DNA Research **11**, 27–35 (2004)

The Genome Sequence of Silkworm, *Bombyx mori*

Kazuei Mita,[1,*] Masahiro Kasahara,[2] Shin Sasaki,[2] Yukinobu Nagayasu,[3] Tomoyuki Yamada,[3] Hiroyuki Kanamori,[4] Nobukazu Namiki,[4] Masanari Kitagawa,[5] Hidetoshi Yamashita,[5] Yuji Yasukochi,[1] Keiko Kadono-Okuda,[1] Kimiko Yamamoto,[1] Masahiro Ajimura,[1] Gopalapillai Ravikumar,[1] Michihiko Shimomura,[6] Yoshiaki Nagamura,[7] Tadasu Shin-i,[8] Hiroaki Abe,[9] Toru Shimada,[10] Shinichi Morishita,[3] and Takuji Sasaki[1]

Genome Research Department, National Institute of Agrobiological Sciences, 1-2 Owashi, Tsukuba, Ibaraki 305-8634, Japan,[1] Department of Computer Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan,[2] Department of Computational Biology, University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8562, Japan,[3] Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, 446-1 Kamiyokoba, Tsukuba, Ibaraki 305-0854, Japan,[4] Dragon Genomics Center, TAKARA BIO Inc., 7870-15 Sakura-cho, Yokkaichi, Mie 512-1211, Japan,[5] Genome Project Department, Tsukuba Division, Mitsubishi Space Software Co., Ltd., 1-6-1 Takezono, Tsukuba, Ibaraki 305-8602, Japan,[6] DNA Bank, Genome Research Department, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan,[7] Center for Genetic Resource Information, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan,[8] Department of Biological Production, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo 183-8509, Japan,[9] and Department of Agricultural and Environmental Biology, University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, Japan[10]

(Received 5 January 2004; revised 2 February 2002)

Abstract

We performed threefold shotgun sequencing of the silkworm (*Bombyx mori*) genome to obtain a draft sequence and establish a basic resource for comprehensive genome analysis. By using the newly developed RAMEN assembler, the sequence data derived from whole-genome shotgun (WGS) sequencing were assembled into 49,345 scaffolds that span a total length of 514 Mb including gaps and 387 Mb without gaps. Because the genome size of the silkworm is estimated to be 530 Mb, almost 97% of the genome has been organized in scaffolds, of which 75% has been sequenced. By carrying out a BLAST search for 50 characteristic *Bombyx* genes and 11,202 non-redundant expressed sequence tags (ESTs) in a *Bombyx* EST database against the WGS sequence data, we evaluated the validity of the sequence for elucidating the majority of silkworm genes. Analysis of the WGS data revealed that the silkworm genome contains many repetitive sequences with an average length of <500 bp. These repetitive sequences appear to have been derived from truncated transposons, which are interspersed at 2.5- to 3-kb intervals throughout the genome. This pattern suggests that silkworm may have an active mechanism that promotes removal of transposons from the genome. We also found evidence for insertions of mitochondrial DNA fragments at 9 sites. A search for *Bombyx* orthologs to *Drosophila* genes controlling sex determination in the WGS data revealed 11 *Bombyx* genes and suggested that the sex-determining systems differ profoundly between the two species.

Key words: silkworm; *Bombyx mori*; WGS; genome sequence

1. Introduction

The domesticated silkworm, *Bombyx mori*, has long been used as a model system for basic studies because of its large body size, ease of rearing in the laboratory, and economic importance in sericulture. The well-developed genetic resources of this species include more than 400 described mutants, which have been mapped to >200 loci, comprising 28 linkage groups,[1] as well as molecular linkage maps developed by using a variety of markers.[2–6] In addition, BAC libraries[7,8] and an EST database based on 36 cDNA libraries totaling more than 35,000 sequences have been constructed.[9] These genetic resources make

Communicated by Michio Oishi

* To whom correspondence should be addressed. Tel. +81-29-838-6120, Fax. +81-29-838-6121, E-mail: kmita@nias.affrc.go.jp

B. mori an ideal reference for Lepidoptera, and thereby this species facilitates studies of comparative genomics and basic research leading toward new genome-based approaches for sericulture and the control of pest species.[10]

With the goal of obtaining a draft sequence of the silkworm genome, we used a whole-genome shotgun (WGS) sequencing strategy. The WGS method has been established as the most powerful tool available for generating a draft genome sequence efficiently, quickly, and more cost-effectively than with the alternative, clone-by-clone sequencing.[11–14] Although the accuracy and quality of the resulting sequence data are lower than those obtained using a BAC-based method, as shown in the rice genome project,[15] a WGS approach is, nevertheless, an efficient way of characterizing the genome structure of an organism. We report here our threefold WGS sequencing of the silkworm genome and subsequent analysis.

2. Materials and Methods

2.1. Shotgun library construction

Genomic DNA was isolated from posterior silkglands of 5th instar larvae on day 3 of strain p50T of *B. mori*, as described previously.[8] Random DNA fragments were generated using the HydroShear process (GeneMachines Inc., USA). The fragmented DNA was fractionated by agarose gel electrophoresis, and approximately 2- to 3-kb fractions and 7- to 10-kb fractions were excised from the gels. Using T4 DNA ligase (Takara Bio Inc., Japan), the short (2 to 3 kb) fragments were ligated into a pUC18 plasmid vector that previously had been digested with *Hinc*II and treated with bacterial alkaline phosphatase. The long (7 to 10 kb) fragments were ligated into a pTWV228 plasmid vector (Takara Bio Inc., Japan) that had been prepared in the same way as pUC18. The ligated DNA samples were introduced into *Escherichia coli* DH10B by electroporation. We constructed two libraries each for the short and long fragments. Insert sizes were estimated by agarose gel electrophoresis after PCR amplification with vector primers. The estimated insert sizes (mean ± 1 standard deviation) for the short-fragment libraries were 2.0 ± 0.35 kb and 3.0 ± 0.35 kb, whereas those for the long-fragment libraries were 7.0 ± 0.65 kb and 11.5 ± 1.8 kb.

