Insect Biochemistry and Molecular Biology 107 (2019) 53–62

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

High-quality genome assembly of the silkworm, *Bombyx mori*

Munetaka Kawamoto[a,1], Akiya Jouraku[b,1], Atsushi Toyoda[c,d], Kakeru Yokoi[b], Yohei Minakuchi[c], Susumu Katsuma[a], Asao Fujiyama[c,d], Takashi Kiuchi[a,***], Kimiko Yamamoto[b,**], Toru Shimada[a,*]

[a] Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo, 113-8657, Japan
[b] Institute of Agrobiological Sciences, National Agriculture and Food Research Organization (NARO), 1-2 Owashi, Tsukuba, Ibaraki, 305-8634, Japan
[c] Comparative Genomics Laboratory, Center for Information Biology, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan
[d] Advanced Genomics Center, National Institute of Genetics, Mishima, Shizuoka, 411-8540, Japan

ARTICLE INFO

Keywords:
Silkworm (*Bombyx mori*)
Genome assembly
Long-read sequencing
Gene prediction

ABSTRACT

In 2008, the genome assembly and gene models for the domestic silkworm, *Bombyx mori*, were published by a Japanese and Chinese collaboration group. However, the genome assembly contains a non-negligible number of misassembled and gap regions due to the presence of many repetitive sequences within the silkworm genome. The erroneous genome assembly occasionally causes incorrect gene prediction. Here we performed hybrid assembly based on 140 × deep sequencing of long (PacBio) and short (Illumina) reads. The remaining gaps in the initial genome assembly were closed using BAC and Fosmid sequences, giving a new total length of 460.3 Mb, with 30 gap regions and an N50 comprising 16.8 Mb in scaffolds and 12.2 Mb in contigs. More RNA-seq and piRNA-seq reads were mapped on the new genome assembly compared with the previous version, indicating that the new genome assembly covers more transcribed regions, including repetitive elements. We performed gene prediction based on the new genome assembly using available mRNA and protein sequence data. The number of gene models was 16,880 with an N50 of 2154 bp. The new gene models reflected more accurate coding sequences and gene sets than old ones. The proportion of repetitive elements was also reestimated using the new genome assembly, and was calculated to be 46.8% in the silkworm genome. The new genome assembly and gene models are provided in SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp).

1. Introduction

The silkworm, *Bombyx mori*, is a model insect used in various fields of biology, including physiology, biochemistry, developmental biology, neurobiology, and pathology. It has been reared worldwide for more than 5000 years for silk production, and is currently used for commercial production of medically or industrially important biomaterials based on genetic engineering. The draft genome sequence of *B. mori* was obtained and published in 2004 by Japanese and Chinese groups (Mita et al., 2004; Xia et al., 2004). Later, a more accurate assembly was performed via an international collaboration (International Silkworm Genome Consortium, 2008), and the results are available at SilkDB and KAIKObase. This assembly is frequently used in research, not only for silkworm, but also for other insects and organisms. In particular, positional cloning of mutants exhibiting various phenotypic traits revealed their crucial biological functions, such as complex larval markings (Yoda et al., 2014), mating behaviour (Fujii et al., 2011), voltinism (Sato et al., 2014), and disease resistance (Ito et al., 2008). In addition, genome information was helpful to elucidate a new mechanism of sex determination (Kiuchi et al., 2014). Simultaneously, many transcriptomic and epigenomic studies have been conducted to investigate important biosystems including metamorphosis (Nakaoka et al., 2017), gustatory perception (Guo et al., 2017), and small RNA regulation (Kawaoka et al., 2013) based on the genome sequence. Over the last several years, reverse genetic techniques have been rapidly developed, and genome editing tools such as TALEN and CRISPR/Cas9 have been commonly used in *B. mori* (Daimon et al., 2014; Sajwan et al., 2013; Wang et al., 2013), all of which require genome information.

Abbreviations: ABC, ATP-binding cassette; COE, carboxylesterase; GST, glutathione S-transferase; piRNA, PIWI-interacting RNA; TE, transposable element

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: kiuchi@ss.ab.a.u-tokyo.ac.jp (T. Kiuchi), kiya@affrc.go.jp (K. Yamamoto), shimada@ss.ab.a.u-tokyo.ac.jp (T. Shimada).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ibmb.2019.02.002
Received 15 December 2018; Received in revised form 13 February 2019; Accepted 18 February 2019
Available online 23 February 2019

Since *B. mori* was domesticated from an ancestral species in China more than 5000 years ago, it has been intensively bred for sericulture worldwide. For comparative genome analyses on variation among strains/races between the domesticated *B. mori* and the wild species *B. mandarina*, the reference genome information needs to be accurate and should cover the sequences of the 28 chromosomes adequately. Since *B. mori* is a model species in Lepidoptera, its genome information needs to be highly accurate and precise to promote comparative genomics toward understanding insect diversity. In addition, genome information is applied not only for basic studies, but also silkworm breeding and biotechnology industries. For these advanced technologies, the current genome information is not sufficient as it still contains unsequenced chromosome regions, incorrect sequence assembly, and erroneous gene predictions.

Here, we present the results of a new genome analysis of the *B. mori* p50T strain based on deep sequencing of short and long reads. We demonstrate that the assembly consists of 460.3 Mb with a small number of gap regions, and its N50 comprises 16.8 Mb in scaffolds and 12.2 Mb in contigs. We also provide high quality of new gene models estimated by the new genome assembly and available mRNA and protein sequence data. The number of gene models is 16,880 with an N50 of 2154 bp.

2. Materials and methods

2.1. Silkworms

The silkworm p50T strain was used for genome sequencing. The strain was maintained at The University of Tokyo and Kyushu University under the support from National Bioresource Project (http://silkworm.nbrp.jp). The larvae used in the present study were fed fresh mulberry leaves under a continuous cycle of 12 h light and 12 h darkness, at 25 °C.

2.2. Sequencing and de novo assembly

The posterior silk gland was collected from a single male fifth instar day-3 larva and high-quality genome DNA was extracted using QIAGEN Genomic DNA Buffer Set (QIAGEN) and Genomic-tip 100/G (QIAGEN) according to the manufacturer's protocol. Whole-genome shotgun sequencing was performed using PacBio and Illumina hybrid sequencing. A SMRTbell library (BluePippin size selection at 8 kb) for P5-C3 chemistry was constructed and run on 85 SMRT cells in a PacBio RS II system (Pacific Biosciences), generating 5,422,487 subreads with a mean read length of 6879 bases for a total of 37,303,333,465 bases. Sequencing data were then assembled using Hierarchical Genome Assembly Process (HGAP) version 3. Next, genomic DNA was sheared using a Focused-ultrasonicator (Covaris). The Illumina paired-end library (400 bp insert size) was prepared using a TruSeq DNA PCR-free LT Sample Prep Kit and run on a HiSeq2500 sequencer (Illumina). A total of 186,007,838 reads, with read length of 150 bases, were mapped against assembled sequences using BWA v0.6.2, and sequence errors were corrected. Gaps in the initial assembly were closed using the complete sequences of 39 BAC and 80 fosmid clones that were further sequenced by PacBio sequencing.

2.3. Mapping of RNA-seq reads

RNA-seq data from brain, early embryo (Kiuchi et al., 2014), epidermis (Zhang et al., 2017), fat body (Zhang et al., 2017), internal genitalia, midgut, anterior silk gland, and middle silk gland of the silkworm p50T strain were mapped onto genome assemblies with HISAT2 (Kim et al., 2015) allowing two nucleotide mismatches with multihits. The total lengths of genomes mapped with RNA reads were calculated.

RNA-seq data were also mapped onto gene models with bowtie (Langmead et al., 2009), allowing no nucleotide mismatches with multihits.

