Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, *Bombyx mori*.


*National Institute of Agrobiological Sciences, Tsukuba 305-8634, Japan; **National Institute of Advanced Industrial Science and Technology, Tsukuba 305-8566, Japan; #Mitsubishi Space Software Co., Ltd., Tsukuba 305-8602, Japan; **Department of Integrated Biosciences, Graduate School of Frontier Sciences, The Univ. of Tokyo, Kashiwa 277-8562, Japan; $$$Faculty of Life and Environmental Sciences, University of Tsukuba 305-8572, Japan; ++Department of Agricultural and Environmental Biology, Graduate School of Agricultural and Life Sciences, The University of Tokyo 113-8657, Japan; ###Department of Computational Biology, Graduate School of Frontier Sciences, The Univ. of Tokyo, Kashiwa 277-0882, Japan; ***Human Genome Center, Institute of Medical Science, The Univ. of Tokyo, Tokyo 108-8639, Japan; +Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500001, India; $$S$$S$$The University of Rhode Island, Kingston, Rhode Island 02881, USA; +++Guangdong Provincial Key Laboratory of Biotechnology for Plant Development, School of Life Sciences, School of Life Science, South China Normal University, Guangzhou 510631, China; ###State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400716, China

This work is dedicated to the late Dr. J. Nagaraju.
2Y. Sue. and R. F. contributed equally to this work.

3To whom correspondence should be addressed:

Kazuei Mita
State Key Laboratory of Silkworm Genome Biology, Southwest University, Chongqing 400716, China
Tel: +86-23-6825-1892
E-mail: mitakazuei@gmail.com

Kallare P Arunkumar
Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad 500001, India
Tel: +91-40-2474 9342
E-mail: arun@cdfd.org.in

Marian R. Goldsmith
Dept. of Biological Sciences, The University of Rhode Island, Kingston, Rhode Island 02881, USA
Tel: 1-401-874-2637
Email: mki101@uri.edu
Running title: 11,000+ Silkworm full-length cDNAs

Keywords: Bombyx mori, large-scale full-length cDNA collection, tissue-specific genes, sexual dimorphism, gene clusters

Abbreviations: UTR = untranslated region; FL-cDNA = full-length complementary DNA; EST = expressed sequence tag; ORF = open reading frame; BLAST = basic local alignment search tool; kb = kilobases; Mb = megabases; QV = quality value; piRNA = PIWI-interacting RNA.

Competing interests: The authors declare no conflict of interest.

Data deposition: All full-length cDNA sequences obtained in this work appear in public databases with acc nos. [DDBJ:AK377185-AK388575]. ESTs used in the paper are deposited in GenBank/EMBL/DDBJ with the accession numbers listed in the cDNA library table found in <http://sgp.dna.affrc.go.jp/EST/page_pub.html>.

Keywords: full-length cDNA, silkworm, tissue-specific gene, gene cluster, sexual dimorphism

Author contributions: TT, HN, MRG, QX, TShim, and KM conceived and designed the project. KK-O, HS, M-GS, TD, TShin, KT, KA, SKat, HF, HK, and KM contributed full-length cDNA libraries. HK, SSa, JN, SSu, and YSuz performed the sequencing. YSue, RF, MA, AJ, NN, MS, MO-F, RN, SKaw, MK, KPA, AT, JN, QF, and KY analyzed the data. RF, YSue, MO-F, SKaw, JN, MRG, TShim, and KM wrote the paper.

E-mail addresses of authors
YSue: suetsugu@affrc.go.jp
RF: ryo-futahashi@aist.go.jp
HK: kanamo@affrc.go.jp
KK-O: kadono@affrc.go.jp
SSa: sasanuma@nias.affrc.go.jp
JN: narukawa@nias.affrc.go.jp
MA: majimura@gmail.com
AJ: joraku@affrc.go.jp
NN: Namiki.Nobukazu@mss.co.jp
MS: Shimomura.Michihiko@mss.co.jp
HS: hsezutsu@affrc.go.jp
MO-F: mfutahashi@nias.affrc.go.jp
M-GS: gakyo@k.u-tokyo.ac.jp
DT: daimontakaaki@affrc.go.jp
TShin: shinoda@affrc.go.jp
KT: taniai@affrc.go.jp
KA: asaoka@affrc.go.jp
RN: nniwa@biol.tsukuba.ac.jp
SKaw: kawaokashinpei@yahoo.co.jp
SKat: katsuma@ss.ab.a.u-tokyo.ac.jp
TT: ttamura@affrc.go.jp
HN: hnada@affrc.go.jp
MK: mkasa@cb.k.u-tokyo.ac.jp
SSu: ssugano@k.u-tokyo.ac.jp
YSuz: ysuzuki@k.u-tokyo.ac.jp
HF: haruh@k.u-tokyo.ac.jp
HK: kataoka@k.u-tokyo.ac.jp
KPA: arun@cdfd.org.in
AT: archanat@cdfd.org.in
JN: jnagaraju@cdfd.org.in
MRG: mki101@uri.edu
QF: qlfeng@scnu.edu.cn
QX: xiaqy@swu.edu.cn
KY: kiya@nias.affrc.go.jp
TShim: shimada@ss.ab.a.u-tokyo.ac.jp
KM: mitakazuei@gmail.com
Establishment of a complete genomic sequence of silkworm, the model species of Lepidoptera, laid a foundation for its functional genomics. A more complete annotation of the genome will benefit functional and comparative studies and accelerate extensive industrial applications for this insect. To realize these goals, we embarked upon a large-scale full-length cDNA collection from 21 full-length cDNA libraries derived from 14 tissues of the domesticated silkworm and performed full-sequencing by primer walking for 11,104 full-length cDNAs. The large average intron size was 1,904 bp, resulting from a high accumulation of transposons. Using gene models predicted by GLEAN and published mRNAs, we identified 16,823 gene loci on the silkworm genome assembly. Orthology analysis of 153 species including 11 insects revealed that among three Lepidoptera including Monarch and Heliconius butterflies, the 403 largest silkworm-specific genes were mainly composed of protective immunity, hormone-related, and characteristic structural proteins. Analysis of testis-/ovary-specific genes revealed distinctive features of sexual dimorphism including depletion of ovary-specific genes on the Z chromosome in contrast to an enrichment of testis-specific genes. Over 40% of genes expressed in specific tissues mapped in tissue-specific chromosomal clusters. The newly obtained FL-cDNA sequences enabled us to annotate the genome of this lepidopteran model insect more accurately, enhancing genomic and functional studies of Lepidoptera and comparative analyses with other insect orders, and yielding new insights into the evolution and organization of lepidopteran-specific genes.
INTRODUCTION

The domesticated silkworm, *Bombyx mori*, is renowned for silk production as well as being a traditional model insect. As the first animal to demonstrate the application of Mendel’s laws (Toyama, 1906), it has served as a subject for genetical, physiological, and developmental studies. A high quality genome sequence “scaffold build2” (The International Silkworm Genome Consortium 2008) combined with comprehensive map information (Yamamoto *et al.* 2008), transcriptome analyses (Mita *et al.* 2003; Xia *et al.* 2007), and transgenic technology (Tamura *et al.* 2000; Uchino *et al.* 2008), has enabled dramatic progress in silkworm studies including positional cloning of mutations (e.g., Ito *et al.* 2008; Liu *et al.* 2010) and functional genomics of other insects, especially Lepidoptera (Beldade *et al.* 2009; Gahan *et al.* 2010; van’t Hof *et al.* 2011), which were integrated into KAIKObase [http://sgp.dna.affrc.go.jp/KAIKObase/](http://sgp.dna.affrc.go.jp/KAIKObase/) and SilkDB [http://silkworm.swu.edu.cn/silkdb/](http://silkworm.swu.edu.cn/silkdb/).

