Insect Biochemistry and Molecular Biology xxx (2009) 1–10

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

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

3.2. Genome organization

3.2.1. Repeats

The *B. mori* genome contains a large number of transposable elements (TEs), but only a few have been previously identified. Hence, starting from raw sequencing reads, we created a *de novo* repeat library using ReAS to identify and construct a consensus for repeat-containing reads having *k*-mers that occur at a higher frequency than expected based on the whole-genome shotgun coverage (Li et al., 2005b). The repeat library, including 17 known TEs in GenBank, contains 1685 sequences. Among these, 827 (35.1% of the whole genome) were confirmed through manual curation. LINEs and SINEs make up the major TE classes and composed 14.5% and 13.3% of the genome, respectively. Including unclassified sequences the repeat content in the genome is about 43.6% (Table 2). The repeat content varies greatly among insects: it is 16% in *A. gambiae* (Holt et al., 2002); 1% in *A. mellifera* (Honeybee Genome Sequencing Consortium, 2006); 33% in *T. castaneum* (Tribolium Genome Sequence Consortium, 2008), and from 2.7 to 25% in the twelve *Drosophila* species (Drosophila 12 Genome Consortium, 2007). The large genome size of *B. mori* may have resulted from a very high accumulation of repetitive sequences mainly composed of transposons (Osanai-Futahashi et al., in this issue). This was also noted for the expansion of the *Aedes aegypti* genome where 47% consists of transposable elements (Nene et al., 2007).

3.2.2. Gene model

De novo gene prediction was performed using the gene finder BGF by pre-filtering classifiable TEs (Li et al., 2005a) and gave a gene count of 16,329. This number is lower than previous estimates (Xia et al., 2004). The difference is likely due to improved genome assembly and better TE-contamination filtering. Assessment of gene prediction is generally carried out using transcription start sites (TSSs). However, given the limited availability of expressed sequence tags (EST) and full-length enriched cDNA sequences, it is currently difficult to determine TSSs. To overcome this limitation, we used a novel method that extends 5′SAGE tags (Hashimoto et al., 2004) and we utilized an Illumina–Solexa sequencer to generate large numbers of 5′-end mRNA tags. We thus aligned 15,744,044 tags of 26 nt length to 3,675,997 unique positions on the silkworm chromosomes, which represent candidate TSSs. We defined these tags as TSSs when they associated with a predicted gene and fell within one of the following ranges upstream of that gene's predicted start codon: 5000 bp, 2000 bp, 1000 bp, or 500 bp. By this method, TSSs supported the BGF predictions at 92.8%, 88.1%, 84.9%, and 83.0%, respectively.

We also constructed several other gene sets using alternative methods, including a set of known cDNAs collected from NCBI, a set of EST/Unigene defined gene fragments, and a gene set based on *Drosophila* and mosquito homology prediction. We then built a consensus gene set by merging all the gene sets using GLEAN (Elsik et al., 2007). Through this method, the estimated gene count in the silkworm is 14,623, which is slightly higher than *D. melanogaster* (14,039 genes) and 44% higher than *A. mellifera* (10,157 genes). Approximately 47% of the GLEAN genes have EST expression evidence, 38% have GO classifications, and 76% have identifiable *D. melanogaster* homologs (Table 3).

3.2.3. Silkworm-specific genes

Orthology relationship of genes from *B. mori*, *D. melanogaster*, *A. aegypti*, *A. gambiae*, *A. mellifera*, *H. sapiens*, *G. gallus* and *C. elegans* genes was assigned by phylogenetic methods. As a result, we obtained 13,450 gene families; this is because one gene family may contain orthologs from different species and paralogs within species. About 11.4% and 25.9% of the gene families are insect and vertebrate specific, respectively. Species-specific genes are particularly interesting because they are more likely to be related to divergent phenotypes. Among the 3223 silkworm-specific genes which have no homologs in any other insects or vertebrates, 1073 belong to multi-gene families. Both EST expression and GO classification data show that multi-gene families are well confirmed (Table S4). Some protein domains including EGF, immunoglobulin subtype 2, lepidopteran low molecular weight lipoprotein have many duplicates in the silkworm while they have not been identified in other insects (Table S5).

3.3. Characteristic biological processes

The *B. mori* genome provides important novel information for genes that participate in characteristic biological processes of this insect. The following papers in this special issue cover the genome-wide analyses on functions of genes controlling biological phenomena (Fujii et al., in this issue; Futahashi et al., in this issue; Roller et al., in this issue; Katsuma et al., in this issue; Chai et al., in this issue; Yu et al., in this issue; Duan et al., in this issue; Tanaka et al., in this issue), regulation of gene expression (Cheng et al., in this issue; Cao et al., in this issue; Kawaoka et al., in this issue), specific genome structure (Osanai-Futahashi et al., in this issue), and application to silkworm transgenesis (Uchino et al., in this issue).

3.3.1. Silk production

The major components of silk are fibroin and sericin. Fig. 1 presents a scheme of silk production and related genes in silk glands. Fibroin and sericin proteins are synthesized and secreted in very high quantities in posterior (PSG) and middle (MSG) parts of the silk gland during the fifth-instar larval stage, pushed forward through the lumen into the anterior silk gland and then pulled out through the spinnerets to form the cocoon filament.

Table 2
Repeat content in the silkworm genome. The repeat library includes 1668 ReAS *de novo* repetitive sequences and 17 known TEs. RepeatMasker program was used to mask repeats in the genome. Among the sequences, 827 are manually curated as TEs which composed 35.1% of the genome. Adding the remaining unclassified repeats, we estimate that repeat content in silkworm genome is about 43.6%.

	Class	Number of consensus	Total length (Mb)	% of genome
Curated	LINEs	408	62.8	14.5
	SINEs	97	57.6	13.3
	LTRs	210	6.6	1.5
	Others	112	24.6	5.7
	Total	827	151.6	35.1
Unclassified		858	36.5	8.5
Total		1685	188.1	43.6

Table 3
Summary of gene prediction. BGF is a *de novo* predicted gene set. GLEAN is an integrated gene set of multiple gene resources including *de novo* predictions, known genes, EST/Unigene defined gene fragments, and a gene set resulting from predicted *Drosophila* and mosquito homologs. Shown is the total predicted gene count, the percentage with EST support, the percentage with GO classification, the percentage with homologs in *Drosophila*, and the statistics of predicted gene models. See http://silkworm.genomics.org.cn, http://silkworm.swu.edu.cn/silkdb.