2.4. Mapping of piRNA-seq reads and piRNA cluster analysis

PIWI-interacting RNA (piRNA) sequences from 0, 6, 12, 24, and 40 h postfertilization eggs (Kawaoka et al., 2011a), ovary, and testis of p50T strain (Kawaoka et al., 2011b) were mapped onto genome assemblies with bowtie (Langmead et al., 2009), allowing no nucleotide mismatches with multihits. The piRNA sequences from ovarian cell line BmN4 GFP #8 (Kawaoka et al., 2012), ovary of MW strain, LY strain, and testis of WF strain (Kawaoka et al., 2011b) were also mapped. Total lengths of genomes mapped with piRNA reads were calculated.

piRNA clusters were defined using a previously described method (Kawaoka et al., 2013). The piRNA sequences uniquely mapped onto the genome assemblies with bowtie (Langmead et al., 2009) allowing two nucleotide mismatches were used for further analysis. When a piRNA overlapped with another piRNA, these were recognized as a single domain. Overlapping piRNA reads were merged by using BEDtools (Quinlan and Hall, 2010). When piRNA domains were located within 300 bp with each other, these were considered as a single piRNA cluster. Sequences of each piRNA cluster were obtained by original Perl script and SAMtools (Li et al., 2009). The normalized expression levels of each piRNA cluster were calculated from output of bowtie (Langmead et al., 2009) allowing two nucleotide mismatches with unique hits and low expressed piRNA clusters (< 100 reads per million) were excluded from the analysis.

2.5. Gene prediction

Genes were predicted with AUGUSTUS (Stanke et al., 2006) using raw genome assembly, species-specific profiles, and hints mentioned below. CEGMA (Parra et al., 2007) output and cDNA sequences from several cDNA libraries were used as species-specific profiles for AUGUSTUS. cDNAs are opened to the public at SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp). RepeatMasker (version 4.0.6) and RepeatModeler (version 1.0.8) (http://repeatmasker.org.) were used to generate the repeat-masked genome and repeat hints for AUGUSTUS. Predicted protein sequences of *B. mori* deposited in the NCBI database were aligned to the repeat-masked genome with exonerate (Slater and Birney, 2005) and coding sequences (CDS) part hints for AUGUSTUS were generated. RNA-seq data were mapped onto the repeat-masked genome with Tophat2 (Kim et al., 2013) and assembled with cufflinks (Trapnell et al., 2010) to create exon hints for AUGUSTUS. Exon–exon sequences were created with AUGUSTUS using Tophat2 output, then RNA-seq data were mapped onto exon–exon sequences with bowtie2 (Langmead and Salzberg, 2012) to generate intron hints. The new gene models were annotated using a blast search against nonredundant (nr) protein data sets in the NCBI database.

2.6. Identification of gene families

Glutathione S-transferase (GST) genes, ATP-binding cassette (ABC) transporter genes, cytochrome P450 genes, and carboxylesterase (COE) genes in the predicted genes of the new *B. mori* genome sequences were searched by blastp search (cutoff e-value: 1e-05) using known genes of *B. mori* as queries (extracted from a comprehensive gene set of *B. mori* (Suetsugu et al., 2013) and NCBI Reference Sequence database, including previously identified genes of the four gene families). A tblastn search was also performed against the new *B. mori* genome sequences to search for unpredicted genes. The identified genes were further examined by HMMER3 (Mistry et al., 2013) search using the Pfam database (http://pfam.xfam.org), and genes containing a conserved essential for each gene family were extracted. Identified genes were manually checked and fixed accordingly using alignment information of the known genes and/or RNA-seq data.

2.7. Identification of repetitive elements

Genome assemblies were analyzed by RepeatMasker (version 4.0.6) and RMBlast (version 2.2.27+) (http://repeatmasker.org.) in default mode with a transposable element (TE) library of *B. mori* in Silkworm Genome Research Program (http://sgp.dna.affrc.go.jp/pubdata/genomicsequences.html) as subjects.

3. Results and discussion

3.1. New genome assembly of the silkworm

In 2008, the International Silkworm Genome Consortium combined two independent data sets from classical whole-genome shotgun sequencing and assembled them with fosmid- and BAC-end sequences. The published silkworm genome contained 43,622 scaffolds, 87.4% of which was assigned to 28 chromosomes. The *B. mori* genome was updated in release 22 of Ensembl Metazoa (ASM15162v1, March 2014). The updated genome contained 43,463 scaffolds with an N50 size of 4.0 Mb; the longest scaffold was 16.2 million bp (Supplemental Table S1). However, the current genome assembly includes unsequenced chromosome regions (gaps) and incorrect sequence assembly, likely due to the enormous repetitive sequences within the genome (43.6% of the whole genome). To improve the quality and accuracy of the *B. mori* genome, we resequenced the genomic DNA from a single male p50T strain larva using PacBio long-read and Illumina short-read sequencer. The hybrid assembly of 37.3 Gb PacBio reads (equal to about 80 × coverage of the silkworm whole genome) and 27.9 Gb pairs of Illumina reads (about 60 ×) resulted in 460.3 Mb of the genomic sequence (combined total coverage of about 140 ×). Furthermore, gaps in the assembly were closed using the complete sequences of BAC and Fosmid clones. Table 1 shows a summary of the new genome assembly, containing 696 scaffolds with an N50 size of 16.8 Mb; the longest scaffold and contig N50 were 21.5 and 12.2 million base pairs, respectively. One chromosome consisted of a single scaffold with 0–5 gaps (30 gaps in total, Table 2). Lepbase (http://lepbase.org) provides published genome information for 24 lepidopteran insects. Table 3 and Supplemental Table S1 clearly shows that *B. mori* new genome assembly is one of the best reference sequences among lepidopteran insects. We provide the new genome assembly data in SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp).

3.2. Comparison of mapping coverage of RNA-seq and piRNA-seq reads onto the new and old genome assemblies

To evaluate the new *B. mori* genome assembly, we mapped available RNA-seq reads from *B. mori* for eight tissues (brain, early embryo, epidermis, fat body, internal genitalia, midgut, anterior silk gland, and middle silk gland) onto the new and old genome assemblies and calculated the total nucleotide lengths of genomic region mapped with all RNA-seq reads. RNA-seq reads were mapped to wider range on the new genome assembly compared with the old genome assembly (Fig. 1A), indicating that the new genome assembly covers wider transcriptionally active regions of the *B. mori* genome than the old genome assembly.

Repeat sequences, such as TEs, are the source of piRNAs (Kawaoka et al., 2013). We mapped piRNA reads from 11 available *B. mori* piRNA libraries (0, 6, 12, 24, and 40 h postfertilization eggs, ovaries from three strains, testes from two strains, and an ovarian cell line BmN4) onto the new and old genome assemblies and calculated the total nucleotide lengths of the genomic region mapped with all piRNA reads. The piRNA reads were then mapped to wider range on the new genome assembly than the old genome assembly (Fig. 1B). Next, we searched for piRNA clusters from the new genome assembly and compared them with those from the old version (Table 4). The number of piRNA clusters on the new *B. mori* genome assembly was remarkably decreased, but the total length, average length, median length, and N50 of the new genome assembly were remarkably increased compared with the old genome assembly (Table 4). This result indicates that separated piRNA clusters on the old genome assembly are connected on the new genome assembly because of the decrease of undetermined bases ("N") on the old genome assembly (Fig. S1). Arrays of piRNA clusters were also found in gap regions between old genome scaffolds (Fig. S2). These results suggest that repeat regions are accurately assembled in the new *B. mori* genome using the long-read sequencer.

Table 2
Overview of chromosome.

Chromosome	Total length (bp)	Number of gaps
chr1	20,666,287	1
chr2	8,396,445	0
chr3	15,212,953	0
chr4	18,737,234	0
chr5	19,061,979	3
chr6	16,650,604	1
chr7	13,944,894	1
chr8	16,262,221	1
chr9	16,796,068	1
chr10	17,614,771	1
chr11	20,440,007	5
chr12	17,580,608	1
chr13	17,735,081	2
chr14	13,345,518	0
chr15	18,440,292	0
chr16	14,337,292	0
chr17	16,840,672	1
chr18	15,699,053	0
chr19	14,801,489	2
chr20	12,370,531	0
chr21	15,310,392	2
chr22	18,482,526	3
chr23	21,465,692	0
chr24	17,359,173	3
chr25	14,548,897	0
chr26	11,473,476	1
chr27	10,930,128	1
chr28	10,609,739	0
Total	445,114,022	30

3.3. Generation and characterization of new gene models

We performed gene prediction using the newly assembled genome and identified 16,880 genes with an N50 of 2154 bp (Table 5). The average number of exons per gene was increased from 5.32 in old gene models to 6.06 in the new gene models. We identified 3636 additional genes that were not predicted in the old genome assembly (Supplemental Table S2). Since repetitive regions were correctly assembled by using the long-read sequencer, a number of transposable

Table 1
Summary of new genome assembly.