2.2. Sequencing

Plasmid DNAs were prepared from overnight cultures by the alkaline lysis method, purified using MultiScreen-FB Plates (Millipore), and used as template DNA in sequencing reactions. Sequencing was performed independently by two sequencing centers. At STAFF-Institute, sequencing was carried out from both ends of plasmid DNAs by using an ABI 3700 capillary sequencer and BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). At Dragon Genomics, sequencing was performed from both ends of amplified DNA with a Templiphi^{TM} DNA Amplification Kit (Amersham Biosciences) using a MegaBACE4000 capillary sequencer and DYEnamic ET Terminator Cycle Sequencing Kit (Amersham Biosciences). The resulting sequence data were evaluated on the basis of Phred scores.[16,17] Sequence data of ≥500 bases in length with a minimum Phred score of 20 were used mainly for sequence assembly. Additionally, sequence data <500 bases in length were used to increase coverage. Sequence data are deposited in DDBJ under accession numbers BAAB01000001 to BAAB01213289.

2.3. RAMEN assembler

The sequences were assembled by RAMEN, a newly developed software program for large-scale whole-genome shotgun sequencing. RAMEN basically follows the overlap layout consensus paradigm, but individual steps have been accelerated by novel or state-of-the-art software implementation ideas such as lookup table generation of seed strings for highly sensitive and rapid detection of overlapping reads, precise alignment by efficient banded dynamic programming, a repeat untangling method of transforming a repeat subcontig flanked by two unique subcontigs into one unique contig, and an efficient multiple alignment algorithm utilizing seeds in the lookup table. These RAMEN algorithms will be described in detail elsewhere (manuscript in preparation).

2.4. Assembly criteria

Considering the low (threefold) coverage and the statistics of the whole-genome shotgun reads, we had to carefully tailor each assembly step to this data set to reduce the rate of misassembly. We briefly describe here the particular treatments of this data set.

In the step for extracting high-quality regions from reads, according to common strict criteria, the longest consecutive bases with quality value[16,17] (QV) scores of >20 were first identified for every raw read. Subsequently, this high-quality range was modestly extended to both sides so that the expectation of error did not exceed three bases according to their QVs. We denote this extended range as the "quality clipped range" (QCR). In the vector masking step, vector sequences around the cloning site were sought in the QCR. Successful identification of the cloning site was followed by elimination of vector sequences from the QCR. Otherwise, to avoid missing vector-derived sequences, we resorted to using the highly sensitive Smith-Waterman algorithm,[18] even though it was computationally costly. When the cloning site was not found by the Smith-Waterman algorithm, we took the deliberate approach of cutting 16 bases from each side of the read to maximize the quality of vector masking. After quality trimming, contaminant removal was carried out by homology search between reads and

known contaminants such as *E. coli*, followed by discarding of reads that had homology to a known contaminant sequence of more than 98% identity over 200 bp.

In the subcontig construction phase, reads that overlapped with each other were combined into subcontigs as long as there was no branch (i.e., repeat boundary[19]). We then discarded short subcontigs that represented fewer than three reads. In the repeat untangling step, at least two end-pairs were required between two unique subcontigs that sandwiched the middle repeat so as to transform the three subcontigs into one unique subcontig. In the scaffolding step, because not all of the links were necessarily correct, due to experimental error or misassembly, confidence scores were given to individual links. Links whose confidence scores were more than a pre-determined threshold were adopted as long as they were consistent with the links of higher confidence scores. In particular, if appropriate, a link that was supported by only one end-pair could be incorporated into the assembly.

2.5. BLAST search

BLAST searches were carried out using BLASTN ver. 2.1.2.[20] Alignments were made using the criteria of >95% identity and >50 bp in length. The coverage was calculated as the ratio of total length of alignments in WGS sequence contigs to the length of the query sequence.

Table 1. Statistics of *Bombyx mori* sequence assembly.

Total number of reads	2,843,020
Number of reads of ~2 kb inserts	2,166,908
Number of reads of ~7 kb inserts	676,112
Number of sequence contigs	213,289
Maximum length of sequence contig	19,243 bp
Minimum length of sequence contig	117 bp
Average length of sequence contigs	1,790 bp
Number of scaffolds	49,345
Maximum scaffold length with gaps	224,537 bp
Maximum scaffold length without gaps	215,846 bp
Average scaffold length with gaps	10,415 bp
Average scaffold length without gaps	7,843 bp
GC content (%)	32.54
Total length of scaffolds with gaps	513,919,331 bp
Total length of scaffolds without gaps	386,552,210 bp
Number of mate pairs in scaffolds	674,490
Number of mate pair orientation violations	6,526
Number of mate pair distance violations	27,818

3. Results and Discussion

3.1. Assembly statistics and quality check

Analysis of the *Bombyx* WGS using the RAMEN assembler program resulted in 213,289 sequence contigs and 49,345 scaffolds. The largest sequence contig was 19,243 bp, and the largest scaffold was 224,537 bp (Table 1). The total scaffold length was 514 Mb; this constitutes 97% of the genome, assuming a size of 530 Mb.[21] The total length of scaffolds without gaps was 387 Mb, suggesting that 73% of the genome has been sequenced. The earlier estimate of genome size (530 Mb) was based on DNA reassociation kinetics,[21] whereas a recent determination using flow cytometry indicated a value of 450 to 493 Mb (J. S. Johnston, personal communication). Based on these new values, the sequence coverage could be as high as 86%.