Lepidoptera is the second-largest order of insects and includes many biologically and economically beneficial species as well as some of the most global and agriculturally destructive pests. Information from the *Bombyx* genome has served as a critical reference for studies of other lepidopterans, and recent reports of chromosomal synteny conservation, even between members of different superfamilies (van’t Hof *et al.* 2011; d’Alençon *et al.* 2010; Zhan *et al.* 2011; The Heliconius Genome Consortium 2012), illustrate the value of the *B. mori* genome as a model for Lepidoptera.

A major goal of genome analysis is complete annotation, i.e., a well-curated gene list. For the published assembly, consensus gene sets were predicted by gene finder programs such as BGF (Li *et al.* 2005) and GLEAN (Elsik *et al.* 2007) and auto-annotated in KAIKObase and SilkDB (Shimomura *et al.* 2009; Duan *et al.* 2010). However, the present silkworm genome assembly still contains inevitable gaps caused by enriched repetitive sequences and tightly-clustered paralogous genes (The International Silkworm Genome Consortium 2008) which are the main sources of incorrect gene prediction. Another source of problems for gene prediction is the limited availability of full-length cDNA (FL-cDNA) sequences which are needed to determine precise gene structures and ortholog-paralog relationships for comparative genomics. Distributing bioresources including FL-cDNA and cDNA clones to researchers is another mission of this project.

Here we report isolation and sequencing of 11,104 FL-cDNA clones chosen from
21 FL-cDNA libraries derived from 14 distinct silkworm tissues of different developmental stages. To determine a complete set of silkworm genes, we used 11,104 FL-cDNAs, 408,172 ESTs, 2,089 mRNA sequences registered in public databases, and 14,625 gene models as shown in Figure 1. We estimated the frequency of expression of a standard set of 16,823 genes, which were mapped on scaffolds and covered complete ORFs culled from the FL-cDNA libraries and a large available EST database, and constructed a map of transcriptional expression for tissue-specific genes along the chromosomes. These data revealed distinctive patterns of tissue-specific and sexually dimorphic gene clusters. The extensive dataset we provide here should be valuable not only for studies in the silkworm but also for other insects.

MATERIALS AND METHODS

Full-length cDNA Library Construction
Twenty-one FL-cDNA libraries were constructed from 14 distinct tissues of the domesticated silkworm, *B. mori*, at various developmental stages; details are listed in Table 1. Total RNA was extracted from various tissues dissected at appropriate stages of larvae, pupae or adult moths and treated with TRIzol (Invitrogen) as described previously (Okamoto et al. 2008). FL-cDNA libraries were constructed with the Oligo-cap method performed by Toyobo Co. Ltd (Suzuki et al. 1997), the G-cap method by Hitachi Science Systems, Ltd., or the V-cap method by Hokkaido System Science Co. Ltd (Kato et al. 2005). More than 10,000 cDNA clones were picked up randomly from each cDNA library.

5’ EST Sequencing and Determination of FL-cDNA Clone Candidates
Approximately 250,000 cDNA clones were randomly selected from 21 FL-cDNA libraries for determination of 5’ ESTs using an ABI 3730 DNA sequencer. All 5’ ESTs were grouped into contigs or singletons under the criteria that if 5’ ESTs shared >95% identity over 100 consecutive bases for aligned regions, they were considered as identical clones transcribed from the same gene. The cDNA clone covering the most upstream region in a contig was chosen as a potential FL-cDNA candidate. All singletons were also potential candidates for FL-cDNAs. All candidate clones were then completely sequenced by a primer-walking method using an ABI 3730 DNA sequencer.
Evaluation of FL-cDNA Sequences
Vector sequences were filtered and low-quality bases (QV <20) were removed. The resulting FL-cDNA sequences were aligned with the silkworm genome assembly (The International Silkworm Genome Consortium 2008) to remove chimeric sequences. Perfect alignment with the genome assembly within 100 kb was used as a provisional criterion for absence of chimerism. The FL-cDNA sequences were also subjected to a BLAST search in the silkworm EST database <http://kaikocdna.dna.affrc.go.jp/> to determine whether mate-paired ESTs transcribed from the same gene could be aligned on the FL-cDNA sequence. If not, the FL-cDNA clone was considered likely to be chimeric or a splicing-isoform.

To facilitate the full sequencing of FL-cDNA candidates, the Illumina Genome Analyzer (GA) was employed to obtain draft sequences of FL-cDNA clones greater than 3 kb in size. The FL-cDNA candidates were divided into several sets, each containing 700 clones. Each set was then shotgun-sequenced using one lane of a GA flow cell. The reads were assembled by MuSICA as described in Kuroshu et al. (2010). We designed primer sets to confirm the assembled sequences by the primer-walking method with the ABI3730 sequencer.

Alignment of FL-cDNAs onto the silkworm genome assembly
Genomic positions of FL-cDNAs were determined with est2genome, a software tool to aid the prediction of genes by sequence homology. est2genome is computationally expensive mainly due to usage of the Smith-Waterman algorithm (Smith and Waterman 1981) in return for higher precision. To reduce computation time, all FL-cDNAs were subjected to BLASTN search against silkworm genomic sequences prior to est2genome analysis in order to estimate their approximate genomic positions and filter out FL-cDNAs which were not aligned to a genomic sequence. We used a percent identity of 95% and alignment coverage of 0.5, the ratio of aligned length to total FL-cDNA length, as filtration criteria. If an FL-cDNA did not meet these criteria, it was classified as “unmapped” and discarded from further processing. We then cut out a subsequence of each FL-cDNA from a genomic sequence around the estimated genomic position for analysis by est2genome.
Construction of an FL-cDNA Database

Complete information for the FL-cDNA sequence data set was compiled into a database <http://sgp.dna.affrc.go.jp/FLcDNA/> with the following characteristics: clone name, full sequence, total size (bp), accession number, mapped scaffold, mapped chromosome number, chromosomal ORF location (start and stop codon positions), BLAST results in protein databases, and GO and InterProScan terms, followed by identical EST numbers in each cDNA library.

Expression Profiles of FL-cDNA Clones

Identical ESTs transcribed from each gene were collected by BLASTn search using criteria of >95% identity in a sequence >100 nucleotides in each cDNA library. File S1 lists the frequency of identical ESTs in each tissue which includes not only FL-cDNA libraries but also cDNA libraries constructed by standard methods (Mita et al. 2003), yielding an expression profile of each FL-cDNA clone. Each FL-cDNA sequence was checked for identical ESTs in a deep database which consisted of ESTs from 21 FL-cDNA libraries, 36 previously published cDNA libraries (Mita et al. 2003), and 12 newly analyzed cDNA libraries <http://sgp.dna.affrc.go.jp/EST/page_pub.html>, 408,172 ESTs in all. Using these expression profiles, tissue-specific gene expression patterns were identified for clones present in more than 3 copies with more than 90% tissue-specificity and compiled in a “silkworm gene set” table (Table S1).