Chromosome	Total length (bp)	Number of gaps	Total length of gaps (bp)	Number of scaffolds	Min length (bp)	Max length (bp)	Scaffold N50 (bp)	Contig N50 (bp)
Whole	460,334,017	30	452,530	696	10,038	21,465,692	16,796,068	12,201,325
Chromosome	445,114,022	30	452,530	28	8,396,445	21,465,692	16,840,672	12,282,768
Unlocated	15,219,995	0	0	668	10,038	294,472	21,743	21,743

Table 3
Comparison of lepidopteran genome assembly.

	Bombyx mori 2016v1.0	*Danaus plexippus* v3	*Heliconius erato lativitta* v1
span (bp)	460,334,017	248,564,116	418,371,739
scaffold count	696	5,397	142
longest scaffold (bp)	21,465,692	6,243,218	18,493,827
scaffold N50 length (bp)	16,796,068	715,606	5,483,780
contig N50 length (bp)	12,201,325	111,043	108,877

Assembly statistics of *D. plexippus* and *H. erato lativitta* were obtained from Lepbase (Richard J Challis, Sujai Kumar, Kanchon Kumar K Dasmahapatra, Chris D Jiggins, Mark Blaxter, Lepbase: the Lepidopteran genome database. bioRxiv 056994. doi: https://doi.org/10.1101/056994).

elements were newly identified. Some duplicate or repetitive protein-coding genes such as olfactory receptor genes (Wanner et al., 2007) and chorion genes (Chen et al., 2015), were also newly identified. Next, we mapped RNA-seq reads onto the new gene models with no nucleotide mismatch and no gap allowed, and calculated the total mapped length. Fig. 2A shows that the new gene models included more RNA-seq reads than the old gene models.

To properly evaluate the quality of new gene models, we compared the length and sequence between the new and old gene models in well-known clock genes (Iwai et al., 2006; Markova et al., 2003; Ou et al., 2014; Suetsugu et al., 2013) (Fig. 2B). Divided old gene models were integrated into a single gene model by new gene prediction in *Bombyx cycle*, *period*, and *timeless*. Although incorrect predictions were observed in the 5′ or 3′ regions of some new gene models, most new gene models were consistent with experimentally sequenced coding sequences.

The accuracy of the new gene models was also confirmed manually in Z (first) and second chromosomes. The structure of the new and old gene models was compared visually using the genome browser provided in SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp/jbrowse/). Most of the both predicted genes were supported by transcriptome data (full-length cDNAs or RNA-seq reads). In the new genome assembly, nearly 20% of the genes (19.1% of Z chromosome and 19.9% of second chromosome-linked gene models) were newly predicted. When the length of CDSs was compared between predicted gene models and

Table 4
Comparison of piRNA clusters.

	New Genome	Old Genome
Total length (bp)	3,879,952	3,555,855
Number of piRNA clusters	794	1,161
Average length (bp)	4,887	3,063
Median length (bp)	3,057	1,691
Max length (bp)	38,938	39,099
Min length (bp)	41	32
N50 (bp)	9,189	5,817

Table 5
Comparison of gene prediction.

	New gene models	Old gene models
Total length (bp)	26,184,828	17,891,307
Number of contigs	16,880	14,623
Average length (bp)	1,551	1,224
Median length (bp)	1,122	867
Max length (bp)	61,167	56,289
Min length (bp)	201	87
N50 (bp)	2,154	1,698
GC content (%)	48.7	47.7

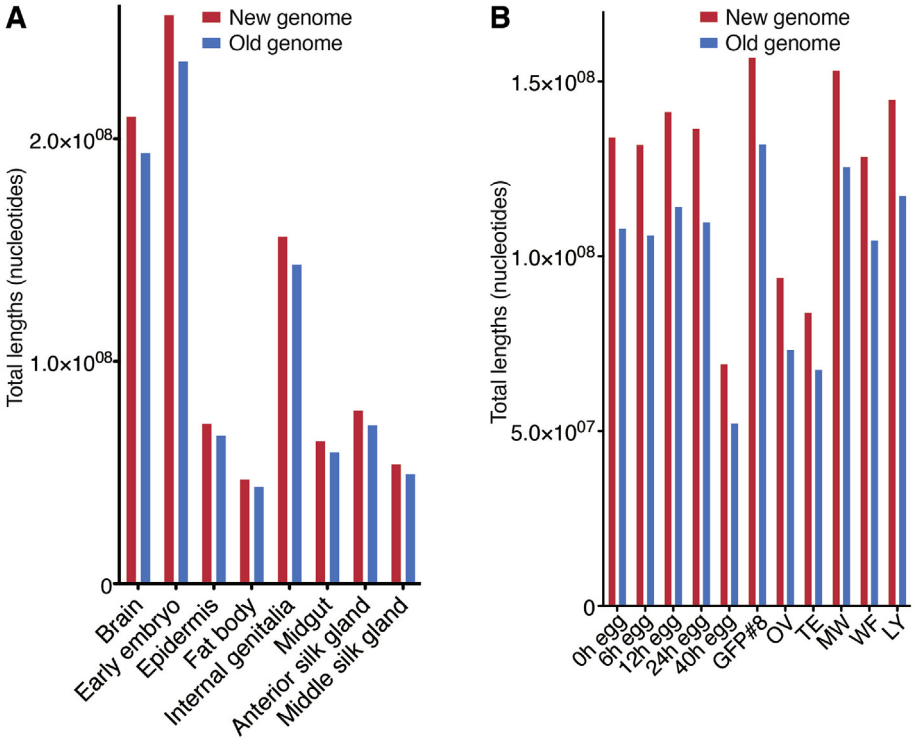

Fig. 1. Comparison of mapping results of RNA-seq and piRNA-seq reads onto the new and old genome assemblies. The vertical axis indicates the total sequence lengths of (A) RNA-seq and (B) piRNA-seq mapped regions.