The assembly was computationally evaluated for integrity of mate pairing. Abnormal mate pairs, either with incorrect orientations or with distances that deviated from the mean plasmid library insert size by three standard deviations, would present an estimation of misassembly rate. Among the 674,490 total mate pairs in scaffolds, only 6526 had orientation violations and 27,818 had distance violations (Table 1). These frequencies of orientation and distance violations are 1.5-fold and 2.4-fold higher, respectively, than those of *Anopheles gambiae*,[12] which may reflect a unique genome structure for the silkworm. Alternatively, the problem could be derived from the abundance and wide distribution of three kinds of high-copy repetitive sequences as discussed later. The difficulty in assembly may be compounded by the holocentric structure of lepidopteran chromosomes, which, in contrast to the monocentric chromosomes of most Metazoa, have many microtubule attachments (dispersed kinetochores) distributed along the poleward chromosome face.[22,23] Centromeres are widely found to consist of tandem repeats of short sequences,[24,25] but detailed sequence information on the holocentric structure of lepidopteran chromosomes has not yet been reported.

To estimate the accuracy in assembly and the precise genome coverage, we aligned sequence contigs and scaffolds in five sequenced BAC clones (4L14, 12L03, 544H24, 559G11, and 534E24; Fig. 1). BAC clones 4L14 and 12L03 were derived from the 320-kb *Bmkettin* locus of the Z chromosome,[8] whereas three other BAC clones from different chromosomes were chosen as references. The red bars in Fig. 1 denote the aligned sequence contigs using the criteria of E-value <e-200 and >99% identity, and the scaffold alignments are shown by green bars. Table 2 summarizes the results of the comparison, in which the matching criteria are >99% identity and >500 bp. Sequence contig coverage ranged from 78.2% to 86.6%. The lowest score was obtained for BAC clone 544H24, which contained a high content of repetitive sequences. The comparison between the reference and aligned sequence

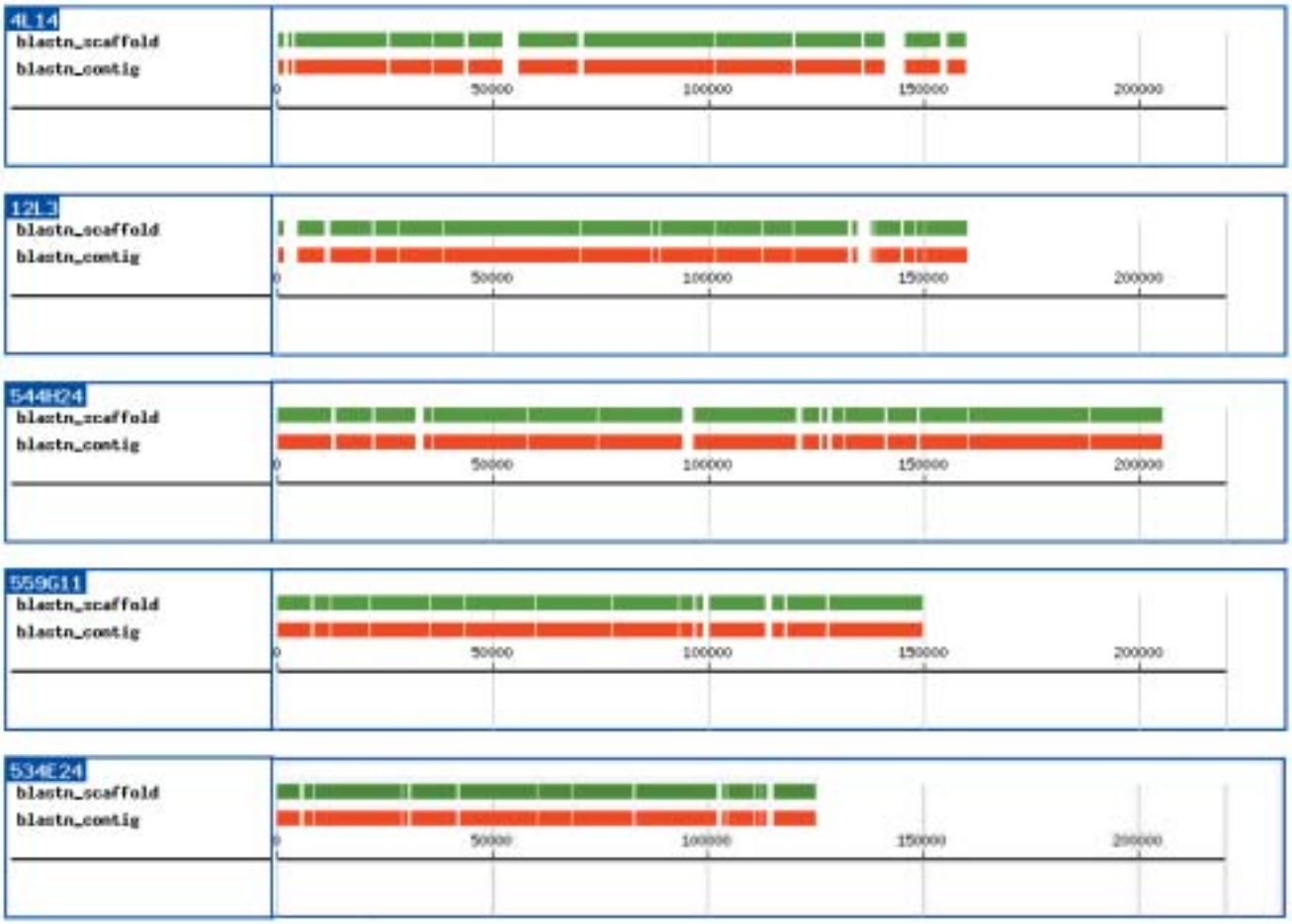

Figure 1. Alignment of sequences in scaffolds (green bars) and sequence contigs (red bars) of five BAC clones. Criteria for alignment: E-value <e-200; identity >99%.