RESULTS AND DISCUSSION

Characteristics of FL-cDNA Sequences.

A total of 248,130 5’ ESTs comprised of at least 100 consecutive nucleotides with QV>20 were obtained from 21 FL-cDNA libraries derived from 14 distinct tissues at different developmental stages of B. mori (Table 1) and assembled into 16,823 unique sequences (7,124 singletons and 9,703 contigs). The cDNA clone that covered the most upstream region of each contig was used as a representative of each group, i.e., an FL-cDNA candidate clone, and was fully sequenced by the primer-walking method. We identified 11,104 FL-cDNA sequences (acc nos. AK377185-AK388575 in GenBank/EMBL/DDBJ), of which 438 could not be aligned on the silkworm genome
assembly (Table 2). The mean length of the aligned FL-cDNAs was 1,813 bp, and 10,838 cDNAs had open reading frames longer than 30 amino acids (Table 2). Ten thousand six hundred and sixty six sequences (96.1%) were aligned at 9,315 transcription sites (TSs) using the est2genome program (Mott 1997) in the EMBOSS (The European Molecular Biology Open Software Suite) package (Rice et al. 2000).

The whole data set including improved annotations was compiled into an FL-cDNA database <http://sgp.dna.affrc.go.jp/FLcDNA/>.

By a comparison of 10,666 mapped FL-cDNA gene structures with previously reported silkworm gene prediction models (The International Silkworm Genome Consortium 2008) we found that 7,504 FL-cDNAs matched with models, while 3,162 (30%) showed no match. Among matched FL-cDNAs, our comparison showed that 1,666 FL-cDNAs provided complete matches; however, 5,059 structures comprising about three-fourths of the predicted genes were misannotated. As a comprehensive silkworm gene set (Table S1), previous annotations were updated by employing available FL-cDNA data instead of predicted gene models and previous predicted gene models were used if there was no transcriptome data. By pair-wise comparison of transcripts mapped to a given locus, 2,072 FL-cDNAs appeared to be derived from alternative splicing. The mean exon number per gene was 4.8, and mean exon and intron sizes were 353 bp and 1,904 bp, respectively (Table 3). Comparison of these values relative to the genome size of eleven other model species showed a good correlation of the intron size with the genome size (Figure 2; R=0.942; p<0.001), indicating that the large introns of silkworm may have contributed to its relatively large genome size. We also compared the intron-genome size with the fraction of transposable elements in each genome (Figure 2), resulting in a rough correlation between genome size and TE content (R=0.558; p<0.059). The deviation in this value was considerably larger than the ratio of the mean intron size:genome size, which suggests that the large introns in the silkworm genome may have arisen in part from a high accumulation of repetitive sequences, mainly composed of transposons (Osanai-Futahashi et al. 2008). In contrast to the intron size, the average exon size had almost no correlation with the genome size (R=-0.487; p=0.109), indicating there is very little variation in average exon length among the species examined.

Obtaining a comprehensive silkworm gene set. To obtain a complete silkworm gene set,
in addition to 11,104 FL-cDNA sequences the available set of 408,172 ESTs, 2,089 mRNA sequences in public databases, and 14,623 gene models predicted by GLEAN (Elsik et al. 2007) were also mapped and grouped. An overview of the procedures and results of this analysis is presented in Figure 1. FL-cDNA sequences, mRNA sequences in public databases, and 370,472 ESTs which could be aligned to FL-cDNAs or mRNAs in public databases and gene models were grouped into 16,823 gene sites (Gene set A in Figure 1); these are compiled in Table S1A with gene IDs and used as the ‘silkworm gene set’. In addition, among 37,700 ESTs that did not align to Gene set A, 27,102 ESTs could be mapped to the genome and were assembled into 7,240 groups which are referred to as “EST-based genes” (Gene set B in Figure 1, sequences of which are available at <http://sgp.dna.affrc.go.jp/FLcDNA/>). It should be noted that Gene set B includes ESTs derived from FL-cDNA libraries for which the full length clone sequences have not yet been determined. The Gene IDs for EST-based genes are prefixed with ‘e’ in Table S1B. Among them 2,268 sequences had open reading frames and were not annotated as TEs or viruses, but may still be missing from Gene set A of the silkworm genome assembly. Gene set C is composed of the remaining 10,598 ESTs, 60 publicly available mRNAs and 438 FL-cDNAs that were not aligned to the silkworm genome assembly. We grouped them into 6,160 genes using CLOBB2 (http://www.nematodes.org/bioinformatics/CLOBB2/) (Table S1C) whose sequences can be found at <http://sgp.dna.affrc.go.jp/FLcDNA/>). BLASTx search against the NCBI nonredundant database using a cutoff of 1e-10 was used to classify genes, indicating 2,710 as "no hit" and 2,612 genes as "other". Of these 5,322 genes, 3,432 had open reading frames which may have arisen from gap regions of the silkworm genome assembly. It should be noted that the silkworm genome assembly did not include the W chromosome due to its extremely high content of repetitive DNAs (Abe et al. 2005; The International Silkworm Genome Consortium 2008); however, transcripts derived from W might be included in cDNA libraries, which would be grouped in Gene set C. Taken together, we consider that the total number of silkworm genes may reach more than 20,000 (e.g., 16,823 + 2,268 + 3,432 = 22,523).

Silkworm FL-cDNA Similarity Comparison with Other Insect Orders.
Comparison of silkworm FL-cDNAs with gene sets of other insect orders with complete genome sequences provides an opportunity for exploring silkworm and
Lepidoptera-specific genes. We compared the silkworm gene set with the gene sets of *Drosophila melanogaster* (Diptera), *Anopheles gambiae* (Diptera), *Aedes aegypti* (Diptera), *Culex quinquefasciatus* (Diptera), *Apis mellifera* (Hymenoptera), *Tribolium castaneum* (Coleoptera), *Acrithosiphon pisum* (Homoptera), *Pediculus humanus* (Phthiraptera), *Danaus plexippus* (Lepidoptera) and *Heliconius melpomene* (Lepidoptera), 11 insect species of six orders in all (Table 4). For orthology analysis, in addition to the 11 insect species we used protein data sets of 142 non-insect species available in OrthoMCL DB (http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&type=summary).