	Total predictions	% with ESTs	% with GO	% with *D. melanogaster* homologs	# of exons, mean (median)	CDS size, mean (median)	Gene size, mean (median)
BGF	16,329	59.37%	45.19%	63.68%	5.11 (4)	1096 (768)	4991 (3066)
GLEAN	14,623	63.52%	48.74%	75.63%	5.44 (4)	1223 (867)	6029 (3553)

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

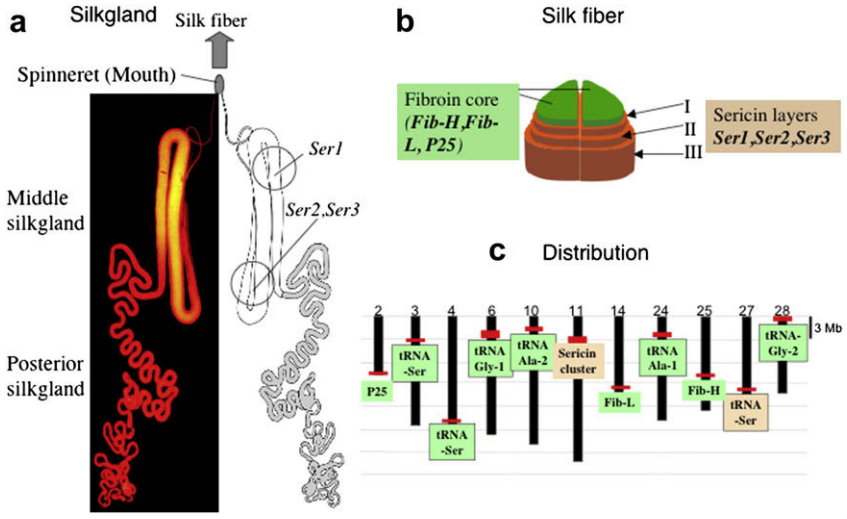

Fig. 1. Silk fiber production. The structure of the silk gland (a), silk fiber (b), and linkage map of genes related to silk production (c) are shown. The silk gland in the photo came from a transgenic silkworm that carries a chimeric fibroin-H chain – DsRed gene. The illustration of the silk gland shows the middle (translucent) and posterior (grey) part of a silk gland, and the regions where *Ser1*, *Ser2* and *Ser3* genes are expressed. The silk fiber consists of two fused filaments covered with three sericin layers. The filament core is the product of three silk genes: *Fib-H*, *Fib-L*, and *P25*. The genes *Ser1*, *Ser2*, and *Ser3* produce the sericin layers. The linkage map shows the location of genes that participate in silk protein production. The genes related to fibroin synthesis in the posterior silk glands are colored green, while the genes for sericin proteins in the middle silk glands are orange. All tRNA genes associated with silk production are clustered in each locus. Of note, the sericin genes cluster on a single chromosome, whereas the fibroin subunit genes are on different chromosomes (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fibroin, the insoluble part of silk, consists of a heavy chain (fibroin-H), a light chain fibroin (fibroin-L) and fibrohexamerin, which combine in a 6–6–1 molar ratio. Fibroin-H and -L are linked by a disulfide bond. The fibroin-H gene is composed of long repetitious domains encoding three main amino acids, Gly (46%), Ala (29%), and Ser (12%) (Zhou et al., 2000). Similarly, sericin consists of motifs enriched in Ser (Hentzen et al., 1981). The biased codon usage for these amino acids in the fibroin-H and sericin protein mRNAs (see Table S6) correlates with the differential accumulation of silk gland-specific tRNAs with cognate anticodons for the major amino acids of the fibroin-H (Zhou et al., 2000) and sericins (Garel et al., 1997; Takasu et al., 2007) during the massive synthesis of the silk proteins. This concerns two tRNA-Gly isotypes, tRNA-Gly1 and tRNA-Gly2, which recognize the major glycine codons of the fibroin-H mRNA, GGY and GGA, respectively (Chevallier and Garel, 1979). Underwood et al. (1998) found that the silk gland-specific tRNA-Ala genes with a GGC anticodon to be tightly clustered in the *B. mori* genome. Hentzen et al. (1981) reported that two species of tRNA-Ser are used for silk production: one recognizes AGY codons for sericin production in the MSG, and the other recognizes the UCA codon for fibroin-H gene expression in the PSG.

A genome-wide search in mapped scaffolds using the program tRNAscan-SE (Lowe and Eddy, 1997) identified 441 tRNA genes (Table S7). One hundred and thirty of these tRNA genes form clusters and most members of each cluster share the same anticodon. Forty-one Gly-tRNA genes were found in the *B. mori* genome, of which 13 of the Gly1 type and 8 of the Gly2 type are clustered on chromosomes 6 and 28, respectively (Table S8). Among them, Gly1 and Gly2 genes clustered in ch.6 and ch.28 (Fig. 1c) and recognize GGC/GGU and GGA codons, respectively, which correspond to the major Gly codons of the fibroin-H gene. Twenty-eighth tRNA-Ser genes are classified into 4 types, *tRNA-Ser1*, *tRNA-Ser2*, *tRNA-Ser3* and *tRNA-Ser4*, with anticodons of CGA, TGA, AGA and GCT, respectively (Table S9). Five tRNA-Ser2 genes predicted to be used for the fibroin-H gene clustered on ch.3, since they decode the UCA codons of fibroin-H mRNA, whereas four tRNA-Ser4 genes, which recognize the major serine codon AGY of sericin genes, are located on ch.27 (Table S9; Fig. 1c). The tight tRNA-Ala1 cluster is located on ch.10 (Fig. 1c). However, the tRNA-Ala sequences in this cluster are found to share the same AGC anticodon (Table S10), which is different from the reports from Underwood et al. (1998). Since the 152-base intervening sequences in the tight cluster are identical to the sequences identified by Underwood et al. (1998), we conclude that this tight tRNA cluster is a tRNA-Ala2 type. The other type of tRNA-Ala1 clusters with a 6-base intron share the anticodon GGC and is located on ch.24 (Table S10; Fig. 1c). The corresponding tRNA recognizes GCC and GCU, the major alanine codons in the fibroin-H gene. We propose that this Ala1 cluster contributes to the silk gland-specific tRNA population in relation with the high rate of fibroin-H synthesis.

All members of each tRNA gene cluster possess the same promoter with the same A-box and B-box sequence, strongly indicating that the copy number of the specific population of tRNA genes used for the synthesis of fibroin and sericin proteins in the silk gland was amplified by duplication, an adaptation to the massive synthesis of the silk proteins. Table S11 lists the tRNA isotypes among five insects with completed genomes. The numbers of silkworm tRNA genes for Gly, Ala and Ser appear to be significantly higher than those for other amino acids, consistent with the need of an adapted tRNA production to sustain the translation of the high cellular mRNA content of the silk proteins.

The fibrohexamerin gene (*Bmfhx*) is expressed with strict spatio-temporal specificity in the posterior silk gland (PSG), where the large complex of silk fibroin is assembled and secreted into the lumen (Durand et al., 1992). Unexpectedly, we identified eight novel *Bmfhx* homologs in the *B. mori* genome. *Bmfhx* is located on ch.2, while the homologs (*Bmfhxh1–8*) form a single gene cluster on ch.14, which may have resulted from gene duplication events (Fig. S2). Northern analysis revealed that *Bmfhxh1* is expressed specifically in the larval fat body and *Bmfhxh4* in the middle silk gland (MSG) (Fig. S3). Since the PSG, MSG and fat body secrete large amounts of proteins, the fibrohexamerin family may be involved in protein biosynthesis and secretion in addition to fibroin assembly in the PSG.

The previously reported sericin genes, *Ser1*, *Ser2* and *Ser3*, are located within 2 Mb of each other on ch.11 (Takasu et al., 2007;

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

Fig. 1c). *Ser1* is expressed in the middle and posterior parts of the MSG (Couble et al., 1987; Takasu et al., 2007) and composes the innermost sericin layer of silk, whereas the outermost sericin layer is comprised of Ser2 and Ser3, which are produced in the MSG anterior part. Since the proteins located in the outermost layer of the silk undergo the highest shear stress, they should have lower crystallinity and higher fluidity. Structural analysis indicates that Ser3 is more hydrophilic and more fluid than Ser1 (Takasu et al., 2007). This suggests that sericin genes underwent selection towards optimal spatial distribution and structural properties.