Table 2. Alignment of sequence contigs in five BAC sequences.

BAC clone	Chromosome	Acc. no.	Clone length (bp)	Total length of alignments (bp)	Mismatches (bp)	Coverage (%)
4L14	1/Z	AB090307	159,412	132,610	78	83.1
12L03	1/Z	AB090308	159,882	132,292	55	82.7
544H24	2	AB159445	204,881	160,272	214	78.2
559G11	11	AB159446	149,562	129,540	102	86.6
534E24	13	AB159447	124,898	103,961	71	83.2
Total			798,635	658,675	520	82.5

Criteria: Identity >99%, alignment length >500 bp.

contigs yielded a total of 0.08% sequence error, which seems reasonable for threefold redundancy. We found 8 misassemblies by checking the 373 alignments of >500 bp in the five reference BACs; this corresponds to a 2% misassembly rate.

3.2. Evaluation of the WGS data by alignment of Bombyx genes and ESTs

Searching for known genes in the WGS sequence contigs enables an evaluation of the effectiveness of the WGS data for finding silkworm genes. For this purpose, we chose 50 characteristic *Bombyx* genes whose complete coding sequences (CDSs) are available in public databases (Table 3). These genes were derived from

Table 3. Coverage of complete coding sequences (CDSs) of 50 characteristic *Bombyx* genes in whole-genome shotgun (WGS) sequence contigs.

Query	A cc. no.	CDS (bp)	WGS contig no.	Matching rate (%/bp)	Coverage (%)
Adipokinetic hormone receptor	AF403542	1,893	33198/75075/43320	97.6/1819	96.0
Allatostatin prehormone	AF303370	752	101316	99.2/752	100.0
Allatostatin receptor	AF254742	1,994	68768	99.7/1805	90.5
Annexin B13a	AB063189	1,461	10904/98531	99.0/1443	98.8
Apolipophorin III	AY341912	561	17124	100/525	93.6
BHR38 (Bombyx hormone receptor 38)	X89247	1,112	117489	99.4/496	44.6
BHR78	AF237663	2,003	60130	99.4/1248	62.3
cdc2 kinase	D85134	960	205444	98.9/960	100.0
cdc2-related kinase	D85135	1,215	30480/159991/20571	99.8/1131	93.7
Cecropin A	D17394	192	2642	100/192	100.0
Corazonin preprohormone	AB106876	862	28479/209928/24902	98.9/657	76.2
Cuticle protein WCP10	AB091694	936	53436/10975	98.6/909	97.1
Cyclin B	D84452	1,578	156579	100/1478	93.7
DH-PBAN	D13437	579	21542	99.8/579	100.0
Dopa decarboxylase	AF372836	1,437	86236	98.4/1192	83.0
E75A	AB024904	2,094	39306	99.2/2000	95.5
Ecdysone receptor B1	D43943	1,632	120690/188952	99.4/1632	100.0
Ecdysteroid-phosphate phosphatase	AB107356	996	23777/1148451/168863	99.5/871	87.4
Eclosion hormone	D10135	718	77145	96.2/532	74.1
Egg-specific protein	D12521	1,680	61111/186640	98.0/1626	96.6
Fibroin heavy chain	AF226688	15,792	162781/166328	99.5/2515	16.0
Fibroin light chain	M76430	789	15673	98.9/789	100.0
Fushitarazu-F1	D10953	2,278	109695	98.8/1959	86.0
Hemocytin (Humoral lectin)	D29738	9,402	89891/146356/12043	99.0/7130	76.6
Hemolin	AB115084	1.233	15854/135995	99.2/858	69.6
Insulin receptor-like protein	AF025542	6,020	90196/153069	98.5/4460	74.1
JH acid methyltrasferase	AB113578	837	157748/103943	99.8/565	67.5
JH diol kinase (JHDK)	AY363308	552	85281	98.0/552	100.0
JH epoxido hydrolase	AY377854	1,386	164130	98.5/1369	98.8
KMO	AB063490	1,371	19644	98.4/1162	100.0
Larval serum protein	D12523	789	30387	98.1/789	100.0
Larval SP-T	AB158646	804	192576/49690	98.1/804	100.0
Leucine-rich repeat GR	AF177772	2,211	22771/74090/173993	97.8/2115	95.7
NADH-cytochrome P450 reductase	AB042615	2,064	50987/173329	100/1793	86.9
Prothoracicostatic peptide	AB073553	1.749	11275	96.5/1606	91.8
PTTH	D90082	675	150798	98.5/673	99.7
Samui (cold-inducible)	AB032717	2,034	124542	99.2/1247	61.8
Sericin 1A	AB112019	2,340	9712	97.2/1414	60.4
Seroin 1	AF352584	327	124563/48920	99.1/327	96.6
Serpin-like protein (SEP-LP)	AB017518	1,164	191638/95694	98.4/1164	100.0
Silk protein P25	X04226	663	25099	98.5/663	100.0
Sorbitol dehydrogenase	D13371	1,047	93955/110658	96.8/964	92.1
Transcription factor BmEts	AB115082	1,170	127236	99.7/730	62.6
Trehalase	D13763	1,740	91396	99.0/1740	100.0
Ultraspiracle	U06073	1,389	398/38247	99.2/1141	82.2
Vap-peptide (BmACP-6.7)	AB001053	255	7536	100/172	67.5
Vitellin-degrading protease	D16232	795	18107	100/795	100.0
Vitelline membrane associated P30	AF294885	819	127957	99.0/819	100.0
Vitellogenin	D13160	5,349	12728	98.7/5320	99.5
Wnt-1	D14169	1,179	44551/160582	97.5/1179	100.0