Orthology analysis of protein data from 153 species including 11 insects using the program orthoMCL (Li et al. 2003) generated 12,637 ortholog groups, 4,480 of which were designated as insect-shared, while 2,803 were Lepidoptera-specific (Figure 3A). Figure 3B presents the classification of Lepidoptera-specific ortholog groups among three species. Silkworm showed the largest number of 403 silkworm-specific (Bm) ortholog groups compared with 49 assigned as Monarch-specific (Dp) and 102 as Heliconius-specific (Hm); the classification of Lepidoptera-specific genes is shown in Figure 3C. In order to check for the presence of paralogs in each group, we compared the number of genes assigned to each ortholog group among the three Lepidoptera (Table 5). It is noteworthy that although the number of “Insect-shared” genes and ortholog groups was almost the same among the three lepidopteran species, with relatively few paralogs (average ratio of 1.58 for #genes/#orthologs), the increase in number of “Lepidoptera-specific genes” and “Species-specific genes” was comparatively high. This might reflect an enhancement of species-specific proteins related to characteristic functions or traits which have evolved through gene duplication events such as the large numbers of phototransduction-related genes in the monarch butterfly putatively involved in sensing skylight cues (Zhan et al. 2011), and chemosensory genes in the monarch and Heliconius butterflies used for host-recognition in feeding and oviposition (Zhan et al. 2011; The Heliconius Genome Consortium 2012). To our surprise, by our criteria silkworm had approximately 8-13-fold more silkworm-specific genes and an approximately 1.5-fold expansion of paralogs among the Lepidoptera-specific group compared with the other two species. These characteristics evoke a distinctive aspect of silkworm, its domestication and complete dependence on human care. Although we are still not clear what produced this
phenomenon, a possible contributing factor may be an acceleration of evolution under artificial selection, which was reported to result in a significant diversity of carotenoid-binding protein genes responsible for cocoon color in the silkworm genome (Sakudoh et al. 2011). A complete genome comparison with B. mandarina, the wild ancestor of B. mori, will be of interest to test this hypothesis.

In order to understand the function of silkworm-specific proteins, we assigned 531 silkworm-specific genes with InterProScan and GO-terms (Table S2) among which 147 groups were annotated with GO-terms. We found a high abundance of genes associated with terms for protective immunity against microbial and viral pathogens such as Moricin, Cecropin, Serpin, Lipoprotein, Glycoside hydrolase, and Guanylate-binding (Tanaka et al. 2008), pheromone/hormone-related functions such as fatty acid CoA reductase, Acyl CoA transferase, JH binding, and carboxylesterase, and characteristic structural proteins such as chorion and cuticle. The fact that protective immunity-related genes were enriched in the silkworm-specific group together with expansion of those genes may again reflect artificial selection for disease-resistant and bacteria-resistant strains during domestication.

**Tissue-Specifically Expressed Genes.**

Tissue-specific genes are of interest because their expression results in tissue-specific functions or traits. To identify them, each FL-cDNA sequence was checked for identical ESTs in the deep EST database and the number of identical ESTs for each FL-cDNA sequence across various tissues constituted its expression profile. If an FL-cDNA sequence with more than 3 identical EST clones had more than 90% identical ESTs derived from only one tissue in the EST database (i.e., fewer than 10% of ESTs were derived from other tissues) we regarded the corresponding gene as tissue-specific. Using these criteria, we identified the tissue-specific genes by automated annotation; the integrated information is presented in Table S1. Table S3 summarizes the chromosomal distribution of tissue-specific genes in each tissue based on the genome alignment.

*Comparison of gonad-specific genes reveals obvious sexual dimorphism.* Silkworm has a female heterogametic sex chromosome system with ZZ in males and ZW in females (Goldsmith et al., 2005; Tanaka, 1916). An obvious difference in tissue-specific genes between ovary and testis was greater than 10-fold more testis-specific genes, 745, than
the 68 ovary-specific genes; this is consistent with the observation that spermatogenesis is a much more complicated structural and developmental process than oogenesis. Similar observations have been made using EST microarray analysis (Xia et al. 2007). Mapping of testis-specific FL-cDNA sequences onto chromosomes yielded the same conclusions as previous work indicating that ch.Z (=ch.1) is enriched in testis-specific genes (Arunkumar et al. 2009). In contrast to testis-specific genes, the spatial distribution of ovary-specific genes provided a highly distinctive feature (Figure 4A). Seventy-four percent of the ovary-specific genes formed apparent gene clusters on only 4 chromosomes, ch.2, 10, 15 and 16 (Figure 4B). Among them, the largest was the chorion gene cluster ch.2 in the region 1.78 – 3.79 Mb. We identified 27 chorion genes in this locus of the genome assembly (The International Silkworm Genome Consortium 2008); however, it has been reported that the chorion locus harbors more than 100 chorion genes (Eickbush and Izzo 2005). This discrepancy was probably caused by a tightly linked cluster of similar sequences, which made a large gap in the automated genome assembly. Just upstream of the chorion gene cluster, 12 ovary-specific genes (Gene000814, Gene000815, Gene000816, Gene000817, Gene000818, Gene000819, Gene000820, Gene000821, Gene000824, Gene000825, Gene000827 and Gene000828) with unknown function formed a cluster in a 100kb region at 1.67 – 1.78 Mb. None of them showed significant homology with sequences in public protein databases, although their presence in the chorion region suggests they may contribute to oogenesis. On ch.10, four ovary-specific genes (Gene005491, Gene005493, Gene005494 and Gene005495) with unknown function formed a cluster in a 16 kb region. Eight ovary-specific genes, of which 2 genes seemed to encode an extensin 2-like protein and others were of unknown function, were located in a 20 kb region of ch.15 forming a tight gene cluster. In addition, an 85 kb region of ch.16 harbored six ovary-specific genes of unknown function. It is noteworthy that more than half of the ovary-specific genes were concentrated on ch.2 within a very narrow region of 2.1 Mb, which made a strong contrast with the dispersed distribution of testis-specific genes enriched on ch.Z. The testis-specific gene density on ch.Z was found to be 1.6 times higher than the average testis-specific gene density (Table S3), similar to the value of 1.75 times enrichment of testis-specific genes on ch.Z reported previously by Arunkumar et al. (2009). In addition, most other ovary-specific genes like the chorion genes formed bigger clustered gene families with overlapping functions than testis-specific genes.
This distinctive nature might be crucial for oogenesis whereby oocytes which are very large compared to spermatocytes must complete development in a relatively short time.

Superimposing the chromosomal distribution images of testis-specific and ovary-specific genes revealed dramatic features of sexual dimorphism (Figure 4A). Ch.Z, where testis-specific genes were enriched, completely lacked ovary-specific genes, whereas testis-specific genes were entirely missing from the ovary-specific gene cluster region of ch.2:1.67–3.79 Mb. Testis-specific genes were also lacking or exclusively depleted in the other ovary-specific gene cluster regions on Chs. 10, 15 and 16. The ovary-specific gene regions were well separated from the testis-specific gene-enriched regions, which may enable efficient gene expression in each type of gonad. For example, ovary-specific gene cluster regions may employ a euchromatic structure leading to efficient gene expression in ovaries, whereas testis-specific genes may be localized in heterochromatic domains for effective repression of expression in ovaries, and vice versa in testes. Analyses of PIWI-interacting RNAs (piRNAs) on ovary-specific regions support the establishment of large euchromatic domains for their expression in ovary (File S1; Figure S1).