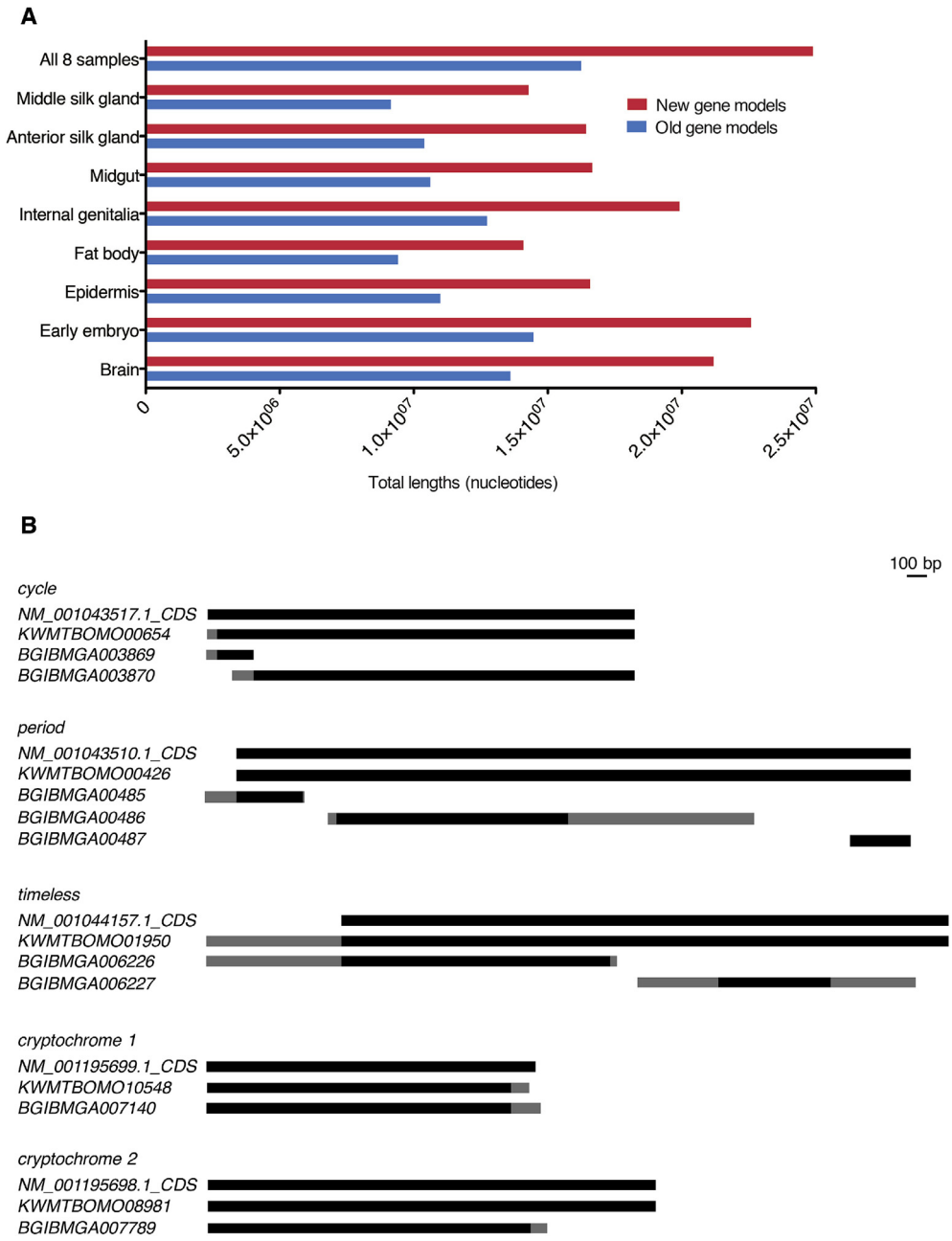

Fig. 2. Accurate gene model prediction using the new genome assembly. (A) Comparison of mapping results of RNA-seq reads onto the new and old gene models. The vertical axis indicates the total sequence lengths of RNA-seq mapped regions. (B) Experimentally cloned clock genes were compared with corresponding new (KWMTBOMO) and old (BGIBMGA) gene models. Correct and incorrect coding sequences (CDSs) are shown by black and gray boxes, respectively. Length is indicated by an upper right bar (100 bp).

native transcripts (i.e., full-length cDNAs and assembled RNA-seq contigs), the new gene models covered a wider range of CDSs than the old ones. About 10% of the new gene models (11.9% of Z chromosome and 9.2% of second chromosome-linked gene models) were assembled gene models connected by new gene prediction. According to the transcriptome data (full-length cDNAs and assembled RNA-seq contigs), more than 80% of these gene models were misassembled in the old gene models (e.g., KWMTBOMO00003, 00057, and 00064). On the other hand, around 4% of the new gene models (4.1% of Z chromosome and 4.0% of second chromosome-linked gene models) were predicted as single gene models in previous version. More than 80% of these gene models were misassembled in the new gene models (e.g., KWMTBOMO00087–00088, 00196–00197, and 00222–00223). More transcriptome data in various tissues will be required to construct more accurate gene models.

These results indicate that the new gene models predicted from the newly assembled genome reflect more accurate transcripts than the old ones. The new gene models with annotations are available in SilkBase (http://silkbase.ab.a.u-tokyo.ac.jp).

3.4. Genome-wide identification of detoxification-related gene families

Lepidopteran insects include many insect pests that have evolved a high resistance to insecticides (Cheng et al., 2017; Gouin et al., 2017; Pearce et al., 2017; You et al., 2013). Such insect pests often possess a greater number of detoxification-related genes, such as cytochrome P450, GST, COE, and ABC transporter. Highly duplicated gene clusters are often found in the gene families of resistant insect pests, which can

Table 6
Comparison of detoxification-related gene families.

Gene family	Old genome sequences	New genome sequences		
		Total	Missing in old genome	Chromosomal position was unknown/wrong in old genome
GST	23	23	0	2
ABC	51	52	1	0
P450	82	83	1	5
COE	78	87	9	4

contribute to high resistance to insecticides. Such high gene duplications of lepidopteran insect pests related with insecticide resistance are often evaluated by synteny analysis with *B. mori*. Therefore, to correctly evaluate the gene duplication level, we searched for previously missing or wrongly mapped detoxification genes in the new genome assembly.

3.4.1. GST genes

The 23 known GST genes (Yu et al., 2008) were all found in the new genome assembly, whereas no missing GST genes were found (Table 6 and Supplemental Table S3). The chromosomal positions of two GST genes were newly identified: *GSTu2* [KWMTBOMO11295 (chr19)] and *GSTe3* [KWMTBOMO13242 (chr22)] (GST gene IDs were defined previously (Yu et al., 2008)). In the old genome assembly, no GST genes were found on chromosome 22, whereas the new genome assembly revealed the chromosomal positions of all GST genes on 13 chromosomes in *B. mori*.

3.4.2. ABC transporter genes

The 51 known ABC genes (Cheng et al., 2017; Liu et al., 2011; Xie et al., 2012) were all found, and one missing ABC gene, KWMTBOMO07445M (where "M" means manually modified), was found on chromosome 12, giving a total of 52 ABC genes in *B. mori* (Table 6 and Supplemental Table S4). While this ABC gene was not identified previously, the corresponding genome sequence was also found in the old genome sequences and the gene was also predicted in the latest *B. mori* gene annotation by NCBI (accession ID: XP_012546162). This shows that gene prediction with comprehensive RNA-seq data, which was not previously used, enabled us to identify the missing gene. The ABC gene was classified into the ABCG subfamily, and shares the highest similarity (40% at the amino acid level) with KWMTBOMO07446M, which is closely and tandemly aligned with KWMTBOMO07445M. Using the new genome assembly, the greatest number of ABCG genes (nine of the 16 genes in total) were found on chromosome 12.

3.4.3. Cytochrome P450 genes

The 82 known P450 genes (Ai et al., 2011) were all found, whereas one missing P450 gene, KWMTBOMO07979-2M, was found on chromosome 13, giving a total of 83 P450 genes in *B. mori* (Table 6 and Supplemental Table S5). KWMTBOMO07979-2M shares high similarity (98% identity at the nucleotide level) with *CYP4G23* (KWMTBOMO07979-1M) (the P450 gene name was defined previously (Ai et al., 2011)) and the two P450 genes are closely and tandemly aligned. The chromosomal positions of the two pairs of tandemly aligned P450 genes were newly identified, which revealed chromosomal positions of all P450 genes on 19 chromosomes in *B. mori*. *CYP367A1* (KWMTBOMO09791) and *CYP367B1* (KWMTBOMO09792M) were found on chromosome 16 (Fig. 3A). Similarly, *CYP4S5* (KWMTBOMO12746M) and *CYP4X1* (KWMTBOMO012747M) were found on chromosome 21 where two tandemly aligned P450 genes, *CYP4AX2* and *CYP4S6*, were previously identified (Fig. 3B), showing that the four P450 genes are tandemly aligned. The two P450 genes pairs were previously unmapped to the chromosomes due to large gap regions in the old genome assembly. In particular, the new genome assembly revealed that the approximately 650 kb genomic region (including 45 genes) next to the *CYP367A1* and *CYP367B1* genes were inversely assembled in the old genome assembly. The genomic region was surrounded by two large gaps in the old genome assembly. Since repetitive sequences were found around the gap region (data not shown), it is likely to be difficult to assemble this kind of complicated genomic region without help of long-read sequencer. Furthermore, the new genome assembly revealed that *CYP340F1* (KWMTBOMO15685M), located on chromosome 18 in the old genome assembly, was actually located on chromosome 26 (Fig. 3C); therefore, all 13 CYP340 genes were located on chromosome 26 (only *CYP340F1* was located on chromosome 18 in the old genome assembly). Similarly, the corresponding genomic region on chromosome 26 in the old genome assembly consists of a large gap with repetitive sequences (data not shown), which is also considered to be difficult to correctly assemble.