Criteria for alignment: >95% identity and >50 bp in length. Coverage was calculated as the ratio of total length of alignments in WGS sequence contigs to the length of the CDS.

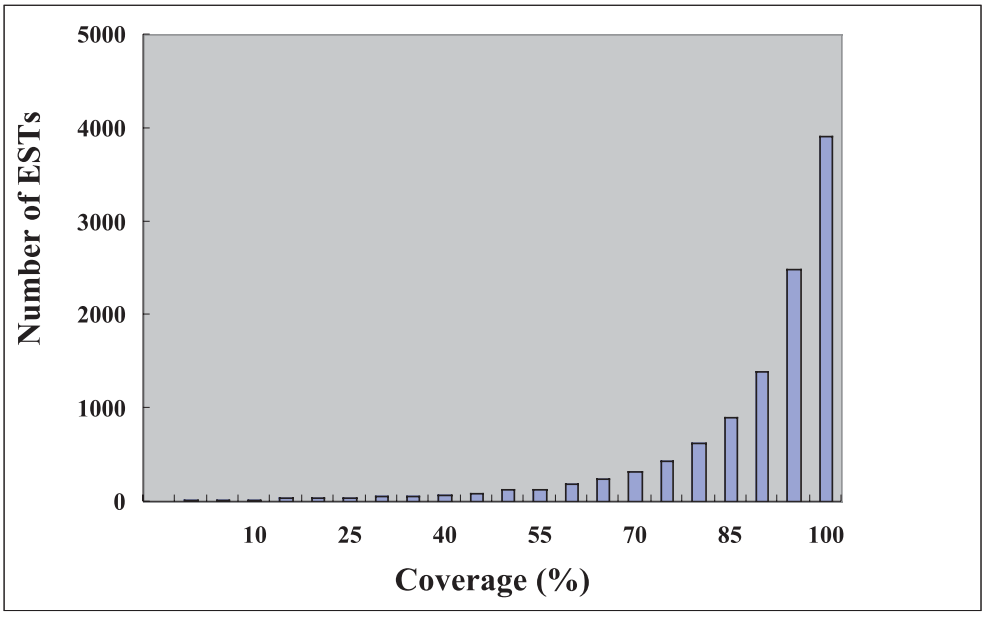

Figure 2. Evaluation of whole-genome shotgun (WGS) sequence contigs by alignment of 11,202 *Bombyx* non-redundant ESTs. Criteria for alignment: >95% identity and >50 bp in length. Coverage was calculated as the ratio of the total length of alignments in WGS sequence contigs to the length of ESTs. The EST accession numbers were reported previously.[9]

many functional classes, such as receptors, silk-related genes, hormone-related genes, diapause-related genes, transcription factors, biological defense-related genes, etc., but none were housekeeping genes. Thirty-two of the 50 genes sampled had more than 90% coverage in WGS sequence contigs, whereas only two genes showed <50% coverage (Table 3). This suggests that the complete structures of about 60% of *Bombyx* genes can be uncovered in the WGS sequence contigs. The lowest matching was the fibroin heavy-chain gene, in which a repetitive sequence tandemly repeats throughout most of the CDS, thereby preventing accurate estimation of coverage.

To achieve the most accurate evaluation of coverage, we carried out alignment of the complete set of 11,202 non-redundant ESTs in SilkBase[9] against the WGS sequence contigs by BLASTN (Fig. 2). Of 11,202 ESTs, only 39 ESTs failed to align significantly with WGS sequence contigs (<10% coverage), whereas 95% of ESTs sampled showed more than 50% coverage. Because the average EST length is about 600 bp, the aligned lengths of most genes would account for more than 300 bp, which is sufficient to find *Bombyx* homologs using amino acid sequences of other species in an amino acid BLAST search, tBLASTN. Thus the present WGS data can be used to identify a large fraction of *Bombyx* genes.

3.3. Transposable elements

The silkworm genome is abundant in repetitive sequences derived from transposons or transposon-related dispersed repeats, long terminal repeat (LTR) retrotransposons, non-LTR retrotransposons, class II transposons, and small interspersed repeat elements (SINEs). From the analysis of DNA reassociation kinetics, Gage[21] estimated the content of repetitive sequences in the silkworm genome to be 45%, with 24% estimated to have 50,000 copies and 21% to have 500 copies per genome. Analysis of 320 kb of the *Bmkettin* region of the Z chromosome[8] revealed that the non-LTR retrotransposons, *BMC1*[26,27] and *L1Bm*,[28] the DNA-type transposon *mariner*,[29] and the SINE *Bm1*,[30] which are dispersed throughout this region, account for 16% of the total sequence. This value is comparable to the transposable element content of characterized segments of *B. mori* autosomes[31] but is different from that in the female-specific W sex chromosome.[32] Table 4 shows the copy numbers estimated from a BLAST search of the *Bombyx* WGS data for identified transposons, including LTR and non-LTR retrotransposons, and the DNA-type transposon *mariner* and SINE *Bm1*, which are highly abundant. Apparently most of these transposons are truncated to 3′ regions of <500 bp in *B. mori*. Altogether there are more than 180,000 copies of repeated elements in the genome, including *Bm1* and 3′ domains of the most abundant *BMC1* and *Bm5886* retrotransposons, which means that well-conserved repetitive sequences are located every 2.5 to 3 kb.