The osiris gene cluster conserves a characteristic structure and expression profile across insects. Osiris genes are highly conserved and clustered in insects; however, their function is still unknown (Dorer et al. 2003; Shah et al. 2012). Interestingly, we found several wing-specific osiris genes in B. mori which were clustered on ch.26. Among them we could identify several Drosophila osiris homologs, although some of them were missing from the silkworm genome, including Osi1, Osi4-6, Osi13-15 and Osi23. It was reported that Drosophila osiris genes encode a novel family of transmembrane proteins (Dorer et al. 2003), all of which contain a well-conserved pair of cysteine amino acid residues. We found silkworm osiris proteins had the same structural features using the protein structure prediction programs PSORT (Nakai and Horton 1999) and SOSUI (Hirokawa et al. 1998). Bmosi2 and Bmosi16 provided no evidence for transcription and may be pseudogenes or misannotated. Although we found no silkworm homologs to Drosophila osiris genes 4-6 and 13-15, the remainder of Bmosi3–Bmosi18 were clustered in a single locus of ch.26:11,546,404-11,956,406 in the same order as Drosophila osiris genes 1-20 (Figure 5; Table S4). In addition, 5
copies of *Bmosi*9 were found in a cluster between *Bmosi*7 and *Bmosi*8 in the region ch26:11,684,158-11,776,381. By phylogenetic analysis, these 5 copies were formed by gene duplication events after separation of Lepidoptera from Diptera. The homologs of *Dmosi*21 and 22 were unlinked to the main silkworm *osiris* gene cluster on ch. 26, which was consistent with their independent genetic linkage mapping on ch.4 and ch.12, respectively. These results indicate an explicit microsynteny between silkworm and fruitfly genomes (Figure 5).

*Bmosi*3–*Bmosi*12 showed wing-specific expression except for *osi9-3* which was epidermis-specific and *Bmosi*10 which had no hit in the EST database. *Bmosi*18, which was located at the very end of the cluster, was also not wing-specific. *Bmosi*19 and *Bmosi*20, which were located in a single site near the other end of ch.26 in the reverse order, were also not wing-specific, similar to *Bmosi*18. *Bmosi*3–*Bmosi*12 in the main gene cluster were conserved to have wing-specific expression in the pupal stage, similar to corresponding homologs in *Drosophila* which show peak expression at the late pupal stage. *Bmosi*18–20 had a different expression profile, suggesting that the original *osiris* gene cluster was partitioned into two parts with different timing of gene expression. Subsequently, the *osi9*/*osi20* region was translocated from its position in the original gene cluster near the end of ch.26 to the other end of the same chromosome during silkworm evolution. The *osiris* gene cluster provides another example suggesting that the same tissue-specific genes may be clustered to form a tissue-specific chromatin domain. The largest cuticular protein gene cluster on ch.22 reinforces this observation (File S1; Figure S2).

**Characteristic expression profile of a 30kDa lipoprotein gene cluster correlates with a distinct function.** Biosynthesis of the Lepidoptera-specific 30kDa lipoprotein gene family whose function is not fully understood occurs in a stage-dependent fashion in fat body (Ujita *et al.* 2002). We found thirty-three 30 kDa protein genes localized in an 820kb region of ch.20 forming two gene clusters (Figure 6, Table S5). The first gene cluster in the region spanning 3,412,956-3,565,568 harbored nineteen 30kDa protein genes; four of these were fat body-specific and one was malpighian tubule (MT)-specific. Interestingly, whereas three of the 19 genes in the cluster seemed not to express, nine genes were transcribed extensively and primarily in the brain-nervous system. The expression levels of these nine genes were significantly higher than in fat body.
body (Additional file 8), indicating the brain-nervous system is a major contributor of 30kDa protein transcripts. In addition, the four fat body and MT-specific genes were localized in the 5’ portion of the gene cluster, whereas the nine genes expressed mainly in brain-nervous system occupied the other half of the cluster (Figure 6). The second gene cluster spanning the region 4,012,903-4,227,867 was composed of fourteen 30kDa protein genes, nine of which expressed exclusively or mainly at an embryonic stage (Figure 6; Table S5). These observations suggest an as yet unknown function for 30kDa proteins which has not been reported previously and further support the idea of tissue-specific chromosome domains.

**Tissue-specific gene clusters**

Mapping of tissue-specific genes onto chromosomes is informative for understanding the relationship between chromosome structure and transcription. As can be clearly seen in ovary (Figure 4B), wing (Figure 5; Figure S2) and fat body (Figure 6), tissue-specific genes showed a tendency to be located close to one another in clusters. In Table 6, we summarize the fraction of tissue-specific genes clustering in each tissue whereby we recognized 2 genes expressed in a given tissue as belonging to a single cluster if they were within 100 kb on the same chromosome. The traditional model describes the higher order structure of interphase chromosomes as formed by a series of loops of about 50-150 kb DNA which are attached to a peripheral lamina or other internal structures, such as scaffolds or skeletons (Cook 1995; Mirshra and Karch 1999; Maeshima and Laemmli 2003). Moreover, large-scale chromatin loops more than 1 Mb have been observed in G1 phase chromosomes of human cells (Yokota et al. 1995) and inducible loci of mammalian cells (Hu et al. 2009). Thus, we employed the criterion of “within 100 kb” for a cluster search of 1,365 tissue-specific genes on the 475 Mb silkworm genome. Table 6 shows that, on average, 42% of tissue-specific genes were in clusters, although the fraction varied from 0.15 to 0.78 among tissues. We checked whether the clustering of tissue-specific genes indeed occurs in silkworm by invoking Fisher's Exact Test (Routlege 2005) to test the null hypothesis for independence of the type of gene and the type of genes adjacent to it. The test rejected the null hypothesis (P-value=2.2e-16), indicating that tissue-specific genes have a tendency of forming clusters (Table S6).

In order to consider the driving factors for clustering, it is highly informative to
check the fraction of tissue-specific genes in clusters formed by gene duplication events, since gene duplication will overwhelmingly produce new gene copies in very close proximity to the original gene locus which are likely to retain their tissue specificity unless they evolve new functions. The rates of gene duplication are summarized in Table 6, showing a deep dependence on tissue. For differentiated tissues such as wing, midgut, malpighian tubule, pheromone gland, silk gland, fat body and epidermis, a very high rate of duplication and clustering of tissue-specific genes was observed, which suggests that the respective tissue produces these characteristic proteins or enzymes at high levels. On the contrary, in ovary, only the chorion gene cluster was formed by gene duplication, and the other gene clusters on ch.10, 15, and 16 were composed of different kinds of genes. The same tendency was observed in testis-specific gene clusters, where the rate of gene duplication was considerably less. Even for differentiated tissues such as wing, midgut and epidermis, some of the clustered tissue-specific genes were not derived from gene duplication events.

FL-cDNA clones and sequence data as bioresources
All FL-cDNA sequence data were incorporated into KAIKObase and can be downloaded from its associated GBrowse. The silkworm gene sets in Table S1 combined with the results of orthology analysis were also integrated into KAIKObase. In addition, a comprehensive Silkworm Genome Annotation Project is in progress by the Silkworm Genome Annotation Consortium using the complete FL-cDNA data. All FL-cDNA and standard cDNA clones are maintained at the National Institute of Agrobiological Sciences and the University of Tokyo and are available on request (see KAIKObase). The distribution of the DNA clones is also supported by the National Bioresource Project (NBRP).

Conclusion
The sequences obtained in this study are critical for more accurate genome annotation of this lepidopteran model insect, enhancing genomic and functional studies of Lepidoptera and comparative analyses of B. mori with other insect orders, and yielding new insights into gene evolution and the existence of lepidopteran-specific genes. Additionally, the data presented here including the organization and expression of testis- and ovary-specific genes, osiris, 30K, and cuticular protein genes (File S1; Table 6) show a clustering tendency for tissue-specific genes in the silkworm genome and
extend earlier observations of striking differences in the organization of sexually dimorphic sets of genes.