3.4.4. Carboxylesterase genes

Nine COE genes were newly identified in addition to the 78 known genes (Tsubota and Shiotsuki, 2010; Yu et al., 2009), giving a total of 87 COE genes in *B. mori* (Table 6 and Supplemental Table S6). Of the new nine COE genes, the corresponding genome sequence of five alpha-esterase genes [KWMTBOMO05962M (chr10), KWMTBOMO05963M (chr10), KWMTOBOMO05965-1M (chr10), KWMTOBOMO05965-2M (chr10), and KWMTOBOMO14092 (chr23)] were newly identified in the new genome assembly. While the corresponding genome sequences of the remaining four new genes [one uncharacterized gene, KWMTBOMO01209 (chr3); three alpha-esterase genes, KWMTBOMO04596 (chr8); KWMTOBOMO05959 (chr10); and *BmCOE-34* (chr12) (*BmCOE-34* was not identified by gene prediction)] were also found in the old genome assembly, these genes were not predicted or identified as COE genes in previous studies (Tsubota and Shiotsuki, 2010; Yu et al., 2009). Furthermore, the chromosomal positions of four COE genes were newly identified: *Bmbe2* [KWMTBOMO00342 (chr1)], *Bmun2* [KWMTBOMO01212 (chr3)], *Bmie2* [KWMTBOMO08294 (chr14)], and *Bmae55* [KWMTBOMO14088 (chr23)] (the COE gene names were defined previously (Yu et al., 2009)). As with the GST and P450 genes, the new genome assembly revealed chromosomal positions of all COE genes on 20 chromosomes in *B. mori* (no COE genes were previously found on chromosome 3). While two COE genes were newly identified on chromosome 3, one of the genes, KWMTBOMO01209, may be a pseudogene because the gene length is too short (only 137 amino acids) compared with other COE genes. The new genome assembly also revealed that *Bmbe2*, which was located on chr19 in the old genome assembly, was actually located on chr1. Five COE genes were newly identified on chr10 where four tandemly aligned COE genes were previously identified (Fig. 3D), revealing that the nine COE genes are tandemly aligned and this is the largest COE gene cluster in the *B. mori* genome. All nine genes share high similarity (93%–99% identity at the nucleotide level) and were classified as alpha-esterase, which is the largest COE class (65 alpha-esterase genes were found, including eight newly identified genes) in *B. mori*. Of the new five genes on chromosome 10, four genes were not identified in the old genome assembly due to a large gap region (Fig. 3D). As with the large gap regions around the newly mapped P450 genes, repetitive sequences were found around the gap region on chromosome 10 (data not shown), indicating a similar contribution of the long-read sequencer. On chromosome 23, two alpha-esterase genes (KWMTBOMO14088 and 014092) were newly identified. These two genes are tandemly aligned and share high similarity (97% identity at the nucleotide level), like the COE genes on chromosome 10.

As shown in detoxification-related gene families, the new genome assembly offers precise gene sets for researchers. Table 7 shows the number of the four detoxification-related gene families in *B. mori* and four important noctuid insect pests of Lepidoptera, *Spodoptera litura*, *Spodoptera frugiperda* (corn and rice strains), *Helicoverpa armigera*, and *Helicoverpa zea* (Cheng et al., 2017; Gouin et al., 2017; Pearce et al., 2017). The four noctuid pests are polyphagous and have developed

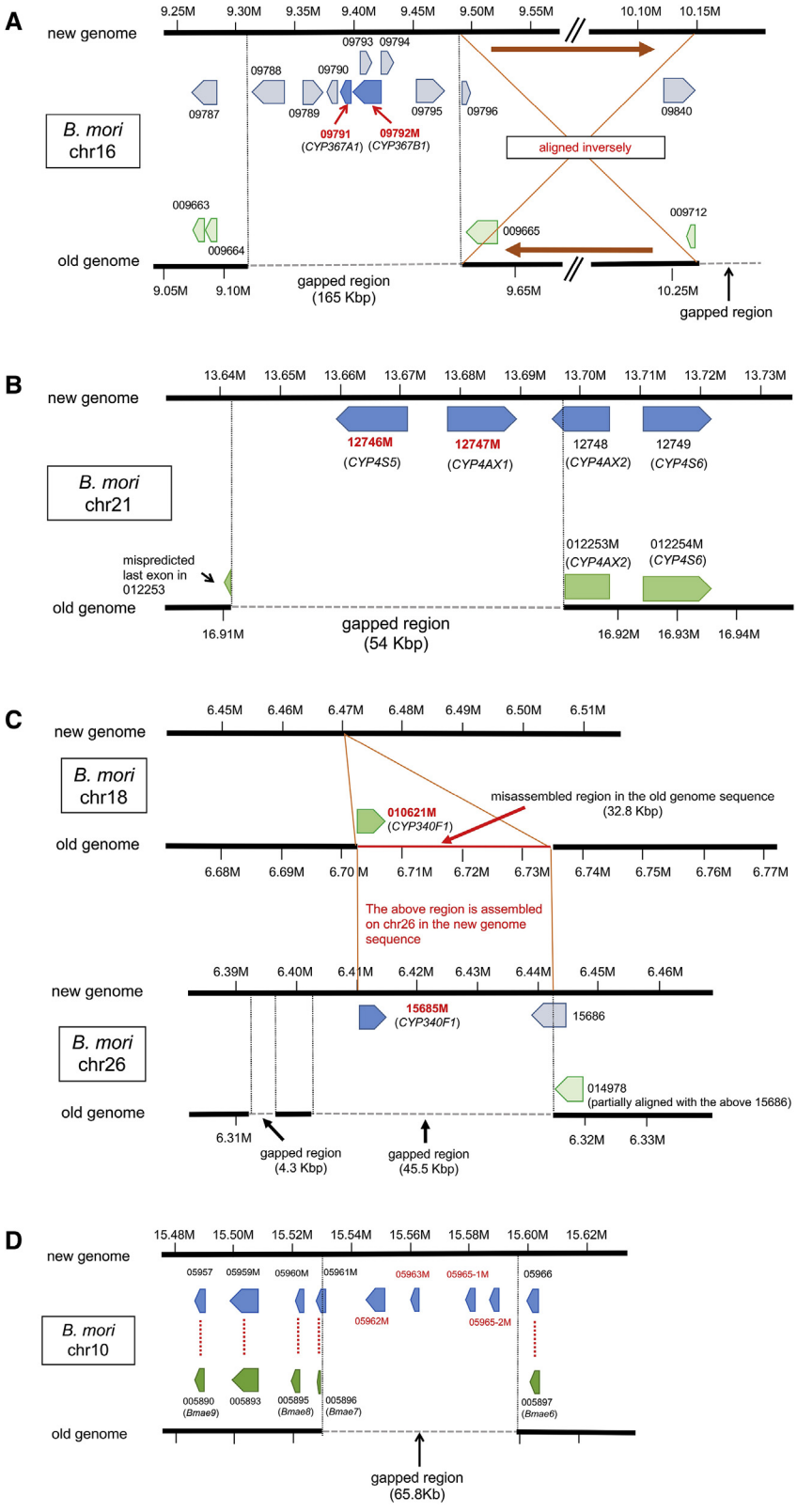

Fig. 3. Newly identified chromosomal positions of cytochrome P450 genes and carboxylesterase genes by new genome assembly. (A) The new genome revealed two tandem P450 genes on chromosome 16. The prefix of the gene ID ["KWMTBOMO" for the new genome and "Gene" (GenesetA (Suetsugu et al., 2013)) for the old genome] was omitted for each gene, and the gene ID "M" indicates manually modified genes (B–D). *CYP367A1* and *CYP367B1* are gene names annotated by Ai et al. (2011). (B) The new genome revealed four tandem P450 genes on chromosome 21. *CYP4S5*, *CYP4AX1*, *CYP4AX2*, and *CYP4S6* are gene names annotated by Ai et al. (2011) (C) The correct position of CYP340F1 on chromosome 26. *CYP340F1* is a gene name annotated by Ai et al. (2011) (D) Nine tandem COE genes including four newly identified COE genes in the chr10. Red dotted line indicated corresponding COE genes between the two genomes. *Bmae6*, *Bmae7*, *Bmae8*, and *Bmae9* are gene names annotated by Yu et al. (2009). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

resistance to insecticides. The number of GST and P450 genes in the four noctuid pests is quite larger than that in monophagous *B. mori*. This is also true for the number of COE genes in *S. litura*, whereas the number of COE genes in the remaining three noctuid pests is slightly larger than that in *B. mori*. On the other hand, the number of ABC genes is similar to each other. Since expansion of the detoxification-related genes has potential to detoxify a wider range of plant toxins and insecticides, such expansion is considered to be associated with evolution of polyphagy and insecticide resistance in noctuid pests (Cheng et al., 2017; Gouin et al., 2017; Pearce et al., 2017). The common large expansion of GST and P450 genes in the noctuid pests may indicate that the two gene families play common important role in their polyphagous

Table 7
Comparison of detoxification-related gene families between *B. mori* and four noctuid insect pests of Lepidoptera.

	GST	ABC	P450	COE	Reference
B. mori	23	52	83	87	This work
S. litura	47	54	138	110	Cheng et al. (2017)
S. frugiperda (corn strain)	46	NA	117	93	Gouin et al. (2017)
S. frugiperda (rice strain)	45	NA	135	90	Gouin et al. (2017)
H. armigera	42	54	114	97	Pearce et al. (2017)
H. zea	40	54	108	93	Pearce et al. (2017)

behaviour and development of resistance to insecticides.

