2019:baz134. doi: 10.1093/database/baz134

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.