SUPPORTING INFORMATION

**Figure S1** Density plots of ovarian piRNAs in the ovary-specific gene regions and a representative large ovarian piRNA cluster. Each bar represents the relative abundance of piRNAs that were uniquely mapped to the silkworm genome (unique mappers). (A-D) Ovary-specific gene clusters of ch.2 (A), ch.10 (B), ch.15 (C) and ch.16 (D). Upper line denotes ovarian piRNA level. The positions of genes are shown by vertical lines in the lower line, where ovary-specific genes are colored red. (E) A representative large ovarian piRNA cluster in the region ch.14:2,547,598-2,847,509. (F) Box plots show relative abundance of sense and antisense piRNAs mapped to 47 ovary-specific ESTs or 121 transposons. Relative abundance of piRNAs matching each element is expressed as reads per million per a kilobase (RPKM).

**Figure S2** The largest cuticular protein gene cluster on ch.22. The largest cuticular protein gene cluster is divided into 4 parts. Both ends are comprised of wing-specific genes (red), whereas the two central parts containing more than 40 cuticular protein genes are larval undifferentiated disc-specific (blue) and embryonic stage-specific (green).

**File S1** A text describing the supporting results of piRNA analysis in ovary, defeminized Z chromosome, wing differentiation, silkgland-specific genes, and other tissue-specific genes.

**Table S1** A dataset table of silkworm gene sets presented in Figure 3: (A) Silkworm Gene Set A, (B) Silkworm Gene Set B, and (C) Silkworm Gene Set C. This table also can be downloaded from [http://sgp.dna.affrc.go.jp/ComprehensiveGeneSet/](http://sgp.dna.affrc.go.jp/ComprehensiveGeneSet/).

**Table S2** Annotated ortholog groups of 531 silkworm-specific orthologs with InterProScan and GO-terms.

**Table S3** Summary of the chromosomal distribution of tissue-specific genes in each
tissue.

**Table S4**  Mapping of *Bm osiris* genes.

**Table S5**  30kDa protein genes forming a gene cluster on chromosome 20.

**Table S6**  Result of Fisher's Exact Test to evaluate the association between type of a gene and type of its adjacent genes.

**ACKNOWLEDGEMENTS**

This work was supported by the Project for Insect Technology of the Ministry of Agriculture, Forestry and Fisheries of Japan. The cDNA clones were partly provided by the National Bioresource Project, Ministry of Education, Culture, Sports, Science and Technology, Japan.

**LITERATURE CITED**


Dorer, D. R., J. A. Rudnik, E. N. Moriyama, and A. C. Christensen, 2003 A family of genes clustered at the Triplo-lethal locus of Drosophila melanogaster has an unusual evolutionary history and significant synteny with Anopheles gambiae. Genetics 165:613-621.


Gahan, L. J., Y. Pauchet, H. Vogel, and D. G. Heckel, 2010 An ABC transporter mutation is correlated with insect resistance to Bacillus thuringiensis Cry1Ac toxin.


FIGURE LEGENDS

Figure 1
Overview of the gene-build for the silkworm genome. The white box in each gene set describes the composition of ESTs derived from FL-cDNA and cDNA libraries. ESTs used in Gene set B include ones derived from FL-cDNA libraries for which the full length clone sequences were not yet determined.

Figure 2
Relationships between genome size and average intron/exon lengths for various model species. To calculate the average intron and exon lengths for 11 species other than B. mori (D. melanogaster, C. elegans, M. musculus, H. sapiens, D. plexippus, H. melpomene, G. gallus, A. aegypti, Acyystisiphon pism, Strongylocentrotus purpuratus and Danio rerio), the GTF-formatted gene annotation files were downloaded from the Ensembl ftp site ftp://ftp.ensembl.org and processed with a custom Perl script. Triangles denote the TE content.

Figure 3
Orthology of 153 species including 11 insect species. (A) Pie chart: Classification of 12,637 ortholog groups generated from protein datasets of 153 species including 11 insect species listed in Table 4. 1:1 indicates universal single-copy genes. N:N indicates universal genes but with paralogs in at least one species. Numbers in the circles denote number of ortholog groups. (B) Bar chart: Orthology assignment of 2,803 Lepidoptera
specific ortholog groups. ■: Lepidoptera specific (1:1); □: Lepidoptera specific (N:N); ■: Bm-Dp specific; □: Bm-Hm specific; □: Dp-Hm specific; ■: Bm specific; ■: Dp specific; □: Hm specific. (C) Classification of Lepidoptera-specific genes in three species. ■: Not involved in ortholog groups; ■: Common to 11 insect species (1:1); ■: Common to 11 insect species (N:N); ■: Common to 6 insect orders but not shared with all 11 species; □: Others; ■: Lepidoptera specific (1:1); ■: Lepidoptera specific (N:N); ■: Species specific; □: Bm-Dp specific for Bm, Dp-Bm specific for Dp, Hm-Bm specific for Hm; ■: Bm-Hm specific for Bm, Dp-Hm specific for Dp, Hm-Dp specific for Hm. Numbers in the boxes denote number of genes. Bm: Bombyx mori, Dp: Danaus plexippus, Hm: Heliconius melpomene.

**Figure 4**

Sexual dimorphism in mapping of ovary-/testis-specific genes. (A) Chromosomal distribution of testis-specific (blue) and ovary-specific (red) genes. The ovary-specific gene clusters on ch.2, 10, 15 and 16 are circled in green and are presented with enlarged views in Fig. 4B. (B) Ovary-specific gene clusters on ch.2, 10, 15 and 16. Red bar, blue bar and black bar denote ovary-specific gene, testis-specific gene and non-tissue-specific gene, respectively.

**Figure 5**

Synteny of the osiris gene cluster between silkworm and fruitfly. Blue, osiris genes transcribed in several silkworm tissues; red, osiris genes transcribed only in wing; black, osiris genes with no evidence for transcription. Silkworm osiris genes are numbered according to the Drosophila gene set; (+) and (–) denote the direction of gene.

**Figure 6**

Map of thirty-three 30kDa protein genes forming a cluster on chromosome 20. In the first gene cluster composed of nineteen 30kDa protein genes, four fat body-specific genes and one malpighian tubule-specific 30kDa protein gene (red) are located in the former half of the cluster, whereas nine 30kDa protein genes mainly expressed in brain-nervous system (blue) occupy the latter half of the cluster. Six genes showed no hit in the EST database (gray). A second 30kDa protein gene cluster composed of fourteen genes was located 440kb from the former gene cluster. Nine of fourteen genes
were specifically expressed in an embryonic stage (green), whereas one was mainly expressed in fat body; the remaining four genes showed no hit in the EST db.
**Table 1** Twenty-one full-length cDNA libraries analyzed.