A comparison of transposons between *B. mori* and *A. gambiae* revealed two distinct aspects. First, the three short repetitive elements derived from *BMC1*, *Bm5886*, and *Bm1* constituted about 12% of the silkworm genome, whereas many transposable elements with lower copy

Table 4. Copies of typical transposons calculated by BLAST search in whole-genome shotgun sequencing (WGS) data set.

Type	Name of transposon	GenBank Acc. no.	Length (bp)	Full/Partial	5' copies	3' copies
LTR retrotransposon	*Pao*	L09635	4824	Full	33	280
LTR retrotransposon	*Kabuki*	AB032718 - AB032724	5308	Full	6	130
LTR retrotransposon	*Kamikaze*	AB042118 - AB042143	7098	Full	26	66
Non-LTR retrotransposon	*BMC1*	AB018558	5248	Full	113	37000
Non-LTR retrotransposon	*Bm5886*	This study	3055	Partial	520	28000
Non-LTR retrotransposon	*HOPE Bm2*	AB090825	4200	Partial	180	3380
Non-LTR retrotransposon	*TREST1*	D55702	1746	Partial	193	590
Class II transposon	Tomita's *mariner*	D88671	1623	Full	127	4700
Class II transposon	Robertson's *mariner*	U47917	480	Partial		2400
Class II transposon	*Bmmar6*	AF461149	1316	Partial	3300	3100
SINEs	*Bm1*	X03542	354	Full		121000

Except for Robertson's *mariner* and *Bm1*, sequences were partitioned at every 300 bp, and BLASTN search was performed at a threshold probability of e-10. Full/Partial denotes whether the sequence entry of transposon from GenBank used as query is a full-length or partial sequence.

numbers than are present in silkworm constituted about 16% of the euchromatic component of the *A. gambiae* genome.[12] Second, in *A. gambiae*, the proportion of LTR retrotransposons is much higher than that of non-LTR retrotransposons, whereas the silkworm genome showed the opposite tendency. In the W chromosome, which appears to lack evidence of recombination,[33,34] LTR and non-LTR retrotransposons account for 20% and 44% of the total sequence, respectively,[32] and most of these retrotransposons are present as full-length sequences. These facts suggest that there may be an active mechanism that promotes removal of transposons from the genome.

3.4. Integration of mitochondrial DNA

A BLAST search of mtDNA sequences in the WGS data set (matching criteria, >90% identity and >50 bp) showed that nine separate sequences of 58 to 179 bp of mtDNA corresponding to ∼7% of the mtDNA were integrated into the *Bombyx* genome. Nuclear insertions of mtDNA are found in many organisms, including fungi, insects, vertebrates, and plants. In the grasshopper, the divergence rate between true mitochondrial and nuclear-mitochondrial sequences is 12.5%,[35] whereas that of human is 33%.[36] In contrast, the fruit fly nuclear genome lacks mtDNA, suggesting that *Drosophila*, with a genome size of 176 Mb, has a much higher nucleotide deletion rate than other organisms. This appears to contribute to the near absence of pseudogene sequences in the genome.[37] It is likely that the pressure to delete inserted mtDNA from the genome is not as high in silkworm because of its large genome size.

3.5. Sex determination—a Bombyx-specific biological system

Sex in *Drosophila* is determined by the "X/A ratio", the numeric balance of the X chromosomes and autosomes.[38] In contrast, *Bombyx* determines sex by the presence of the female-specific W chromosome.[39] To clarify the genetic mechanism of sex determination in *Bombyx*, we looked for homologs of the sex-determining genes of *Drosophila* somatic cells in the *Bombyx* WGS sequence (Table 5). More than 50% of the *Drosophila* sex-determining homologs were found in the *Bombyx* WGS sequences, although the overall mechanism of sex determination in *Bombyx* is different from that of *Drosophila*, whose sex-determining system is closely related to the gene dosage compensation of the sex chromosomes. In *Drosophila*, five proteins—Male-specific lethal-1 (MSL1), MSL2, MSL3, Males absent on the first (MOF), and Maleless (MLE)—compose a male-specific protein complex with two RNA molecules (*roX1* and *roX2*) coded on the X chromosome to activate transcription of the X-linked genes only in males. This yields equal amounts of mRNA per cell in both sexes.[40] In contrast, dosage compensation appears to be absent in *Bombyx*,[8,41] suggesting that the *Bombyx* orthologs of MSL-3, MOF, and MLE serve different functions.