3.5. Screening of repetitive elements in the new genome assembly

In the previous version of the *B. mori* genome, repetitive elements were estimated to account for 43.6% of the whole genome (International Silkworm Genome Consortium, 2008). As mentioned above, repeat regions were accurately assembled in the new genome assembly. To accurately estimate the proportion of repetitive elements in the *B. mori* genome, we searched them using RepeatMasker with the TE sequence library of *B. mori* as subjects. We also applied the same program to the old genome assembly to compare the results. All repetitive elements with detailed description in the new and old genome assemblies are shown in Supplemental Table S7 and S8, respectively, and the results are summarized in Table 8. The total bases of repetitive elements were increased from 39.30% in the old version to 46.84% in the new version. Repetitive elements were also screened in the new genome assembly using RepeatMasker with Repbase library of repetitive elements (i.e., default setting), and the percentage of repetitive elements was estimated to be 46.45% (data not shown). This suggests that the basic results are not entirely dependent on the source of TE libraries. The percentage of repetitive elements in the old genome assembly was previously estimated to be 43.6% (International Silkworm Genome Consortium, 2008; Osanai-Futahashi et al., 2008). The differences between the current and previous data may be due to the version of RepeatMasker and the search algorism used. In one previous report, the authors used RepeatMasker version 3.1.6 with WU-BLAST version 2.2.6, while we used RepeatMasker version 4.0.6 with RMBlast version 2.2.27+. The number of repetitive elements in all TE categories in the new genome assembly was smaller than in the old version, while the total length was larger. Therefore, the average length of each element identified by RepeatMasker in the new genome assembly was longer than in the old version. Additionally, the percentage of each type of TE in the new genome assembly and other repetitive elements was higher than in the old. In LINEs and LTR elements, both the average length and percentage of sequence between the old and the new version were significantly different. Table 9 shows the distribution of the numbers of TEs in each repeat class or families registered in Repbase between both genomes. More TEs and repetitive elements were identified in the old genome assembly compared with the new version in all repeat classes/families, except the LINE/Daphne and simple repeat. These results may be due to incorrect identification or an overestimation of TEs or repeat elements in the old genome assembly, constructed mainly from short-read sequence data.

Accession numbers

Raw sequence data have been submitted to DDBJ under accession number DRA004839 (PacBio) and DRA004840 (Illumina). The new *B. mori* genome assembly (Bmori_2016v1.0) have been submitted to DDBJ under accession numbers BHWX01000001-BHWX01000696. The old *B. mori* genome assembly is available under accession number GCA_000151625.1. RNA-seq data are available under accession numbers DRA006778 (brain, internal genitalia), DRA001104 (early embryo) (Kiuchi et al., 2014), DRA005299 (fat body, epidermis) (Zhang et al., 2017), DRA007063 (midgut), and DRA007064 (anterior silk gland, middle silk gland). piRNA-seq data are available under accession numbers DRA000317 (0, 6, 12, 24, and 40-h postfertilization eggs) (Kawaoka et al., 2011a), DRA000374 (ovarian cell line BmN4 GFP #8) (Kawaoka et al., 2012), DRA000173 (ovary of p50T strain, MW strain, LY strain, and testis of p50T strain, WF strain) (Kawaoka et al., 2011b).

Table 8
Classification of repeat elements in the new and old genome assembly.

New genome assembly

Total length: 460,334,017 bp
Total bases of TEs and repeat contents (% in the genome): 215,619,839 bp (46.84)

	Number of elements detected (A)	Total length (bp) (B)	Average length per element (B/A)	% of sequence
Class I (Retrotransposons)				
SINEs:	325,434	54,540,975	167.6	11.85
LINEs:	245,088	80,492,706	328.4	17.49
LTR elements:	12,112	10,733,148	886.2	2.33
Class II (DNA transposons)				
DNA elements:	44,093	12,299,355	278.9	2.67
Unclassified*:	240,013	53,982,991	224.9	11.73
Total interspersed repeats:		212,049,175		46.06
Simple repeats:	83,408	3,858,209	46.2	0.84
Low complexity:	12,377	572,001	46.2	0.12

Old genome assembly

Total length: 481,803,763 bp
Total bases of TEs and repeat contents (% in the genome): 189,363,890 bp (39.30)

	Number of elements detected (A)	Total length (bp) (B)	Average length per element (B/A)	% of sequence
Class I (Retrotransposons)				
SINEs:	334,762	54,276,626	162.1	11.27
LINEs:	265,782	59,637,161	224.4	12.38
LTR elements:	16,597	6,911,127	416.4	1.43
Class II (DNA transposons)				
DNA elements:	51,106	12,496,937	244.5	2.59
Unclassified*:	257,862	52,827,296	204.9	10.96
Total interspersed repeats:		186,149,147		38.64
Simple repeats:	81,054	3,363,293	41.5	0.7
Low complexity:	12,093	554,947	45.9	0.12

* Unclassified in repeat class or family means sequences which had homology to sequences in TE library of *B .mori* and could not be classified into any repeat classes or families registered in Repbase.

Author contributions

A.F., Ki.Y., and T.S. conceived the project. T.K. prepared the genome sample and Ki.Y. provided BAC and Fosmid clones. A.T. and Y.M. performed sequencing and genome assembly. M.K. constructed new gene models. A.J. manually identified detoxification-related gene families. M.K., A.J., A.T., Ka.Y., S.K., T.S., and T.K. evaluated the new genome assembly. All authors discussed the data and assisted with manuscript preparation. M.K., A.J., A.T., Ka.Y., S.K., T.K., and T.S. wrote the draft manuscripts, M.K. collected them, and T.K. integrated and revised them with intellectual input from all authors. T.K., Ki.Y., and T.S. supervised the project.

Table 9
Detailed information on repeat elements.

Repeat class/family	Number of elements in the library	Number of elements in the new genome	Number of elements in the old genome
SINEs	**100**		
SINE/Bm1	65	270,383	276,409
SINE/SINE	16	7,167	8,392
SINE/*Unclassified	19	60,570	61,467
LINEs	**411**		
LINE/chimera	2	29	61
LINE/CR1	25	50,667	51,816
LINE/CRE	2	138	156
LINE/Daphne	1	1,041	1,001
LINE/DNA	1	89	146
LINE/I	7	953	1,000
LINE/Jockey	63	55,490	62,225
LINE/L1	3	72	100
LINE/L2	37	13,542	15,177
LINE/LTR	3	1,926	2,516
LINE/R1	75	20,560	24,752
LINE/R2	1	27	46
LINE/R4	15	7,953	8,107
LINE/RTE	111	52,671	55,030
LINE/*Unclassified	65	50,727	56,081
LTR retrotransposons	**222**		
LTR/Copia	12	443	499
LTR/Gypsy	82	4,382	5,539
LTR/Helitron	1	21	29
LTR/Micropia	1	48	63
LTR/Pao	81	4,790	7,199
LTR/*Unclassified	45	4,351	5,409
DNA transposons	**76**		
DNA/BMC1	1	581	863
DNA/Harbinger	4	233	413
DNA/hAT	4	2,101	2,359
DNA/helitron	8	2,587	3,011
DNA/P	1	44	92
DNA/piggybac	7	401	472
DNA/Tc1 mariner	51	40,172	46,701
Others			
Penelope/*Unclassified	1	267	269
*Unclassified	880	254,380	272,739
Low complexity		12,437	12,144
Simple repeat		83,851	81,452

* Unclassified in repeat class or family means sequences which had homology to sequences in TE library of *B .mori* and could not be classified into any repeat classes or families registered in Repbase.

Conflicts of interest

None declared.

Funding

This study was supported by MEXT KAKENHI grant number JP221S0002 to A.T. and T.S., JSPS KAKENHI grant numbers JP15H02482 to T.K., S.K., and T.S., and Research Grant from NARO Gender Equality Program to Ki.Y.

Acknowledgements

We thank the Institute for Sustainable Agro-ecosystem Services, The University of Tokyo, for facilitating the mulberry cultivation. We also thank Biotron Facility at the University of Tokyo for rearing the silkworms.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ibmb.2019.02.002.

References

Ai, J., Zhu, Y., Duan, J., Yu, Q., Zhang, G., Wan, F., Xiang, Z. huai, 2011. Genome-wide analysis of cytochrome P450 monooxygenase genes in the silkworm, *Bombyx mori*. Gene 480, 42–50. https://doi.org/10.1016/j.gene.2011.03.002.