<table>
<thead>
<tr>
<th>Name of Full-length cDNA library</th>
<th>No. of clones</th>
<th>Vector</th>
<th>Method (-cap)</th>
</tr>
</thead>
<tbody>
<tr>
<td>famL (antenna &amp; maxill. gl., larva)</td>
<td>14,044</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>fner (nerve system + brain, larva)</td>
<td>18,772</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>fcaL (corpora allata-cardiaca, larva)</td>
<td>17,222</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fmXG (maxillary galea, larva)</td>
<td>4,366</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fwd (wing disc)</td>
<td>11,515</td>
<td>pCMVFL</td>
<td>Oligo</td>
</tr>
<tr>
<td>fwgP (wing, pupa day2-8)</td>
<td>17,521</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>ftes (testis, 5th larva)</td>
<td>16,487</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>bmte (testis, pupa day 4)</td>
<td>10,367</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>bmov (ovary, pupa day 4)</td>
<td>10,375</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>fmGV (midgut, 5th larva)</td>
<td>19,015</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>bmmt (malpighian tubule, larva)</td>
<td>10,104</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>fphe (pheromone gland, adult)</td>
<td>7,600</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>ffbm (male fat body, larva)</td>
<td>18,380</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>MFB (fat body, microbe-infected)</td>
<td>5,846</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fem (epidermis 4th molt)</td>
<td>6,588</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fprW (prothoracic gland, W-stage)</td>
<td>6,000</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fufe (unfertilized egg)</td>
<td>17,095</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>fdpe (diapaused egg)</td>
<td>5,714</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>e100 (embryo 100h)</td>
<td>5,025</td>
<td>pGCAP1</td>
<td>G-cap</td>
</tr>
<tr>
<td>fe8d (embryo day 8)</td>
<td>17,551</td>
<td>pGCAP10</td>
<td>V-cap</td>
</tr>
<tr>
<td>BmN (cultured cell)</td>
<td>8,543</td>
<td>pCMVFL</td>
<td>Oligo</td>
</tr>
</tbody>
</table>

Total: 21 FL-cDNA libraries        Total: 248,130 clones
**Table 2**  Basic statistics of *B. mori* full-length cDNAs.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of FL-cDNAs</strong></td>
<td>11,104</td>
</tr>
<tr>
<td><strong>Mean length (bp)</strong></td>
<td>1,813</td>
</tr>
<tr>
<td><strong>GC content</strong></td>
<td>0.38</td>
</tr>
<tr>
<td><strong>Maximum length (bp)</strong></td>
<td>10,430</td>
</tr>
<tr>
<td><strong>Minimum length (bp)</strong></td>
<td>141</td>
</tr>
<tr>
<td><strong>FL-cDNAs with CDS</strong></td>
<td>10,838</td>
</tr>
<tr>
<td><strong>FL-cDNAs with poly-A</strong></td>
<td>8,789</td>
</tr>
<tr>
<td><strong>FL-cDNAs with CDS and poly-A</strong></td>
<td>8,789</td>
</tr>
<tr>
<td><strong>Mapped FL-cDNAs</strong></td>
<td></td>
</tr>
<tr>
<td>mapped onto chromosomes</td>
<td>10,075</td>
</tr>
<tr>
<td>ORF +/-</td>
<td>9,840/235</td>
</tr>
<tr>
<td>Poly-A +/-</td>
<td>8,191/1,884</td>
</tr>
<tr>
<td>Scaffolds (unmapped on chromosomes)</td>
<td>591</td>
</tr>
<tr>
<td>ORF +/-</td>
<td>569/22</td>
</tr>
<tr>
<td>Poly-A +/-</td>
<td>485/106</td>
</tr>
<tr>
<td>Mapped total</td>
<td>10,666</td>
</tr>
<tr>
<td>ORF +/-</td>
<td>10,409/257</td>
</tr>
<tr>
<td>Poly-A +/-</td>
<td>8,676/1,990</td>
</tr>
<tr>
<td><strong>Unmapped FL-cDNAs</strong></td>
<td>438</td>
</tr>
<tr>
<td><strong>Low coverage</strong></td>
<td>380 (3.4%)</td>
</tr>
<tr>
<td><strong>Low homology</strong></td>
<td>58 (0.5%)</td>
</tr>
<tr>
<td><strong>Mean alignment coverage</strong></td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Mean percent identity</strong></td>
<td>98.8</td>
</tr>
</tbody>
</table>
Table 3  Characteristics of *B.mori* exons and introns.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exon</strong></td>
<td></td>
</tr>
<tr>
<td>Total number of exons</td>
<td>51,590</td>
</tr>
<tr>
<td>No of exons per transcript</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>4.8</td>
</tr>
<tr>
<td>Median</td>
<td>4.0</td>
</tr>
<tr>
<td>Exon length (bp)</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>8,783</td>
</tr>
<tr>
<td>Mean</td>
<td>353</td>
</tr>
<tr>
<td>Median</td>
<td>179</td>
</tr>
<tr>
<td>Min</td>
<td>27</td>
</tr>
<tr>
<td><strong>Intron</strong></td>
<td></td>
</tr>
<tr>
<td>Intron length (bp)</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>109,257</td>
</tr>
<tr>
<td>Mean</td>
<td>1,904</td>
</tr>
<tr>
<td>Median</td>
<td>730</td>
</tr>
<tr>
<td>Min</td>
<td>34</td>
</tr>
</tbody>
</table>
Table 4: Protein datasets of 11 insect species tested for orthology using orthoMCL.

<table>
<thead>
<tr>
<th>Species</th>
<th>Order</th>
<th># of proteins</th>
<th>Data source (URL)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. gambiae</td>
<td>Diptera</td>
<td>12,457</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>A. aegypti</td>
<td>Diptera</td>
<td>15,419</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>C. pipiens</td>
<td>Diptera</td>
<td>18,883</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>Diptera</td>
<td>14,076</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>A. mellifera</td>
<td>Hymenoptera</td>
<td>9,257</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>A. pisum</td>
<td>Hemiptera</td>
<td>10,466</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>T. castaneum</td>
<td>Coleoptera</td>
<td>16,645</td>
<td><a href="http://beetlebase.org/">http://beetlebase.org/</a></td>
</tr>
<tr>
<td>P. humanus</td>
<td>Phthiraptera</td>
<td>10,773</td>
<td><a href="http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary">http://orthomcl.org/cgi-bin/OrthoMclWeb.cgi?rm=genome&amp;type=summary</a></td>
</tr>
<tr>
<td>B. mori</td>
<td>Lepidoptera</td>
<td>22,163</td>
<td><a href="http://sgp.dna.affrc.go.jp/KAIKObase/">http://sgp.dna.affrc.go.jp/KAIKObase/</a></td>
</tr>
<tr>
<td>D. plexippus</td>
<td>Lepidoptera</td>
<td>15,130</td>
<td><a href="http://monarchbase.umassmed.edu">http://monarchbase.umassmed.edu</a></td>
</tr>
<tr>
<td>H. melpomene</td>
<td>Lepidoptera</td>
<td>12,829</td>
<td><a href="http://butterflygenome.org/">http://butterflygenome.org/</a></td>
</tr>
</tbody>
</table>
Table 5  Comparison of orthologs and paralogs among three Lepidoptera.

<table>
<thead>
<tr>
<th>Species</th>
<th>Insect-common</th>
<th>Lep-specific N:N</th>
<th>Species-specific</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># Orthologs</td>
<td># Genes(^b)</td>
<td># Orthologs</td>
</tr>
<tr>
<td>Bm</td>
<td>3042</td>
<td>4780 (1.57)</td>
<td>401</td>
</tr>
<tr>
<td>Dp</td>
<td>3042</td>
<td>5061 (1.66)</td>
<td>401</td>
</tr>
<tr>
<td>Hm</td>
<td>3042</td>
<td>4638 (1.52)</td>
<td>401</td>
</tr>
</tbody>
</table>

\(^{a}\)Bm, *Bombyx mori*; Dp, *Danaus plexippus*; Hm, *Heliconius melpomene.*

\(^{b}\)Parenthesis denotes the ratio of [\# Genes]/[# Orthologs].
Table 6  Fractions of clustered tissue-specific genes.

