The construction of an EST database for *Bombyx mori* and its application

Kazuei Mita*, Mitsuoki Morimyo†, Kazuhiro Okano‡§, Yoshiko Koike¶‖, Junko Nohata*, Hideki Kawasaki, Keiko Kadono-Okuda*, Kimiko Yamamoto*, Masataka G. Suzuki‡¶, Toru Shimada¶, Marian R. Goldsmith††‡‡, and Susumu Maeda‡§§**

*Laboratory of Insect Genome, National Institute of Agrobiological Sciences, Owashi 1-2, Tsukuba, Ibaraki 305-8634, Japan; †National Institute of Radiological Sciences, Anagawa 4-9-1, Inage-ku, Chiba 263-8555, Japan; ‡Laboratory of Molecular Entomology and Baculovirology, The Institute of Physical and Chemical Research (RIKEN), Hirosawa 2-1, Wako, Saitama 351-0198, Japan; ¶Department of Agricultural and Environmental Biology, University of Tokyo, Yayoi 1-1-1, Bunkyo-ku, Tokyo 113-8657, Japan; **Faculty of Agriculture, Utsunomiya University, 350 Mine, Utsunomiya, Tochigi 321-8505, Japan; and ††Biological Sciences Department, University of Rhode Island, 100 Flagg Road, Kingston, RI 02881-0816

Edited by Fotis C. Kafatos, European Molecular Biology Laboratory, Heidelberg, Germany, and approved September 15, 2003 (received for review August 16, 2002)

To build a foundation for the complete genome analysis of *Bombyx mori*, we have constructed an EST database. Because gene expression patterns deeply depend on tissues as well as developmental stages, we analyzed many cDNA libraries prepared from various tissues and different developmental stages to cover the entire set of Bombyx genes. So far, the *Bombyx* EST database contains 35,000 ESTs from 36 cDNA libraries, which are grouped into ≈11,000 nonredundant ESTs with the average length of 1.25 kb. The comparison with FlyBase suggests that the present EST database, SilkBase, covers >55% of all genes of *Bombyx*. The fraction of library-specific ESTs in each cDNA library indicates that we have not yet reached saturation, showing the validity of our strategy for constructing an EST database to cover all genes. To tackle the coming saturation problem, we have checked two methods, subtraction and normalization, to increase coverage and decrease the number of housekeeping genes, resulting in a 5–11% increase of library-specific ESTs. The identification of a number of genes and comprehensive cloning of gene families have already emerged from the SilkBase search. Direct links of SilkBase with FlyBase and WormBase provide ready identification of candidate Lepidoptera-specific genes.

In eukaryotes, the genome projects of various species have been vigorously pushed forward. Among model organisms, genome sequences have been completed in the following two insects: *Drosophila melanogaster* (1), and the malaria mosquito, *Anopheles gambiae* (2), whereas those of *Drosophila pseudoobscura* (www.hgsc.bcm.tmc.edu/projects/drosophila) and honey bee (www.hgsc.bcm.tmc.edu/projects/honeybee) will be finished shortly. In addition, whole-genome sequencing is underway or is planned in many other species. Genome information provides powerful tools for understanding biological mechanisms and functions and is essential for biology, medical science, and agriculture.

The Lepidoptera include the most highly destructive agricultural pests; hundreds of species of caterpillars cause widespread economic damage on food and fiber crop plants, fruit trees, forests, and stored grains. They are also important indicators of ecosystem diversity and health, serving as both pollinators and prey. Lepidopteran genome information will make a strong impact on insect science and industries such as insecticide, pest control, and silk production. In Lepidoptera, however, genome information is quite limited so far. As Diptera and honey bee are fairly distant from Lepidoptera evolutionarily (≈250 and 300 million years ago, respectively), the genome analysis of species closely related to Lepidoptera has not yet been performed. The domesticated silkworm, *Bombyx mori*, has been used as a model for basic studies and provides a number of mutants and genetically improved strains. In addition, several groups have engaged in the construction of molecular linkage maps in the silkworm by using a variety of markers, with the aims of providing a framework for positional cloning of specific genes and mutations, large-scale physical map construction, analysis of quantitative trait loci, and comparative genomics. To date, ≈1,500 markers based on random amplified polymorphic DNAs (3, 4), restriction fragment length polymorphisms (5, 6), amplified fragment length polymorphisms (7), and microsatellites (8) are available for the construction of molecular linkage maps, which now cover all 28 *Bombyx* chromosomes at an average spacing of 2 cM, which is equivalent to ≈500 kb (4). The extensive genetic resources for *B. mori* make it an ideal reference for the Lepidoptera, where comparative genetics and genomics can work together to elucidate conserved evolutionary pathways and their diversification, identify new genes and gene systems as targets for transgenesis, and provide basic research leading toward new genome-based approaches for the control of pest species (9).