Chen, Z., Nohata, J., Guo, H., Li, S., Liu, J., Guo, Y., Yamamoto, K., Kadono-Okuda, K., Liu, C., Arunkumar, K.P., Nagaraju, J., Zhang, Y., Liu, S., Labropoulou, V., Swevers, L., Tsitoura, P., Iatrou, K., Gopinathan, K.P., Goldsmith, M.R., Xia, Q., Mita, K., 2015. A comprehensive analysis of the chorion locus in silkmoth. Sci. Rep. 5, 16424. https://doi.org/10.1038/srep16424.

Cheng, T., Wu, J., Wu, Y., Chilukuri, R.V., Huang, L., Yamamoto, K., Feng, L., Li, W., Chen, Z., Guo, H., Liu, J., Li, S., Wang, X., Peng, L., Liu, D., Guo, Y., Fu, B., Li, Z., Liu, C., Chen, Y., Tomar, A., Hilliou, F., Montagné, N., Jacquin-Joly, E., D'Alençon, E., Seth, R.K., Bhatnagar, R.K., Jouraku, A., Shiotsuki, T., Kadono-Okuda, K., Promboon, A., Smagghe, G., Arunkumar, K.P., Kishino, H., Goldsmith, M.R., Feng, Q., Xia, Q., Mita, K., 2017. Genomic adaptation to polyphagy and insecticides in a major East Asian noctuid pest. Nat. Ecol. Evol. 1, 1747–1756. https://doi.org/10.1038/s41559-017-0314-4.

Daimon, T., Kiuchi, T., Takasu, Y., 2014. Recent progress in genome engineering techniques in the silkworm, *Bombyx mori*. Dev. Growth Differ. 56, 14–25. https://doi.org/10.1111/dgd.12096.

Fujii, T., Fujii, T., Namiki, S., Abe, H., Sakurai, T., Ohnuma, A., Kanzaki, R., Katsuma, S., Ishikawa, Y., Shimada, T., 2011. Sex-linked transcription factor involved in a shift of sex-pheromone preference in the silkmoth *Bombyx mori*. Proc. Natl. Acad. Sci. U. S. A 108, 18038–18043. https://doi.org/10.1073/pnas.1107282108.

Gouin, A., Bretaudeau, A., Nam, K., Gimenez, S., Aury, J.M., Duvic, B., Hilliou, F., Durand, N., Montagné, N., Darboux, I., Kuwar, S., Chertemps, T., Siaussat, D., Bretschneider, A., Moné, Y., Ahn, S.J., Hänniger, S., Grenet, A.S.G., Neunemann, D., Maumus, F., Luyten, I., Labadie, K., Xu, W., Koutroumpa, F., Escoubas, J.M., Llopis, A., Maïbèche-Coisne, M., Salasc, F., Tomar, A., Anderson, A.R., Khan, S.A., Dumas, P., Orsucci, M., Guy, J., Belser, C., Alberti, A., Noel, B., Couloux, A., Mercier, J., Nidelet, S., Dubois, E., Liu, N.Y., Boulogne, I., Mirabeau, O., Le Goff, G., Gordon, K., Oakeshott, J., Consoli, F.L., Volkoff, A.N., Fescemyer, H.W., Marden, J.H., Luthe, D.S., Herrero, S., Heckel, D.G., Wincker, P., Kergoat, G.J., Amselem, J., Quesneville, H., Groot, A.T., Jacquin-Joly, E., Nègre, N., Lemaitre, C., Legeai, F., D'Alençon, E., Fournier, P., 2017. Two genomes of highly polyphagous lepidopteran pests (*Spodoptera frugiperda*, Noctuidae) with different host-plant ranges. Sci. Rep. 7, 11816. https://doi.org/10.1038/s41598-017-10461-4.

Guo, H., Cheng, T., Chen, Z., Jiang, L., Guo, Y., Liu, J., Li, S., Taniai, K., Asaoka, K., Kadono-Okuda, K., Arunkumar, K.P., Wu, J., Kishino, H., Zhang, H., Seth, R.K., Gopinathan, K.P., Montagné, N., Jacquin-Joly, E., Goldsmith, M.R., Xia, Q., Mita, K., 2017. Expression map of a complete set of gustatory receptor genes in chemosensory organs of *Bombyx mori*. Insect Biochem. Mol. Biol. 82, 74–82. https://doi.org/10.1016/j.ibmb.2017.02.001.

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

Ito, K., Kidokoro, K., Sezutsu, H., Nohata, J., Yamamoto, K., Kobayashi, I., Uchino, K., Kalyebi, A., Eguchi, R., Hara, W., Tamura, T., Katsuma, S., Shimada, T., Mita, K., Kadono-Okuda, K., 2008. Deletion of a gene encoding an amino acid transporter in the midgut membrane causes resistance to a *Bombyx* parvo-like virus. Proc. Natl. Acad. Sci. U. S. A 105, 7523–7527. https://doi.org/10.1073/pnas.0711841105.

Iwai, S., Fukui, Y., Fujiwara, Y., Takeda, M., 2006. Structure and expressions of two circadian clock genes, *period* and *timeless* in the commercial silkmoth, *Bombyx mori*. J. Insect Physiol. 52, 625–637. https://doi.org/10.1016/j.jinsphys.2006.03.001.

Kawaoka, S., Arai, Y., Kadota, K., Suzuki, Y., Hara, K., Sugano, S., Shimizu, K., Tomari, Y., Shimada, T., Katsuma, S., 2011a. Zygotic amplification of secondary piRNAs during silkworm embryogenesis. RNA 17, 1401–1407. https://doi.org/10.1261/rna.2709411.

Kawaoka, S., Hara, K., Shoji, K., Kobayashi, M., Shimada, T., Sugano, S., Tomari, Y., Suzuki, Y., Katsuma, S., 2013. The comprehensive epigenome map of piRNA clusters. Nucleic Acids Res. 41, 1581–1590. https://doi.org/10.1093/nar/gks1275.

Kawaoka, S., Kadota, K., Arai, Y., Suzuki, Y., Fujii, T., Abe, H., Yasukochi, Y., Mita, K., Sugano, S., Shimizu, K., Tomari, Y., Shimada, T., Katsuma, S., 2011b. The silkworm W chromosome is a source of female-enriched piRNAs. RNA 17, 2144–2151. https://doi.org/10.1261/rna.027565.111.

Kawaoka, S., Mitsutake, H., Kiuchi, T., Kobayashi, M., Yoshikawa, M., Suzuki, Y., Sugano, S., Shimada, T., Kobayashi, J., Tomari, Y., Katsuma, S., 2012. A role for transcription from a piRNA cluster in de novo piRNA production. RNA 18, 265–273. https://doi.org/10.1261/rna.029777.111.

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12, 357–360. https://doi.org/10.1038/nmeth.3317.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36. https://doi.org/10.1186/gb-2013-14-4-r36.

Kiuchi, T., Koga, H., Kawamoto, M., Shoji, K., Sakai, H., Arai, Y., Ishihara, G., Kawaoka, S., Sugano, S., Shimada, T., Suzuki, Y., Suzuki, M.G., Katsuma, S., 2014. A single female-specific piRNA is the primary determiner of sex in the silkworm. Nature 509, 633–636. https://doi.org/10.1038/nature13315.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S., 2009. Ultrafast and memory-efficient

alignment of short DNA sequences to the human genome. Genome Biol. 10, R25. https://doi.org/10.1186/gb-2009-10-3-r25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome project data Processing Subgroup, 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

Liu, S., Zhou, S., Tian, L., Guo, E., Luan, Y., Zhang, J., Li, S., 2011. Genome-wide identification and characterization of ATP-binding cassette transporters in the silkworm, *Bombyx mori*. BMC Genomics 12, 491. https://doi.org/10.1186/1471-2164-12-491.

Markova, E.P., Ueda, H., Sakamoto, K., Oishi, K., Shimada, T., Takeda, M., 2003. Cloning of *Cyc* (*Bmal1*) homolog in *Bombyx mori*: Structural analysis and tissue specific distributions. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 134, 535–542. https://doi.org/10.1016/S1096-4959(03)00004-6.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 41, e121. https://doi.org/10.1093/nar/gkt263.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y., Kadono-Okuda, K., Yamamoto, K., Ajimura, M., Ravikumar, G., Shimomura, M., Nagamura, Y., Shin-I, T., Abe, H., Shimada, T., Morishita, S., Sasaki, T., 2004. The genome sequence of silkworm, *Bombyx mori*. DNA Res. 11, 27–35. https://doi.org/10.1093/DNARES/11.1.27.