<table>
<thead>
<tr>
<th>Tissue</th>
<th>No of tissue-specific genes</th>
<th>No of clustered tissue-specific genes</th>
<th>Fraction in cluster (%)</th>
<th>Total # of tissue-specific gene clusters</th>
<th>Fraction of gene duplicated in clusters (no. genes)</th>
<th>Tissue-characteristic proteins expressed by duplicated genes in clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>38</td>
<td>9</td>
<td>25</td>
<td>4</td>
<td>0.22 (2)</td>
<td>Adipokine/Adipokinetic hormone</td>
</tr>
<tr>
<td>Compound eyes</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>Cuticular protein/Asirin</td>
</tr>
<tr>
<td>Corpora allata-cardiacia complex</td>
<td>52</td>
<td>8</td>
<td>15</td>
<td>4</td>
<td>0.25 (2)</td>
<td>Paralytic peptide binding protein</td>
</tr>
<tr>
<td>Maxillary galea</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>Cuticular protein/Asirin</td>
</tr>
<tr>
<td>Wing</td>
<td>67</td>
<td>30</td>
<td>45</td>
<td>9</td>
<td>0.57 (17)</td>
<td>Carboxypeptidase/ beta-tubulin/ pyruvate kinase</td>
</tr>
<tr>
<td>Wing disc</td>
<td>54</td>
<td>8</td>
<td>15</td>
<td>4</td>
<td>0.25 (2)</td>
<td>Carboxypeptidase/ beta-tubulin/ pyruvate kinase</td>
</tr>
<tr>
<td>Testis</td>
<td>745</td>
<td>327</td>
<td>44</td>
<td>113</td>
<td>0.20 (66)</td>
<td>Carboxypeptidase/ beta-tubulin/ pyruvate kinase</td>
</tr>
<tr>
<td>Ovary</td>
<td>568</td>
<td>50</td>
<td>79</td>
<td>8</td>
<td>0.46 (24)</td>
<td>Chorion</td>
</tr>
<tr>
<td>Midgut</td>
<td>185</td>
<td>88</td>
<td>48</td>
<td>27</td>
<td>0.89 (78)</td>
<td>Carboxypeptidase/ lipase/ glucosidase/ 30kP protease/ fatty acid-binding protein/ trypsin/ cuticle/ SEC14/ Bm122</td>
</tr>
<tr>
<td>Malpighian tubules</td>
<td>48</td>
<td>11</td>
<td>23</td>
<td>4</td>
<td>1.0 (11)</td>
<td>Sugar transporter/ Na+ -dependent transporter/ synaptic vesicle transporter/ adenylyl cyclase</td>
</tr>
<tr>
<td>Pheromone gland</td>
<td>14</td>
<td>6</td>
<td>43</td>
<td>3</td>
<td>0.67 (4)</td>
<td>Fatty-acid reductase/ Aldehyde oxidase</td>
</tr>
<tr>
<td>Silk gland</td>
<td>22</td>
<td>4</td>
<td>18</td>
<td>2</td>
<td>1.0 (4)</td>
<td>Serrin-like</td>
</tr>
<tr>
<td>Fat body</td>
<td>13</td>
<td>5</td>
<td>38</td>
<td>2</td>
<td>0.6 (3)</td>
<td>30k protein</td>
</tr>
<tr>
<td>Epidermis</td>
<td>26</td>
<td>17</td>
<td>93</td>
<td>6</td>
<td>0.52 (14)</td>
<td>Cuticular protein</td>
</tr>
<tr>
<td>Yersin’s gland</td>
<td>11</td>
<td>0</td>
<td>27</td>
<td>1</td>
<td>1.0 (3)</td>
<td>Serpin</td>
</tr>
<tr>
<td>Total</td>
<td>1,305</td>
<td>569</td>
<td>42</td>
<td>167</td>
<td>0.40 (230)</td>
<td></td>
</tr>
</tbody>
</table>

*These numbers are low estimates since many chorion genes were known to be missing from the assembly (Eickbush and Izzo 2005).
**Gene set A** (mapped to scaffold)
- 16,823 genes
  - FL-cDNAs: 10,666
  - mRNA(RefSeq): 2,029
  - Gene model: 14,623
  - EST: 370,472
- FL-cDNA library: 308,196
- EST library: 62,276

**Gene set B** (EST-based)
- 7,240 genes
  - FL-cDNAs: 0
  - mRNA(RefSeq): 0
  - Gene model: 0
  - EST: 27,102
- FL-cDNA library: 16,467
- EST library: 10,629

**Gene set C** (unmapped to scaffold)
- 6,160 genes
  - FL-cDNAs: 438
  - mRNA(RefSeq): 60
  - Gene model: 0
  - EST: 10,598
- FL-cDNA library: 7,455
- EST library: 3,143

FL-cDNA (11,104)

EST (408,172)
<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (Mb)</th>
<th>Mean exon length (bp)</th>
<th>Mean intron length (bp)</th>
<th>% of TEs in genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. elegans</td>
<td>100.3</td>
<td>147</td>
<td>410</td>
<td>12.0</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>168.7</td>
<td>455</td>
<td>1,141</td>
<td>5.5</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>269.0</td>
<td>284</td>
<td>1,044</td>
<td>24.9</td>
</tr>
<tr>
<td>D. melanogaster</td>
<td>273.0</td>
<td>206</td>
<td>809</td>
<td>11.0</td>
</tr>
<tr>
<td>B. mori</td>
<td>431.7</td>
<td>353</td>
<td>1,904</td>
<td>35.1</td>
</tr>
<tr>
<td>A. pisum</td>
<td>464.0</td>
<td>394</td>
<td>1,719</td>
<td>38.0</td>
</tr>
<tr>
<td>S. purpuratus</td>
<td>514.0</td>
<td>200</td>
<td>1,728</td>
<td>ND</td>
</tr>
<tr>
<td>G. gallus</td>
<td>1,200.0</td>
<td>127</td>
<td>2,662</td>
<td>5.0</td>
</tr>
<tr>
<td>A. aegypti</td>
<td>1,200.0</td>
<td>400</td>
<td>4,789</td>
<td>47.0</td>
</tr>
<tr>
<td>D. rerio</td>
<td>1,505.0</td>
<td>219</td>
<td>2,877</td>
<td>26.5</td>
</tr>
<tr>
<td>M. musculus</td>
<td>2,317.0</td>
<td>111</td>
<td>5,293</td>
<td>37.0</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>3,101.8</td>
<td>157</td>
<td>6,290</td>
<td>45.0</td>
</tr>
</tbody>
</table>

Correlation coefficient R and P-value:

- Mean exon length and genome size: R = -0.487, P-value = 0.109
- Mean intron length and genome size: R = 0.942, P-value < 0.001
- % of TEs in genome and genome size: R = 0.558, P-value = 0.059

Graph showing the relationship between genome size (Mb) and mean exon/intron length (bp), with R² values for each relationship.