3.3.2. Seven-transmembrane domain proteins (7TMP)

Through comparative analyses of several insect genomes, we surveyed seven-transmembrane domain proteins (7TMPs), several of which serve as the receptors for a variety of biologically active compounds. Some of the insect 7TMPs are not coupled with G-proteins, although major subsets of this group are G-protein coupled receptors (GPCRs). To identify 7TMPs from genome sequences of the insects, we applied an 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).

In each insect genome we found a few hundred reliable full-length 7TMP genes: *B. mori* (185 genes), *D. melanogaster* (257 genes), *A. mellifera* (263 genes), and *A. gambiae* (266 genes) (Ono et al., 2005). Table 4 provides a summary of the *B. mori* 7TMP gene family distribution. Of the 185 *B. mori* 7TMP genes, the largest subfamilies were Class E (78 chemosensory receptors including olfactory and gustatory receptors) and Class A family, followed by Class B, Class C and family Frizzled/Smoothened (Table 4). We also detected 14 genes that were annotated as "Unclassified" or "Orphan". The numbers of 7TMP genes in Class A, B, C, and D families were nearly identical in the four insect species, whereas the number of chemosensory receptors is lowest in *B. mori* (Table 5). This is because *B. mori* and *A. mellifera* have fewer GRs [14] than ORs [64] (Table 5), whereas *D. melanogaster* and *A. gambiae* have about the same number of each (Hill et al., 2002; Robertson et al., 2003; Robertson and Wanner, 2006). A similar case is observed in *A. mellifera* which also has a small number of 13 GRs in comparison to the 170 ORs. We predicted a higher number [64] of ORs than that [48] predicted by Wanner et al. (2007). The difference can probably be attributed to the higher quality sequence assembly used in our study. Based on the approximation that the number of ORs agrees well with glomeruli number in the antennal lobe (60 in *B. mori*) in some insects, Wanner et al. (2007) expected the *B. mori* genome to encode close to 60 ORs. In other words our forecast is reasonable.

Table 4
7-Transmembrane helical receptor family in *B. mori*.

Class	Family	Number
A (Rhodopsin-like)	Biogenic amine	20
	Glycoprotein hormone	2
	Neuropeptide	42
	Purine	1
	Opsin	6
B (Secretin-like)	Neuropeptide	2
	HE6 like	2
	Latrophilin	2
	Methuselah-like	3
C (Metabotropic glutamate-like)	Metabotropic glutamate	5
	GABA-B	3
D (Atypical 7TMPs)	Frizzled/Smoothened	5
E (Chemosensory 7TMPs)	Odorant	60
	Gustatory	11
Orphan		9
Unclassified		5
	Total	185

Table 5
Odorant/gustatory receptor numbers in four insects.

Family	Species			
	Bombyx mori	*Drosophila melanogaster*[a]	*Anopheles gambiae*[b]	*Apis mellifera*[c]
Odorant Receptors	64	62	79	170
Gustatory Receptors	14	68	72	13
Total	78	130	151	183

[a] Cited from Hill et al. (2002).
[b] Cited from Ono et al. (2005).
[c] Cited from Robertson et al. (2003).

Phylogenetic analysis showed that olfactory receptors (ORs, red for *B. mori*; yellow for other insects) and gustatory receptors (GRs, blue for *B.* mori; green for other insects) segregated into distinct clades (Fig. 2). Several clades, including genes from different insects, show an orthologous relationship with *B. mori*. Interestingly, we observed five *B. mori*-specific clades for ORs indicated by arrows (Fig. 2).

The genomic positional information of each insect ORs/GRs is shown in Fig. 3 where ORs and GRs are colored in green and red, respectively. It is of note that *B. mori* ORs/GRs are distributed through most chromosomes, and there are fewer tight gene clusters than observed in both *D. melanogaster* and *A. gambiae*. Even higher density clusters of recently duplicated genes are found in vertebrates than in *D. melanogaster* and *A. gambiae*.

3.3.3. Host-plant specialization

B. mori subsists solely on mulberry leaves, whose latex contains very high concentrations of alkaloidal sugar-mimic alpha-glycosidase inhibitors, such as 1,4-dideoxy-1,4-imino-D-arabinitol (D-AB1), 1-deoxynojirimycin (DNJ), and 1,4-dideoxy-1,4-imino-D-ribitol.

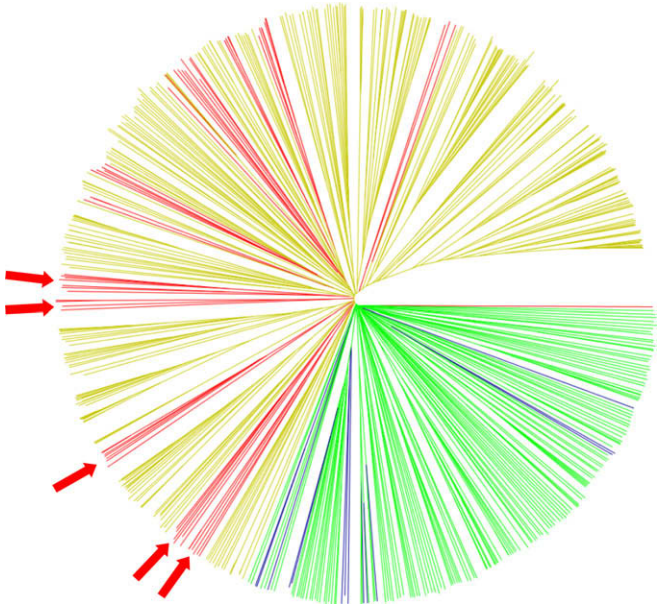

Fig. 2. Phylogenetic trees of ORs and GRs of *B. mori*, *D. melanogaster*, *A. gambiae* and *A. mellifera*. Branches of ORs and GRs are colored in red and blue for *Bombyx mori*, while orange and green for other three insects, respectively. The red arrows indicate the five *Bombyx mori*-specific clades (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

