Insect Biochemistry and Molecular Biology 38 (2008) 1036–1045

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

The genome of a lepidopteran model insect, the silkworm *Bombyx mori*

The International Silkworm Genome Consortium [1]

ARTICLE INFO

Article history:
Received 27 November 2008
Received in revised form
28 November 2008
Accepted 28 November 2008

Keywords:
Bombyx mori
Silkworm
Genome
Transposable elements
Silk production
Gene duplication

ABSTRACT

Bombyx mori, the domesticated silkworm, is a major insect model for research, and the first lepidopteran for which draft genome sequences became available in 2004. Two independent data sets from whole-genome shotgun sequencing were merged and assembled together with newly obtained fosmid- and BAC-end sequences. The remarkably improved new assembly is presented here. The 8.5-fold sequence coverage of an estimated 432 Mb genome was assembled into scaffolds with an N50 size of ~3.7 Mb; the largest scaffold was 14.5 million base pairs. With help of a high-density SNP linkage map, we anchored 87% of the scaffold sequences to all 28 chromosomes. A particular feature was the high repetitive sequence content estimated to be 43.6% and that consisted mainly of transposable elements. We predicted 14,623 gene models based on a GLEAN-based algorithm, a more accurate prediction than the previous gene models for this species. Over three thousand silkworm genes have no homologs in other insect or vertebrate genomes. Some insights into gene evolution and into characteristic biological processes are presented here and in other papers in this issue. The massive silk production correlates with the existence of specific tRNA clusters, and of several sericin genes assembled in a cluster. The silkworm's adaptation to feeding on mulberry leaves, which contain toxic alkaloids, is likely linked to the presence of new-type sucrase genes, apparently acquired from bacteria. The silkworm genome also revealed the cascade of genes involved in the juvenile hormone biosynthesis pathway, and a large number of cuticular protein genes.

1. Introduction

The silkworm, *Bombyx mori*, has been used for silk production for about 5000 years. As a fully domesticated insect, it is dependent on humans for its survival and reproduction. It is also of great economic importance, particularly in developing countries, owing to its ease of large-scale propagation and use in silk production for the textile-industry. Moreover, with the development of biotechnology, *B. mori* has become an important bioreactor for recombinant protein production (Tamura et al., 2000; Tomita et al., 2003).

Given its history and current reliance on humans, the availability of the silkworm genome will facilitate investigations into its domestication and its comparison to the wild ancestor, *Bombyx mandarina*. In addition, *B. mori* is a model organism for Lepidoptera, the second largest insect order, which includes the most disruptive agricultural pests. This genomic resource will likely aid in understanding and combating the devastating impact of these organisms on the world's food and fiber production.

In 2004, two whole-genome shotgun (WGS) sequencing projects for male *B. mori* were reported independently by Chinese and Japanese teams (Mita et al., 2004; Xia et al., 2004). These independent data sets, however, were insufficient for building long scaffolds due to low sequencing coverage and/or lack of fosmid or BAC libraries. Here the two data sets have been merged and assembled through an international collaboration between these two groups. We first present the new assembly and features of the *B. mori* genome, then discuss some genes relevant to silkworm biology. The genetic resources, ease of transformation and extensive physiological, biochemical and molecular knowledge base on the silkworm, and now its genome sequence, all contribute to make this insect the model for Lepidoptera.

2. Methods

2.1. Annotation of repeats

An improved version of ReAS program was used to detect and assemble repeat consensus sequences from the raw 8.5× shotgun sequencing reads (Wang et al., 2002). Vector sequences were screened with Cross_match (http://www.phrap.org/) and reads

* Corresponding authors.
E-mail addresses: xbxzh@swu.edu.cn (Z. Xiang), kmita@nias.affrc.go.jp (K. Mita), xiaqy@swu.edu.cn (Q. Xia), wangjian@genomics.org.cn (J. Wang), moris@cb.k.u-tokyo.ac.jp (S. Morishita), shimada@ss.ab.a.u-tokyo.ac.jp (T. Shimada)
[1] Lists of participants and affiliations appear in Appendix section.

0965-1748/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ibmb.2008.11.004

shorter than 100 bp were removed. Candidate repeat-containing reads were identified as those having *k*-mers that occur at a frequency higher than expected based on the whole-genome shotgun coverage. Reads sharing high-depth *k*-mers were aligned to each other using Cross_match with the mat70 similarity matrix; dust was used to filter simple-sequence alignments; joining information between each pair of repeat segments was determined by refining pairwise alignment, and complete joining information among all repeat segments was used to form a connection network; finally, consensus sequences were created through searching the paths in the connection network using MUSCLE as the multi-alignment engine. The parameters for the repeat assemblies in ReAS were: 1) *k*-mer size, $K = 17$; 2) depth threshold, $D = 16$; 3) identity threshold of pairwise alignment hits, 70%. RepeatMasker v3.1.6 with WU-BLAST v2.2.6 as the search engine was used to annotate repeats in genome with the ReAS repeat library.