3.6. Conclusions

We carried out a threefold shotgun sequencing to generate a draft sequence of the silkworm genome. The WGS data were assembled into 49,345 scaffolds. About 97% of the genome was organized in scaffolds, of which approximately 75% were sequenced. Homology searches using 50 characteristic *Bombyx* genes and 11,202 non-redundant ESTs indicated the validity of the WGS sequence data for identifying *Bombyx* genes. Analysis of transposons

Table 5. Results of search for *Bombyx* genes orthologous to *Drosophila* genes controlling sex determination and gene dosage compensation in somatic cells

Function	*Drosophila* gene as query	Acc. no.	Scaffold	E value in tBLASTN from a *Drosophila* protein to *Bombyx* Ramen scaffolds	E value in BLASTX from a *Bombyx* Ramen contig to *Drosophila* proteins
X chromosome counting	*sis-a*	NP_511116	(no hit)		
	sis-b (sc)	NP_476803	scaffold2730	3e-18	4e-16
	sis-c (os)	NP_525095	(no hit)		
	run	AAC27786	scaffold1093	4e-58	8e-58
	Dpn	Q26263	scaffold4892	2e-23	4e-14
Somatic sex determination	*Sxl*	P19339	scaffold7852	1e-22	2e-21
	Da	P11420	scaffold8463	2e-18	5e-20
	tra	NP_524114	(no hit)		
	tra2	P19018	scaffold4201	3e-20	2e-18
	dsx	NP_731197	scaffold80	2e-20	2e-21
	ix	AAN37397	(no hit)		
	fru	NP_732349	scaffold2784	2e-30	2e-29
	her	Q9VJJ6	(no hit)		
Dosage compensation	*roX1*	U85980	(no hit in BLASTN)		
	roX2	NR_001307	(no hit in BLASTN)		
	msl-1	P50535	(no hit)		
	msl-2	A56720	(no hit)		
	msl-3	P50536	scaffold6061	3e-13	7e-15
	mof	O02193	scaffold15631	8e-92	2e-92
	mle	B40025	scaffold1256	2e-61	3e-61

revealed distinct features of the silkworm genome, including a high frequency (12% of the genome) of three short repetitive elements derived from *BMC1*, *Bm5886*, and *Bm1* distributed approximately every 2.5 to 3 kb. The WGS data provide a powerful basis for the analysis of many biological phenomena specific to *Bombyx* and other Lepidoptera.

Acknowledgements: We thank Dr. Hiroshi Yoshikawa of JT Biohistory Research Hall, Dr. Yuji Kohara of the National Institute of Genetics, Dr. Toshinobu Yaginuma of Nagoya University, and Drs. Chikayoshi Kitamura, Kenjiro Kawasaki, Baltazar A. Antonio and Koh-ichi Kadowaki of the National Institute of Agrobiological Sciences for valuable comments and suggestions. Special thanks to the Human Genome Center, Institute of Medical Science, University of Tokyo for allowing us to execute RAMEN assembler on their parallel computer to accelerate the WGS assembly. This work was supported by the Insect Technology Project of the Ministry of Agriculture, Forestry and Fisheries of Japan.

References

1. Fujii, H., Banno, Y., Doira, H., Kihara, H., and Kawaguchi, Y. 1998, Genetic Stocks and Mutation of *Bombyx mori*. Important Genetic Resources, 2nd Ed., Institute of Genetic Resources, Kyushu University (in Japanese).
2. Promboon, A., Shimada, T., Fujiwara, H., and Kobayashi, M. 1995, Linkage map of random amplified polymorphic DNAs (RAPDs) in the silkworm, *Bombyx mori*, *Genet. Res.*, **66**, 1–7.
3. Yasukochi, Y. 1998, A dense genetic linkage map of the silkworm, *Bombyx mori*, covering all chromosomes based on 1018 molecular markers, *Genetics*, **150**, 1513–1525.
4. Shi, J., Heckel, D. G., and Goldsmith, M. R. 1995, A genetic linkage map for the domesticated silkworm, *Bombyx mori*, based on restriction fragment length polymorphisms, *Genet. Res.*, **66**, 109–126.
5. Kadono-Okuda, K., Kosegawa, E., Mase, K., and Hara, W. 2002, Linkage analysis of maternal EST cDNA clones covering all twenty-eight chromosomes in the silkworm, *Bombyx mori*, *Insect Mol. Biol.*, **11**, 443–451.
6. Reddy, K. D., Abraham, E. G., and Nagaraju, J. 1999, Microsatellites in the silkworm, *Bombyx mori*: abundance, polymorphism, and strain characterization, *Genome*, **42**, 1057–1065.
7. Wu, C., Asakawa, S., Shimizu, N., Kawasaki, S., and Yasukochi, Y. 1999, Construction and characterization of bacterial artificial chromosome libraries from the silkworm, *Bombyx mori*, *Mol. Gen. Genet.*, **261**, 698–706.
8. Koike, Y., Mita, K., Suzuki, M. G. et al. 2003, Genome sequence of a 320-kb segment of the Z chromosome of *Bombyx mori* containing a*kettin* ortholog, *Mol. Gen. Genomics*, **269**, 137–149.
9. Mita, K., Morimyo, M., Okano, K. et al. 2003, The construction of an EST database for *Bombyx mori* and its application, *Proc. Natl. Acad. Sci. USA*, **100**, 14121–14126.