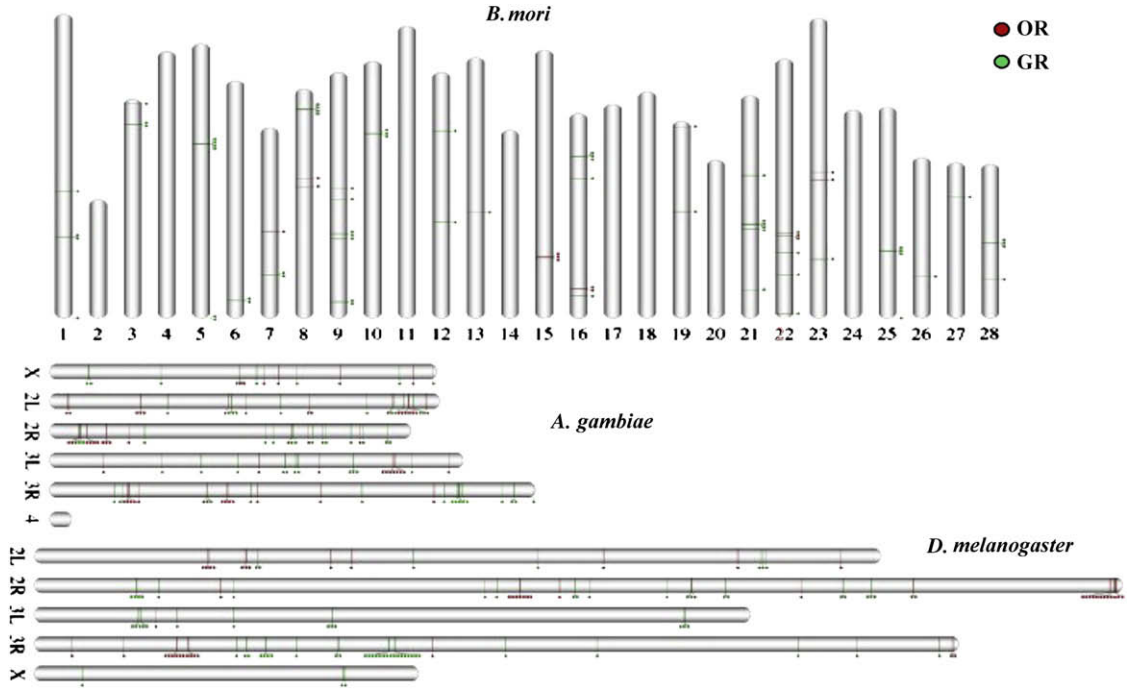

Fig. 3. Chromosomal distribution of 7TMPs of *B. mori*, *A. gambiae* and *D. melanogaster*. Green lines and dots denote OR genes, and red lines and dots represent GRs. Dots beside chromosomes indicate respective OR/GR genes (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

These sugar-mimic alkaloids are highly toxic to caterpillars other than the silkworm. *B. mori*, thus, has mechanisms to overcome these toxic substances (Konno et al., 2006), and its genome provides evidence for the adaptations that underlie these mechanisms. D-AB1 and DNJ are strong inhibitors of the sucrase alpha-glucosidase (EC 3.2.1.20), but do not inhibit the activity of beta-fructofuranosidase (EC 3.2.1.26). Alpha-glucosidase genes are present in various organisms, including bacteria, fungi, plants and animals; however, beta-fructofuranosidase genes have only been found in *B. mori*. Indeed, Daimon et al. (2008) have identified two beta-fructofuranosidase genes designated *BmSuc1* and *BmSuc2*, both of which have been putatively acquired by horizontal gene transfers from bacteria during evolution. Our present genome assembly confirms that *BmSuc1* and *BmSuc2* are the only genes coding for beta-fructofuranosidases, and both of them are located on ch.17. *BmSuc1* is expressed in the midgut, and encodes a functional beta-fructofuranosidase, whose activity is not inhibited by the mulberry latex alkaloids D-AB1 or DNJ (Daimon et al., 2008). On the other hand, the genome assembly reveals the presence of four alpha-glucosidases that are homologous to *Drosophila* maltases. These four genes, *BmMal1*, *BmMal2*, *BmMal3*, and *BmMal4*, are located on ch.4. Based on the EST database, three of them are expressed specifically in midgut, suggesting that they are functional alpha-glucosidase genes. These findings indicate that *B. mori* is a unique organism that utilizes beta-fructofuranosidases as digestive enzymes, in addition to alpha-glucosidases that are commonly present in the animal kingdom (Table S12). Further analyses of these two types of sucrases may elucidate how *B. mori* evolved to bypass the mulberry plants' defense mechanism.

3.3.4. Juvenile hormone pathway

Juvenile hormone (JH) in insects plays important roles in regulating metamorphosis, reproduction, diapause, and other physiological processes. Lepidopteran insects produce unique ethyl-branched JHs that are not found in any other insect order, indicating the need for specific biochemical adaptations (Goodman and Granger, 2005), perhaps including the evolution of genes specific to the JH biosynthesis pathway. We identified most of the key genes involved in JH biosynthesis, including epoxidase (Helvig et al., 2004) and JH acid O-methyltransferase (JHAMT) (Shinoda and Itoyama, 2003) in *B. mori* genome (Fig. 4). All these genes are present in a single copy in the *D. melanogaster*, *A. gambiae*, and *A. mellifera* genomes, but in the *B. mori* genome there are three copies of farnesylpyrophosphate synthase (FPPS 1–3) and six copies of JHAMT-like methyltransferase genes. Although the function of these genes is still unclear, we speculate that the precise sequential assembly of homoisoprenoid units into the higher homologs of JH III in Lepidoptera may involve specialized FPPS enzymes.

3.3.5. Cuticular proteins

Our analysis predicted 220 putative genes encoding cuticular proteins in the *B. mori* genome. More than 80% of these are present in gene clusters (Fig. 5; Futahashi et al., in this issue). Gene clusters of cuticular protein genes have been also reported in *Drosophila* (Karouzou et al., 2007), *Anopheles* (Cornman et al., 2008), and *Tribolium* (Tribolium Genome Sequence Consortium, 2008).

There are several different types of protein motifs found in cuticular proteins. One of the most common is the chitin-binding R&R consensus motif (Willis et al., 2005), which can be classified into RR1, RR2 and RR3 motifs. *B. mori* genome encodes at least 148 putative RR proteins (56 RR1, 89 RR2 and 3 RR3), which is far larger than the number of genes found in *D. melanogaster* (101 RR; Karouzou et al., 2007), *T. castaneum* (102 RR; Tribolium Genome Sequence Consortium, 2008), and *A. mellifera* (28 RR; Honeybee Genome Sequencing Consortium, 2006). RR1 and RR2 genes formed large clusters (Figs. 5 and 6). Glycine-rich repeat cuticular proteins and other cuticular proteins also showed distinct gene clustering in the *B. mori* genome (Fig. 5; Futahashi et al., in this issue). Ch.18 contained such a 7- and 3-gene cluster of glycine-rich