2.2. Construction of gene model

The widely spread TE sequences always confuse gene finders, which will result in many false positive predictions entirely composed of TEs or with partial TEs and partial real protein-coding gene sequences. During the rice gene annotation, we found that by selectively masking TEs prior to gene prediction largely allow the removal of TE contaminations while having very little effect on real genes (Li et al., unpublished). So the same strategy was used here for the silkworm. The goal was to remove as many TEs as possible, while not over masking any real gene region. On the other hand, we had to balance the false positive and false negative rates. The greatest contamination came from the ORFs inside TEs, so only for these TEs, which covered 30% of the whole genome, they were pre-masked before gene prediction.

2.3. Identification of orthologs between two species

Protein sequences of *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Gallus gallus*, and *Homo sapiens* were obtained from Ensembl (http://www.ensembl.com/), while those of *Apis mellifera* were from the honeybee official gene set (release 1). The BLASTP alignments between any two species were performed, and all reciprocally best-matching gene pairs were treated as orthologs if *E*-value $< 1e{-}10$.

2.4. Identification of seven-transmembrane domain proteins genes

To identify seven-transmembrane helix protein (7TMR) genes from *B. mori* genome sequences, we applied the automated gene discovery pipeline that had been specifically designed to find 7TMP genes (see http://sevens.cbrc.jp and Ono et al. (2005) for details). The automated gene discovery pipeline, which is composed of the gene finding stage and the 7TMP gene screening stage, was repeated until no additional 7TMP was detected. The most accurate data set ("level A" data) was obtained by the "AND" combination of the two outputs that were obtained by using the *best specificity* threshold of the sequence similarity search (*E*-value $< 10^{-80}$) and the Pfam domain (Bateman et al., 2000) assignments (*E*-value $< 10^{-10}$). This data set had an accuracy level of 99.4% sensitivity and 96.6% specificity when applied to reference data sets. The same strategy was applied to genome sequences of *D. melanogaster*, *A. mellifera* and *A. gambiae* for comparative genome analysis; their genome sequences were obtained from UCSC (http://genome.ucsc.edu/) and BeeBase (http://racerx00.tamu.edu/bee_resources.html).

3. Results and discussion

3.1. Genome assembly

In this assembly, the merged WGS read set, together with newly obtained paired ends from fosmids and BACs, provided an 8.48× sequence coverage. The fosmid and BAC clone coverage is 12.6× and 24.7×, respectively (Table S1). Genome assembly was performed using in-house RAMEN and RePS assemblers (Wang et al., 2002). Since both softwares produced similar output results, subsequent analyses used only the RAMEN assembly data set.

The assembled genome size is 432 Mb, which is consistent with the previous estimation (Xia et al., 2004). The N50 contig and scaffold size is 15.5 Kb and 3.7 Mb, respectively (Table 1). N50 contig or scaffold size is such that half of the assembled sequence is included in contigs or scaffolds of equal or larger size. The increased scaffold size provided significant improvement over the draft assembly, through the contribution of fosmid- and BAC-end data. Sequence comparison between WGS assembled data of China (*Dazao*; Xia et al., 2004) and of Japan (p50T; Mita et al., 2004) revealed 0.2% changes at nucleotide level, which caused no serious problem in assembling.

Using a high-density SNP linkage map consisting of 1577 markers (Yamamoto et al., 2008), about 87.4% of the sequence was anchored to the 28 *Bombyx* chromosomes (Table S2). Comparison between the linkage map and the assembly showed that 1532 (99.2%) of the 1544 unique markers were linearly consistent inside the scaffolds, indicating that both the assembly and the genetic map were reliable for subsequent analyses. Whole marker information is available at KAIKObase (http://sgp.dna.affrc.go.jp/KAIKObase/). Table S3 summarizes the 12 unreliable markers.

We aligned 53 independently finished BACs (total length – 8.37Mb) to the assembly and found no misjoined contigs (Fig. S1). Known cDNAs were used as an independent means to measure gene-region completeness within the assembly. Among the 767 cDNAs collected from GenBank, 96% (738) could be fully aligned on the genome with correct order and exon orientation. More than 98% of the nucleotides in the cDNA set are covered by the current assembly; a similar estimate was obtained using the 16,425 EST clusters. The sequence has been submitted to GenBank (accession numbers: DF090316–DF092116 for scaffolds; BABH01000001–BABH01088672 for contigs) and is also available for download at http://silkworm.genomics.org.cn; http://silkworm.swu.edu.cn/silkdb; and http://sgp.dna.affrc.go.jp/KAIKObase/.

Table 1
Size of assembled scaffolds and contigs. The total length of contigs amounted to 431.8 Mb, and this figure is used as the silkworm genome size. In this table, the N50 scaffold (contig, resp.) size, for example, indicates that 50% of nucleotides in the assembly occur in scaffolds (contigs) of length more than or equal to the N50 size. The number of scaffolds (contigs) longer than or equal to the N50 size is also displayed in the last column.

	Scaffold (wo/g)		Contig	
	Size (bp)	Number	Size (bp)	Number
Max	14,496,184	1	139,031	1
N10	7,612,736	5	41,915	785
N20	6,299,201	11	30,773	2006
N30	5,377,136	18	24,330	3590
N40	4,475,702	27	19,439	5588
N50	3,716,872	37	15,506	8077
N60	2,574,369	51	11,989	11,248
N70	1,776,626	72	8792	15,441
N80	1,110,220	103	5605	21,521
N90	43,109	282	1934	33,670
Total	431,756,343	43,622	431,756,343	88,842