10. Heckel, D. G. 2003, Genomics in pure and applied entomology, *Ann. Rev. Entomol.*, **48**, 236–260.
11. Adams, M. D., Celniker, S. E., Holt, R. A. et al. 2000, The genome sequence of *Drosophila melanogaster*, *Science*, **287**, 2185–2195.
12. Holt, R. A., Suburamanian, G. M., Harpern, A. et al. 2003, The genome sequence of the Malaria mosquito *Anopheles gambiae*, *Science*, **298**, 129–149.
13. Venter, J. C., Adams, M. D., Myers, E. W. et al. 2001, The sequence of the human genome, *Science*, **291**, 1304–1351.
14. Mural, R. J., Adams, M. D., Myers, E. W. et al. 2002, A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome, *Science*, **296**, 1661–1671.
15. Sasaki, T., Matsumoto, T., Yamamoto, K. et al. 2002, The genome sequence and structure of rice chromosome 1, *Nature*, **420**, 312–316.
16. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. 1998, Base-calling of automated sequencer traces using Phred. I. Accuracy assessment, *Genome Res.*, **8**, 175–185.
17. Ewing, B., and Green, P. 1998, Base-calling of automated sequencer traces using Phred. II. Error probabilities, *Genome Res.*, **8**, 186–194.
18. Smith, T. F., and Waterman, M. S. 1981, Comparison of biosequences, *Adv. Appl. Math.*, **2**, 482–489.
19. Myers, E. W. 1995, Toward simplifying and accurately formulating fragment assembly. *J. Comp. Biol.*, **2**, 275–290.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A. et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucl. Acids Res.*, **25**, 3389–3402.
21. Gage, L. P. 1974, The *Bombyx mori* genome analysis by DNA reassociation kinetics, *Chromosoma*, **45**, 27–42.
22. Wolf, K. W., and Traut, W. 1991, Cytology of Lepidoptera. VII The restructuring of eupyrene prophase 1 spermatocytes and its relationship to meiotic chromosome and spindle organization in *Ephestia kuehniella Z, Protoplasma*, **165**, 51–63.
23. Wolf, K. W., Novak, K., and Marec, F. 1997, Kinetic organization of metaphase I bivalents in spermatogenesis of Lepidoptera and Trichoptera species with small chromosome numbers, *Heredity*, **79**, 135–143.
24. Friedlander, M., and Wahrman, J. 1970, The spindle as a basal body distributor. A study in the meiosis of the male silkworm moth, *Bombyx mori*, *J. Cell Sci.*, **7**, 65–89.
25. Zhou, C. Z., and Liu, B. 2001, Identification and characterization of a silkgland-related matrix association region in *Bombyx mori*, *Gene*, **277**, 139–144.
26. Ogura, T., Okano, K., Tsuchida, K. et al. 1994, A defective non-LTR retrotransposon is dispersed throughout the genome of the silkworm, *Bombyx mori*, *Chromosoma*, **103**, 311–323.
27. Abe, H., Ohbayashi, F., Shimada, T., Sugazaki, T., Kawai, S., and Oshiki, T. 1998, A complete full-length non-LTR retrotransposon, *BMC1*, on the W chromosome of silkworm, *Bombyx mori*, *Genes Genet. Syst.*, **263**, 916–924.
28. Ichimura, S., Mita, K., and Sugaya, K. 1997, A major non-LTR retrotransposon of *Bombyx mori*, *L1Bm*, *J. Mol. Evol.*, **45**, 253–264.
29. Robertson, H. M., and Asplund, M. L. 1996, *Bmmar1*: a basal lineage of the mariner family of transposable elements in the silkworm moth, *Bombyx mori*, *Insect Biochem. Mol. Biol.*, **26**, 945–954.
30. Adams, D. S., Eickbush, T. H., Herrera, R. I., and Lizardi, P. M. 1986, A highly reiterated family of transcribed oligo(A)-terminated, interspersed DNA elements in the genome of *Bombyx mori*, *J. Mol. Biol.*, **187**, 465–478.
31. Zhou, C. Z., Confalonieri, F., Medina, N. et al. 2000, Fine organization of *Bombyx mori* fibroin heavy chain gene, *Nucleic Acids Res.*, **25**, 2413–2419.
32. Abe, H., Mita, K., Yasukochi, Y., Oshiki, T., and Shimada, T. 2003, Retrotransposable elements on the W chromosome of the silkworm, *Bombyx mori*, *Cytogenet. Genome Res.*, in press.
33. Sturtevant, A. H. 1915, No crossing over in the female of the silkworm moth, *Am. Nat.*, **49**, 42–44.
34. Rasmussen, S. W. 1977, The transformation of the synaptonemal complex into the 'elimination chromatin' in *Bombyx mori* oocytes, *Chromosoma*, **60**, 205–221.
35. Bensasson, D., Zhang, D., and Hewitt G. M. 2000, Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes, *Mol. Biol. Evol.*, **17**, 406–415.
36. Woischnik, M., and Moraes, C. T. 2002, Pattern of organization of human mitochondrial pseudogenes in the nuclear genome, *Genome Res.*, **12**, 885–893.
37. Bensasson, D., Zhang, D., Hartl, D. I., and Hewitt, G. M. 2001, Mitochondrial pseudogenes: Evolution's misplaced witness, *Trends Ecol. Evol.*, **16**, 314–321.
38. Liu, Y., and Belote, J. M., 1995, Protein-protein interactions among components of the *Drosophila* primary sex determination signal, *Mol. Gen. Genet.*, **248**, 182–189.
39. Ohbayashi, F., Suzuki, M. G., Mita, K., Okano, K., and Shimada, T., 2001, A homologue of the *Drosophila doublesex* gene is transcribed into sex-specific mRNA isoforms in the silkworm, *Bombyx mori*, *Comp. Biochem. Physiol. B*, **128**, 145–158.
40. Kelley, R. L., Meller, V. H., Gordadze, P. R., Roman, G., Davis, R. L., and Kuroda, M. I. 1999, Epigenetic spreading of the *Drosophila* dosage compensation complex from *roX* RNA genes into flanking chromatin, *Cell*, **98**, 513–522.
41. Suzuki, M. G., Shimada, T., and Kobayashi, M. 1998, Absence of dosage compensation at the transcription level of a sex-linked gene in a female heterogametic insect, *Bombyx mori*, Heredity, **81**, 275–283.