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

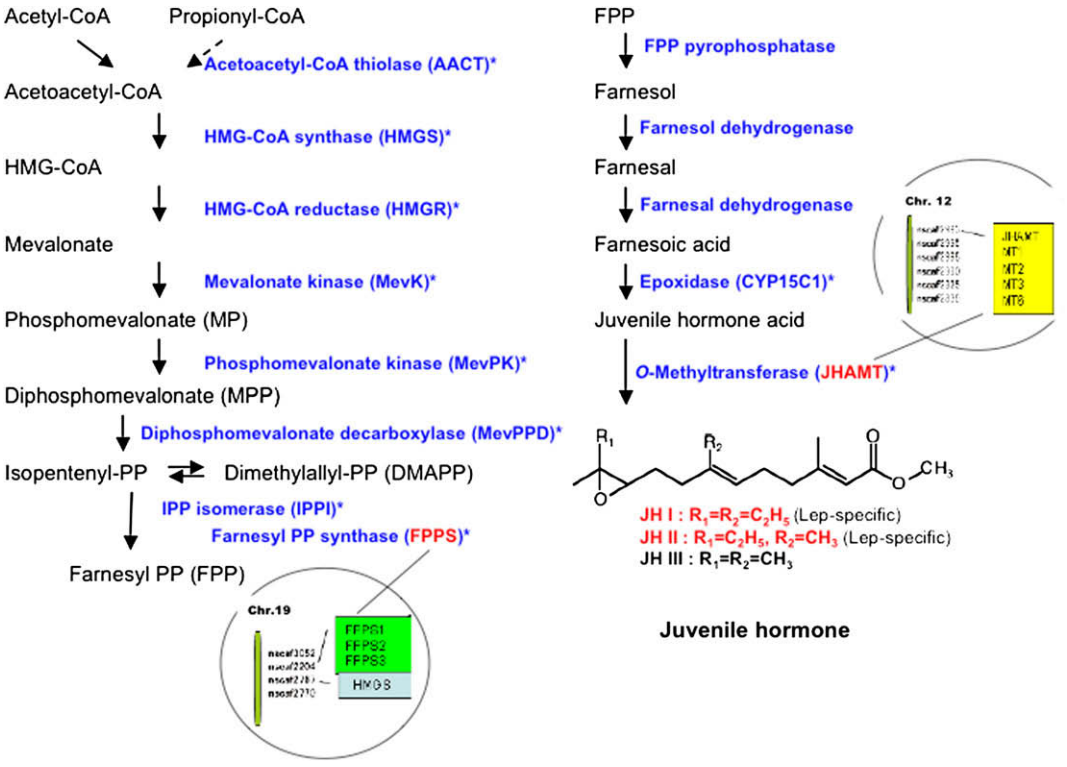

Fig. 4. JH biosynthetic pathway and genes. JH is a unique hormone in insects regulating metamorphosis, reproduction, diapause and other important physiological processes. JH is produced in the corpora allata via a mevalonate pathway and the following JH-specific pathway (Goodman and Granger, 2005). The genes identified in this work are indicated with asterisks. FPP pyrophosphatase, farnesol dehydrogenase, and farnesal dehydrogenase are putative enzymes that are supposed to catalyze the pathway from FPP to farnesoic acid. Most insects produce only a single form of JH, JH III. However, the Lepidoptera produce four ethyl-branched homologs of JH III: JH0, JH I, 4-methyl JH I, and JH II. FPPS 1–3 localize to a small region on chromosome 19 (nscaf2204). An authentic JHAMT and four homologous methyltransferase genes localize on chromosome 12 (nscaf2993).

cuticular proteins. Other currently unclassified putative cuticular protein genes also showed extensive gene clustering (e.g. a 22-gene cluster on Ch.11) (Fig. 5). The four Tweedle-motif cuticular protein genes are not clustered in the genome. *B. mori* EST comparison and RT-PCR analyses (Mita et al., 2003; Futahashi et al., in this issue; Okamoto et al., 2008) showed distinct expression patterns for each cuticular protein gene in a cluster, even between adjacent genes, indicating that the structure and regulation of each cuticular protein gene in the clusters may have been rearranged or altered during evolution.

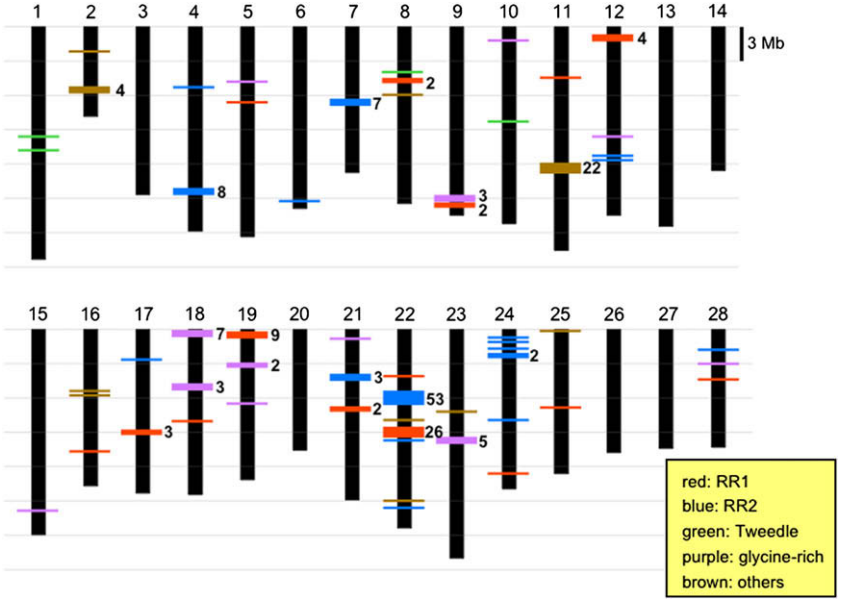

Fig. 5. Clustering of *Bombyx* cuticular protein genes. 220 predicted cuticular protein genes were distributed on the different chromosomes. The genes composing a cluster share the same motif such as RR1, RR2, RR3, or glycine-rich repeat, while Tweedle-motif genes (Guan et al., 2006) are dispersed in the *B. mori* genome. The number of genes in each cluster is indicated.

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004

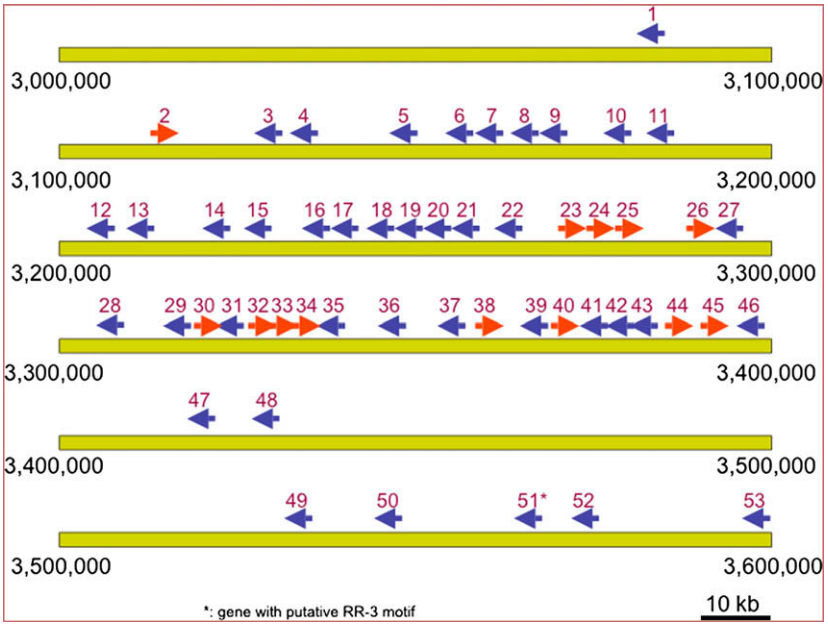

Fig. 6. The largest Cluster of *B. mori* cuticular protein genes is on chromosome 22. Most of the cuticular protein genes (52 of 53 genes) belong to the RR2 class.

Acknowledgments

This project was supported by Ministry of Science and Technology of the People's Republic of China (2005CB121000, 2007CB815701, 2007CB815703, 2007CB815705, 2006AA10A117, 2006AA10A118, 2006AA02Z334, 2006AA10A121), National Natural Science Foundation of China (30471313, 30571407, 90403130, 90608010, 30221004, 90612019), the Chinese Academy of Science (GJHZ0701-6; KSCX2-YWN-023), the 111 Project (B07045), Ministry of Education of the People's Republic of China (Program for Changjiang Scholars and Innovative Research Team in University), Chongqing Municipal Government, Ministry of Agriculture of the People's Republic of China, and the Chinese Municipal Science and Technology Commission (D07030200740000). Also we thank the Ministry of Agriculture, Forestry and Fisheries of Japan, a Grant-in-Aid for Scientific Research on Priority Areas "Genome" from the Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT), the Japan Science and Technology Corporation (JST), Human Genome Center, University of Tokyo, and grants from the Program for Promotion of Basic Research Activities for Innovated Biosciences (PROBRAIN) , and the National Bioresource Project, MEXT. The final manuscript was edited with the assistance of Laurie Goodman.