Nakaoka, T., Iga, M., Yamada, T., Koujima, I., Takeshima, M., Zhou, X., Suzuki, Y., Ogihara, M.H., Kataoka, H., 2017. Deep sequencing of the prothoracic gland transcriptome reveals new players in insect ecdysteroidogenesis. PLoS One 12. https://doi.org/10.1371/journal.pone.0172951.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., Fujiwara, H., 2008. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. 38, 1046–1057. https://doi.org/10.1016/j.ibmb.2008.05.012.

Ou, J., Deng, H.M., Zheng, S.C., Huang, L.H., Feng, Q.L., Liu, L., 2014. Transcriptomic analysis of developmental features of *Bombyx mori* wing disc during metamorphosis. BMC Genomics 15, 820. https://doi.org/10.1186/1471-2164-15-820.

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

Pearce, S.L., Clarke, D.F., East, P.D., Elfekih, S., Gordon, K.H.J., Jermiin, L.S., Mcgaughran, A., Oakeshott, J.G., Papanikolaou, A., Perera, O.P., Rane, R.V., Richards, S., Tay, W.T., Walsh, T.K., Anderson, A., 2017. Genomic innovations , transcriptional plasticity and gene loss underlying the evolution and divergence of two highly polyphagous and invasive *Helicoverpa* pest species. BMC Biol. 15, 1–30. https://doi.org/10.1186/s12915-017-0402-6.

Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

Sajwan, S., Takasu, Y., Tamura, T., Uchino, K., Sezutsu, H., Zurovec, M., 2013. Efficient disruption of endogenous *Bombyx* gene by TAL effector nucleases. Insect Biochem. Mol. Biol. 43, 17–23. https://doi.org/10.1016/j.ibmb.2012.10.011.

Sato, A., Sokabe, T., Kashio, M., Yasukochi, Y., Tominaga, M., Shiomi, K., 2014. Embryonic thermosensitive TRPA1 determines transgenerational diapause phenotype of the silkworm, *Bombyx mori*. Proc. Natl. Acad. Sci. Unit. States Am. 111, E1249–E1255. https://doi.org/10.1073/pnas.1322134111.

Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinf. 6, 31. https://doi.org/10.1186/1471-2105-6-31.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., Morgenstern, B., 2006. AUGUSTUS: *Ab initio* prediction of alternative transcripts. Nucleic Acids Res. 34, 435–439. https://doi.org/10.1093/nar/gkl200.

Suetsugu, Y., Futahashi, R., Kanamori, H., Kadono-Okuda, K., Sasanuma, S., Narukawa, J., Ajimura, M., Jouraku, A., Namiki, N., Shimomura, M., Sezutsu, H., Osanai-Futahashi, M., Suzuki, M.G., Daimon, T., Shinoda, T., Taniai, K., Asaoka, K., Niwa, R., Kawaoka, S., Katsuma, S., Tamura, T., Noda, H., Kasahara, M., Sugano, S., Suzuki, Y., Fujiwara, H., Kataoka, H., Arunkumar, K.P., Tomar, A., Nagaraju, J., Goldsmith, M.R., Feng, Q., Xia, Q., Yamamoto, K., Shimada, T., Mita, K., 2013. Large Scale full-length cDNA sequencing reveals a unique genomic landscape in a Lepidopteran model insect, *Bombyx mori*. G3 (Bethesda) 3, 1481–1492. https://doi.org/10.1534/g3.113.006239.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., Pachter, L., 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28, 511–515. https://doi.org/10.1038/nbt.1621.

Tsubota, T., Shiotsuki, T., 2010. Genomic analysis of carboxyl/cholinesterase genes in the silkworm *Bombyx mori*. BMC Genomics 11, 377. https://doi.org/10.1186/1471-2164-11-377.

Wang, Y., Li, Z., Xu, J., Zeng, B., Ling, L., You, L., Chen, Y., Huang, Y., Tan, A., 2013. The CRISPR/Cas System mediates efficient genome engineering in *Bombyx mori*. Cell Res. https://doi.org/10.1038/cr.2013.146.

Wanner, K.W., Anderson, A.R., Trowell, S.C., Theilmann, D.A., Robertson, H.M., Newcomb, R.D., 2007. Female-biased expression of odourant receptor genes in the adult antennae of the silkworm, *Bombyx mori*. Insect Mol. Biol. 16, 107–119. https://doi.org/10.1111/j.1365-2583.2007.00708.x.

Xia, Q., Zhou, Z., Lu, C., Cheng, D., Dai, F., Li, B., Zhao, P., Zha, X., Cheng, T., Chai, C., Pan, G., Xu, J., Liu, C., Lin, Y., Qian, J., Hou, Y., Wu, Z., Li, G., Pan, M., Li, C., Shen, Y., Lan, X., Yuan, L., Li, T., Xu, H., Yang, G., Wan, Y., Zhu, Y., Yu, M., Shen, W., Wu, D., Xiang, Z., Yu, J., Wang, J., Li, R., Shi, J., Li, H., Li, G., Su, J., Wang, X., Li, G., Zhang, Z., Wu, Q., Li, J., Zhang, Q., Wei, N., Xu, J., Sun, H., Dong, L., Liu, D., Zhao, S., Zhao, X., Meng, Q., Lan, F., Huang, X., Li, Y., Fang, L., Li, C., Li, D., Sun, Y., Zhang, Z., Yang, Z., Huang, Y., Xi, Y., Qi, Q., He, D., Huang, H., Zhang, X., Wang, Z., Li, W., Cao, Y., Yu, Y., Yu, H., Li, J., Ye, J., Chen, H., Zhou, Y., Liu, B., Wang, J., Ye, J., Ji, H., Li, S., Ni, P., Zhang, J., Zhang, Y., Zheng, H., Mao, B., Wang, W., Ye, C., Li, S., Wang, J., Wong, G.K.S., Yang, H., 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306, 1937–1940. https://doi.org/10.1126/science.1102210.

Xie, X., Cheng, T., Wang, G., Duan, J., Niu, W., Xia, Q., 2012. Genome-wide analysis of the ATP-binding cassette (ABC) transporter gene family in the silkworm, *Bombyx mori*. Mol. Biol. Rep. 39, 7281–7291. https://doi.org/10.1007/s11033-012-1558-3.

Yoda, S., Yamaguchi, J., Mita, K., Yamamoto, K., Banno, Y., Ando, T., Daimon, T., Fujiwara, H., 2014. The transcription factor Apontic-like controls diverse colouration pattern in caterpillars. Nat. Commun. 5. https://doi.org/10.1038/ncomms5936.

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., Zhan, D., Baxter, S.W., Vasseur, L., Gurr, G.M., Douglas, C.J., Bai, J., Wang, P., Cui, K., Huang, S., Li, X., Zhou, Q., Wu, Z., Chen, Q., Liu, C., Wang, B., Li, X., Xu, X., Lu, C., Hu, M., Davey, J.W., Smith, S.M., Chen, M., Xia, X., Tang, W., Ke, F., Zheng, D., Hu, Y., Song, F., You, Y., Ma, X., Peng, L., Zheng, Y., Liang, Y., Chen, Y., Yu, L., Zhang, Y., Liu, Y., Li, G., Fang, L., Li, J., Zhou, X., Luo, Y., Gou, C., Wang, J., Wang, J., Yang, H., Wang, J., 2013. A heterozygous moth genome provides insights into herbivory and detoxification. Nat. Genet. 45, 220–225. https://doi.org/10.1038/ng.2524.

Yu, Q., Lu, C., Li, B., Fang, S., Zuo, W., Dai, F., Zhang, Z., Xiang, Z., 2008. Identification, genomic organization and expression pattern of glutathione S-transferase in the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. 38, 1158–1164. https://doi.org/10.1016/j.ibmb.2008.08.002.

Yu, Q.Y., Lu, C., Li, W. Le, Xiang, Z.H., Zhang, Z., 2009. Annotation and expression of carboxylesterases in the silkworm, *Bombyx mori*. BMC Genomics 10, 553. https://doi.org/10.1186/1471-2164-10-553.

Zhang, H., Kiuchi, T., Wang, L., Kawamoto, M., Suzuki, Y., Sugano, S., Banno, Y., Katsuma, S., Shimada, T., 2017. *Bm-muted*, orthologous to mouse *muted* and encoding a subunit of the BLOC-1 complex, is responsible for the otm translucent mutation of the silkworm *Bombyx mori*. Gene 629, 92–100. https://doi.org/10.1016/j.gene.2017.07.071.