The International Silkworm Genome Consortium

Chinese group

Qingyou Xia[a,b,2,*], Jun Wang[c,d,2], Zeyang Zhou[a,2], Ruiqiang Li[c,d,2], Wei Fan[c], Daojun Cheng[a], Tingcai Cheng[a,b], Junjie Qin[c,e], Jun Duan[a,b], Hanfu Xu[a], Qibin Li[c,e], Ning Li[c,e], Mingwei Wang[c], Fangyin Dai[a], Chun Liu[a], Ying Lin[a], Ping Zhao[a], Huijie Zhang[a], Shiping Liu[a], Xingfu Zha[a], Chunfeng Li[a], Aichun Zhao[a], Minhui Pan[a], Guoqing Pan[a], Yihong Shen[a], Zhihong Gao[a], Zilong Wang[a], Genhong Wang[a,b], Zhengli Wu[a], Yong Hou[a], Chunli Chai[a], Quanyou Yu[a,b], Ningjia He[a], Ze Zhang[a,b], Songgang Li[c], Huanming Yang[c], Cheng Lu[a], Jian Wang[c,*] and Zhonghuai Xiang[a,*]

Japanese group

Kazuei Mita[f,2,*], Masahiro Kasahara[g,2], Yoichiro Nakatani[g,2], Kimiko Yamamoto[f,2], Hiroaki Abe[h], Brudrul Ahsan[g], Takaaki Daimon[i], Koichiro Doi[g], Tsuguru Fujii[i], Haruhiko Fujiwara[j], Asao Fujiyama[k], Ryo Futahashi[j], Shin-ichi Hashimoto[l], Jun Ishibashi[f], Masafumi Iwami[m], Keiko Kadono-Okuda[f], Hiroyuki Kanamori[n], Hiroshi Kataoka[j], Susumu Katsuma[i], Shinpei Kawaoka[i], Hideki Kawasaki[o], Yuji Kohara[p], Toshinori Kozaki[f], Reginaldo M. Kuroshu[g], Seigo Kuwazaki[f], Kouji Matsushima[l], Hiroshi Minami[q], Yukinobu Nagayasu[g], Tatsuro Nakagawa[j], Junko Narukawa[f], Junko Nohata[f], Kazuko Ohishi[p], Yukiteru Ono[r], Mizuko Osanai-Futahashi[j], Katsuhisa Ozaki[s], Wei Qu[g], Ladislav Roller[f,t], Shin Sasaki[g], Takuji Sasaki[u], Atsushi Seino[f], Masaru Shimomura[f], Michihiko Shimomura[q], Tadasu Shin-I[p], Tetsuro Shinoda[f], Takahiro Shiotsuki[f], Yoshitaka Suetsugu[f], Sumio Sugano[v], Makiko Suwa[r], Yutaka Suzuki[v], Shigeharu Takiya[w], Toshiki Tamura[f], Hiromitsu Tanaka[f], Yoshiaki Tanaka[f], Kazushige Touhara[j], Tomoyuki Yamada[g], Minoru Yamakawa[f], Naoki Yamanaka[j], Hiroshi Yoshikawa[s], Yang-Sheng Zhong[o], Toru Shimada[i,*] and Shinichi Morishita[g,*]

[2]These authors contributed equally to this work.

Affiliations

[a]The Key Sericultural Laboratory of Agricultural Ministry, Southwest University, Chongqing 400715, China. [b]Institute of Agronomy and Life Sciences, Chongqing University, Chongqing 400030,China. [c]Beijing Genomics Institute at Shenzhen, Shenzhen 518083, China. [d]Department of Biochemistry and Molecular Biology, University of Southern Denmark, Odense M DK-5230, Denmark. [e]Beijing Institute of Genomics of the Chinese Academy of Sciences, Beijing Genomics Institute, Beijing 101300, China. [f]National Institute of Agrobiological Sciences, Owashi 1-2, Tsukuba, Ibaraki 305-8634, Japan. [g]Department of Computational Biology, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-0882, Japan. [h]Department of Biological Production, Faculty of Agriculture, Tokyo University of Agriculture and Technology, Fuchu, Tokyo 183-8509, Japan. [i]Department of Agricultural and Environmental Biology,

Graduate School of Agricultural and Life Sciences, The University of Tokyo, Tokyo 113-8657, Japan. [j]Department of Integrated Biosciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8562, Japan. [k]National Institute of Informatics, Tokyo 101-8430, Japan. [l]Department of Molecular Preventive Medicine, School of medicine, The University of Tokyo, Tokyo 113-0033, Japan. [m]Division of Life Sciences, Graduate School of Science and Technology, Kanazawa University, Kanazawa 920-1192, Japan. [n]Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba 305-0854, Japan. [o]Faculty of Agriculture, Utsunomiya University, Utsunomiya 321-8505, Japan. [p]Center for Genetic Resource Information, National Institute of Genetics, Mishima 411-8540, Japan. [q]Genome Project Department, Tsukuba Division, Mitsubishi Space Software Co., Ltd., Tsukuba 305-8602, Japan. [r]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan. [s]JT Biohistory Research Hall, Takatsuki 569-1125, Japan. [t]Institute of Zoology, Slovak Academy of Sciences, Dubravska 9, 84506 Bratislava, Slovakia. [u]National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan. [v]Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan. [w]Division of Genome Dynamics, Hokkaido University, Sapporo 060-0810, Japan.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ibmb.2008.11.004.

References

Bateman, A., Bimey, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266.

Cao, J., Tong, C., Wu, X., Lv, J., Yang, Z., Jin, Y. Identification of conserved microRNAs in *Bombyx mori* (silkworm) and regulation of fibroin L chain production by microRNAs in heterologous system. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.09.008.

Chai, C., Zhang, Z., Huang, F., Wang, X., Yu, Q., Liu, B., Tian, T., Xia, Q., Lu, C., Xiang, Z. A genomewide survey of homeobox genes and identification of novel structure of the Hox cluster in the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.06.008.

Cheng, D., Xia, Q., Duan, J., Wei, L., Huang, C., Li, Z., Wang, G., Xianga, Z. Nuclear receptors in *Bombyx mori*: insights into genomic structure and developmental expression. Insect Biochem. Mol. Biol., doi:10.1016/j.ibmb.2008.09.013.

Chevallier, A., Garel, J.P., 1979. Studies on tRNA adaptation, tRNA turnover, precursor tRNA and tRNA gene distribution in *Bombyx mori* by using two-dimensional polyacrylamide gel electrophoresis. Biochemie 61, 245–262.

Cornman, R.S., Togawa, T., Dunn, W.A., Dunn, W.A., He, N., Emmons, A.C., Willis, J.H., 2008. Annotation and analysis of a large cuticlular protein family with the R&R consensus in *Anopheles gambiae*. BMC Genomics 9, 22.

Couble, P., Michaille, J.J., Garel, A., Couble, M.L., Prudhomme, J.C., 1987. Developmental switches of sericin mRNAs splicing in individual cells of *Bombyx mori* silkgland. Dev. Biol. 124, 431–440.

Daimon, T., Taguchi, T., Meng, Y., Katsuma, S., Mita, K., Shimada, T., 2008. Beta-fructofuranosidase genes of the silkworm, *Bombyx mor*i: insight into enzymatic adaptation of *B. mori* to toxic alkaloids in mulberry latex. J. Biol. Chem. 283, 15271–15279.

Drosophila 12 Genome Consortium, 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450, 203–218.

Duan, J., Xia, Q., Cheng, D., Zha, X., Zhao, P., Xiang, Z. Species-specific expansion of C2H2 zinc-finger genes and their expression profiles in silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.08.005.

Durand, B., Drevet, J., Couble, P., 1992. P25 gene regulation in *Bombyx mori* silk gland: two promoter-binding factors have distinct tissue and developmental specificities. Mol. Cell. Biol. 12, 5768–5777.

Elsik, C.G., Mackey, A.J., Reese, J.T., Milshina, N.V., Roos, D.S., Weinstock, G.M., 2007. Creating a honey bee consensus gene set. Genome Biol. 8, R13.

Fujii, T., Abe, H., Katsuma, S., Mita, K., Shimada, T. Mapping of sex-linked genes onto the genome sequence using various aberrations of the Z chromosome in *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.03.004.

Futahashi, R., Okamoto, S., Kawasaki, H., Zhong, Y.S., Iwanaga, M., Mita, K., Fujiwara, H. Genome-wide identification of cuticular protein genes in the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.05.007.

Garel, A., Deleage, G., Prudhomme, J.C., 1997. Structure and organization of the *Bombyx mori* sericin 1 gene and of the sericins 1 deduced from the sequence of the Ser 1B cDNA. Insect Biochem. Mol. Biol. 27, 469–477.

Goodman, W.G., Granger, N.A., 2005. The juvenile hormones. In: Comprehensive Molecular Insect Science, vol. 3. Elsevier, Amsterdam, pp. 319–408.

Guan, X., Middlebrooks, B.W., Alexander, S., Wasserman, S.A., 2006. Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. 103, 16794–16799.

Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S., Matsushima, K., 2004. 5′-End SAGE for the analysis of transcriptional start sites. Nature Biotechnol. 22, 1146–1149.

Helvig, C., Koener, J.F., Unnithan, G.C., Feyereisen, R., 2004. CYP15A1, the cytochrome P450 that catalyzes epoxidation of methyl farnesoate to juvenile hormone III in cockroach corpora allata. Proc. Natl. Acad. Sci. U.S.A. 101, 4024–4029.

Hentzen, D., Chevallier, A., Garel, J.P., 1981. Differential usage of isoaccepting tRNA Ser species in silk glands of *Bombyx mori*. Nature 290, 267–269.

Hill, C.A., Fox, A.N., Pitts, R.J., Kent, L.B., Tan, P.L., Chrystal, M.A., Cravchik, A., Collins, F.H., Robertson, H.M., Zwiebel, L.J., 2002. G protein-coupled receptors in *Anopheles gambiae*. Science 298, 176–178.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R., Wincker, P., Clark, A.G., Ribeiro, J.M., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Lai, Z., Kraft, C.L., Abril, J.F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H., de Berardinis, V., Baldwin, D., Benes, V., Biedler, J., Blass, C., Bolanos, R., Boscus, D., Barnstead, M., Cai, S., Center, A., Chaturverdi, K., Christophides, G.K., Chrystal, M.A., Clamp, M., Cravchik, A., Curwen, V., Dana, A., Delcher, A., Dew, I., Evans, C.A., Flanigan, M., Grundschober-Freimoser, A., Friedli, L., Gu, Z., Guan, P., Guigo, R., Hillenmeyer, M.E., Hladun, S.L., Hogan, J.R., Hong, Y.S., Hoover, J., Jaillon, O., Ke, Z., Kodira, C., Kokoza, E., Koutsos, A., Letunic, I., Levitsky, A., Liang, Y., Lin, J.J., Lobo, N.F., Lopez, J.R., Malek, J.A., McIntosh, T.C., Meister, S., Miller, J., Mobarry, C., Mongin, E., Murphy, S.D., O'Brochta, D.A., Pfannkoch, C., Qi, R., Regier, M.A., Remington, K., Shao, H., Sharakhova, M.V., Sitter, C.D., Shetty, J., Smith, T.J., Strong, R., Sun, J., Thomasova, D., Ton, L.Q., Topalis, P., Tu, Z., Unger, M.F., Walenz, B., Wang, A., Wang, J., Wang, M., Wang, X., Woodford, K.J., Wortman, J.R., Wu, M., Yao, A., Zdobnov, E.M., Zhang, H., Zhao, Q., Zhao, S., Zhu, S.C., Zhimulev, I., Coluzzi, M., della Torre, A., Roth, C.W., Louis, C., Kalush, F., Mural, R.J., Myers, E.W., Adams, M.D., Smith, H.O., Broder, S., Gardner, M.J., Fraser, C.M., Birney, E., Bork, P., Brey, P.T., Venter, J.C., Weissenbach, J., Kafatos, F.C., Collins, F.H., Hoffman, S.L., 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129–149.

Honeybee Genome Sequencing Consortium, 2006. Insights into social insects from the genome of the honeybee *Apis mellifera*. Nature 443, 931–949.

Karouzou, M.V., Spyropoulos, Y., Iconomidou, V.A., Cornman, R.S., Hamodrakas, S.J., Willis, J.H., 2007. *Drosophila* cuticular proteins with the R&R consensus annotation and classification with a new tool for discriminating RR-1 and RR-2 sequences. Insect Biochem. Mol. Biol. 37, 754–760.

Katsuma, S., Kawaoka, S., Mita, K., Shimada, T. Genome-wide survey for baculoviral host homologs using the *Bombyx* genome sequence. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.05.008.

Kawaoka, S., Hayashi, N., Katsuma, S., Kishino, H., Kohara, Y., Mita, K., Shimada, T. *Bombyx* small RNAs: genomic defense system against transposons in the silkworm, *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.03.007.

Konno, K., Ono, H., Nakamura, M., Tateishi, K., Hirayama, C., Tamura, Y., Hattori, M., Koyama, A., Kohno, K., 2006. Mulberry latex rich in antidiabetic sugar-mimic alkaloids forces dieting on caterpillars. Proc. Natl. Acad. Sci. U.S.A. 31, 1337–1341.

Li, H,, Liu, J., Xu, Z., 2005a. Test data sets and evaluation of gene prediction programs on the rice genome. J. Computer Sci. Technol. 10, 446–453.

Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G.K., Wang, J., 2005b. ReAS: recovery of ancestral sequences for transposable elements from the unassembled reads of a whole genome shotgun. PLoS Comput. Biol. 1, e43.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

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.

Mita, K., Morimyo, M., Okano, K., Koike, Y., Nohata, J., Kawasaki, H., Kadono-Okuda, K., Yamamoto, K., Suzuki, M.G., Shimada, T., Goldsmith, M.R., Maeda, S., 2003. The construction of an EST database for *Bombyx mori* and its application. Proc. Natl. Acad. Sci. U.S.A. 100, 14121–14126.

Nene, V., Wortman, J.R., Lawson, D., Haas, B., Kodira, C., Tu, Z.J., Loftus, B., Xi, Z., Megy, K., Grabherr, M., Ren, Q., Zdobnov, E.M., Lobo, N.F., Campbell, K.S., Brown, S.E., Bonaldo, M.F., Zhu, J., Sinkins, S.P., Hogenkamp, D.G., Amedeo, P., Arensburger, P., Atkinson, P.W., Bidwell, S., Biedler, J., Birney, E., Bruggner, R.V., Costas, J., Coy, M.R., Crabtree, J., Crawford, M., Debruyn, B., Decaprio, D., Eiglmeier, K., Eisenstadt, E., El-Dorry, H., Gelbart, W.M., Gomes, S.L., Hammond, M., Hannick, L.I., Hogan, J.R., Holmes, M.H., Jaffe, D., Johnston, J.S., Kennedy, R.C., Koo, H., Kravitz, S., Kriventseva, E.V., Kulp, D., Labutti, K., Lee, E., Li, S., Lovin, D.D., Mao, C., Mauceli, E., Menck, C.F., Miller, J.R., Montgomery, P., Mori, A., Nascimento, A.L., Naveira, H.F., Nusbaum, C., O'leary, S., Orvis, J.,

Pertea, M., Quesneville, H., Reidenbach, K.R., Rogers, Y.H., Roth, C.W., Schneider, J.R., Schatz, M., Shumway, M., Stanke, M., Stinson, E.O., Tubio, J.M., Vanzee, J.P., Verjovski-Almeida, S., Werner, D., White, O., Wyder, S., Zeng, Q., Zhao, Q., Zhao, Y., Hill, C.A., Raikhel, A.S., Soares, M.B., Knudson, D.L., Lee, N.H., Galagan, J., Salzberg, S.L., Paulsen, I.T., Dimopoulos, G., Collins, F.H., Birren, B., Fraser-Liggett, C.M., Severson, D.W., 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 318, 1718–1723.

Okamoto, S., Futahashi, R., Kojima, T., Mita, K., Fujiwara, H., 2008. A catalogue of epidermal genes: genes expressed in the epidermis during larval molt of the silkworm *Bombyx mori*. BMC Genomics 9, 396.

Ono, Y., Fujibuchi, W., Suwa, M., 2005. Automatic gene collection system for genome-scale overview of G-protein coupled receptors in eukaryote. Gene 364, 63–73.

Osanai-Futahashi, M., Suetsugu, Y., Mita, K., Fujiwara, H. Genome-wide screening and characterization of transposable elements and their distribution analysis in the silkworm. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.05.012.

Roller, L., Yamanaka, N., Watanabe, K., Daubnerová, I., Žitňan, D., Kataoka, H., Tanaka, Y., 2008. The unique evolution of neuropeptide genes in the silkworm *Bombyx mori*. Insect Biochem. Mol. Biol., in this issue.

Robertson, H.M., Wanner, K.W., 2006. The chemoreceptor superfamily in the honey bee, *Apis mellifera*: expansion of the odorant, but not gustatory, receptor family. Genome Res. 16, 1395–1403.

Robertson, H.M., Warr, C.G., Carlson, J.R., 2003. Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. 100, 14537–14542.

Shinoda, T., Itoyama, K., 2003. Juvenile hormone acid methyltransferase: a key regulatory enzyme for insect metamorphosis. Proc. Natl. Acad. Sci. U.S.A. 100, 11986–11991.

Takasu, Y., Yamada, H., Tamura, T., Sezutsu, H., Mita, K., Tsubouchi, K., 2007. Identification and characterization of a novel sericin gene expressed in the anterior middle silk gland of the silkworm *Bombyx mori*. Insect Biochem. Mol. Biol. 37, 1234–1240.

Tamura, T., Thibert, C., Royer, C., Kanda, T., Abraham, E., Kamba, M., Komoto, N., Thomas, J.L., Mauchamp, B., Chavancy, G., Shirk, P., Fraser, M., Prudhomme, J.C., Couble, P., 2000. Germline transformation of the silkworm *Bombyx mori* L. Nature Biotechnol. 18, 81–84.

Tanaka, H., Ishibashi, J., Fujita, K., Nakajima, Y., Sagisaka, A., Tomimoto, K., Suzuki, N., Yoshiyama, M., Kaneko, Y., Iwasaki, T., Sunagawa, T., Yamaji, K., Asaoka, A., Mita, K., Yamakawa, M. A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.09.001.

Tomita, M., Munetsuna, H., Sato, T., Adachi, T., Hino, R., Hayashi, M., Shimizu, K., Nakamura, N., Tamura, T., Yoshizato, K., 2003. Transgenic silkworms produce recombinant human type III procollagen in cocoons. Nature Biotechnol. 21, 52–56.

Tribolium Genome Sequence Consortium, 2008. The genome of the model beetles and pest *Tribolium castaneum*. Nature 452, 949–955.

Uchino, K., Sezutsu, H., Imamura, M., Kobayashi, I., Tatematsu, K., Iizuka, T., Yonemura, N., Mita, Tamura, T. Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.09.009.

Underwood, D.C., Knickerbocker, H., Gardner, G., Condriffe, D.P., Sprague, K.U., 1998. Silk gland-specific tRNA-Ala genes are tightly clustered in the silkworm genome. Mol Cell Biol. 8, 5504–5512.

Wang, J., Wong, G.K., Ni, P., Han, Y., Huang, X., Zhang, J., Ye, C., Zhang, Y., Hu, J., Zhang, K., Xu, X., Cong, L., Lu, H., Ren, X., Ren, X., He, J., Tao, L., Passey, D.A., Wang, J., Yang, H., Yu, J., Li, S., 2002. RePS: a sequence assembler that masks exact repeats identified from the shotgun data. Genome Res. 12, 824–831.

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.

Willis, J.H., Iconomidou, V.A., Smith, R.F., Hamodrasks, S.J., 2005. Cuticular proteins. In: Gilbert, L.I., Iatrou, K., Gill, S.S. (Eds.), Comparative Molecular Insect Science, vol. 4. Elsevier, Oxford, UK, pp. 79–110.

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., Yang, H., 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306, 1937–1940.

Yamamoto, K., Nohata, J., Kadono-Okuda, K., Narukawa, J., Sasanuma, M., Sasanuma, S., Minami, H., Shimomura, M., Suetsugu, Y., Osoegawa, K., de Jong, P.J., Goldsmith, M.R., Mita, K., 2008. A BAC-based integrated linkage map of the silkworm *Bombyx mori*. Genome Biol. 9, R21.

Yu, Q., Lu, C., Li, B., Fang, S., Zuo, W., Dai, P., Zhang, Z., Xianga, Z. Identification, genomic organization and expression pattern of glutathione S-transferase in the silkworm. Insect Biochem. Mol. Biol. doi:10.1016/j.ibmb.2008.08.002.

Zhou, C.Z., Confalonieri, F., Medina, N., Zivanovic, Y., Esnault, C., Yang, T., Jacquet, M., Janin, J., Duguet, M., Perasso, R., Li, Z.G., 2000. Fine organization of *Bombyx mori* fibroin heavy chain gene. Nucleic Acids Res. 28, 2413–2419.

Please cite this article in press as: The International Silkworm Genome Consortium, The genome of a lepidopteran model insect, the silkworm *Bombyx mori*, Insect Biochemistry and Molecular Biology (2009), doi:10.1016/j.ibmb.2008.11.004