Aiming for the complete analysis of the *Bombyx* genome, we are taking the following strategy: (*i*) the construction of an EST database, followed by sequencing of full-length cDNAs, (*ii*) the construction of a *Bombyx* bacterial artificial chromosome (BAC) library, (*iii*) making BAC contigs based on DNA fingerprinting and EST markers anchored to linkage maps, and (*iv*) genomic sequencing by BAC shotgun sequencing. In this article, we report on the progress of the EST database as a step toward the complete analysis of the *Bombyx* genome, and some of its initial applications.

Materials and Methods

Strategy for Preparing a Comprehensive *Bombyx* cDNA Catalog. A cDNA catalog is the comprehensive identification of all expressed genes by large-scale cDNA sequencing. It is an essential tool for genome annotation and an important resource for studies of gene structure, regulation, and function. Identifying individual cDNAs is most easily accomplished by first constructing an EST database. There are now relatively few comprehensive EST projects reported for insects, notably, for *D. melanogaster* (≈261,000 ESTs in public databases as of April, 2003; refs.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: BAC, bacterial artificial chromosome.

§Present address: Department of Microbiology, Nash Hall 220, Oregon State University, Corvallis, OR 97331-3804.

‖Present address: International Medical Center of Japan, Toyama 1-21-1, Shinjuku-ku, Tokyo 162-8655, Japan.

‡‡To whom correspondence should be addressed. E-mail: mki101@uri.edu.

§§Deceased March 27, 1998.

Table 1. List of *Bombyx* cDNA libraries for EST database

Library	Tissue/developmental stage	*Bombyx* strain	Accession nos. in GenBank/EMBL/DDBJ
N---	Cultured cell, BmN	Unknown*	AU002477–AU003299
NV02	Baculovirus-infected BmN, 2 h postinfection	Unknown*	AV398029–AV398586
NV06	Baculovirus-infected BmN, 6 h postinfection	Unknown*	AV398587–AV399270
NV12	Baculovirus-infected BmN, 12 h postinfection	Unknown*	AV399271–AV399916
br—	Brain, P0	p50	AV399917–AV400934
brS-	Brain, mixed stages fifth-instar, day 3 to S3	p50	BP116407–BP117204
brP-	Brain, mixed pupal stages after P1	p50	BP114820–BP116406
ce–	Compound eyes, mixed stages fifth-instar to pupa	C202 × J201	BP117205–BP118782
ceN-	The same as above but a different preparation	C202 × J201	BP118783–BP122594
e40h	Embryo, 40 h after fertilization	p50	AU000001–AU000762
e96h	Embryo, 96 h after fertilization	p50	AV400935–AV401647
epV3	Epidermis, fifth-instar day 3	p50	BP122595–BP124563
fbS2	Fat body, S2 (mixed sexes)	C202 × J201	BP124564–BP124844
fbVf	Female fat body, fifth-instar day 3	p50	AU000763–AU001065
fbVm	Male fat body, fifth-instar day 3	p50	AU001066–AU001867
fbpv	Baculovirus-infected fat body, S2, 2 h postinfection	Shuko × Ryuhaku	BP124845–BP125532
heS0	Hemocytes, S0	C108	AV401648–AV402478
heS3	Hemocytes, S3	C108	AV402479–AV403095
maV3	Malpighian tubule, fifth-instar day 3	p50	BP177602–BP179298
mg–	Midgut, fifth-instar	p50	AU001868–AU002476
msgV	Middle silkgland, fifth-instar	N02 × C02	AV403096–AV403745
ovS3	Ovary, S3	p50	BP179299–BP182208
pg–	Pheromone gland, moth	Shuko × Ryuhaku	AV403746–AV404455
NRPG	Pheromone gland, moth, normalized library	p50	BP182009–BP183529
P5PG	Pheromone gland, moth	p50	BP183530–BP184340
prgv	Prothoracic gland, fifth-instar	p50	AV404456–AV405268
ps4M	Posterior silkgland, fourth molt	p50	BP126121–BP126372
psV3	Posterior silkgland, fifth-instar day 3	p50	BP126373–BP127101
tesS	Testis, spinning stage	p50	BP127102–BP127988
tesV	Testis, fifth-instar	p50	BP127989–BP128308
wdV1	Wing disc, fifth-instar day 1	C108	AV405269–AV405596
wdV3	Wing disc, fifth-instar day 3	C108	AV405597–AV406327
wdV4	Wing disc, fifth-instar day 4	C108	AU005709–AU006458
wdS0	Wing disc, S0	C108	AU003300–AU004166
wdS2	Wing disc, S2	C108	AU004167–AU004926
wdS3	Wing disc, S3	C108	AU004927–AU005708

Stages: P0, beginning of pupation; P1, 1 day after pupation; S0, beginning of spinning; S2, 2 days after spinning; S3, 3 days after spinning.
*BmN cultured cell was derived from ovary (19).

10 and 11) and *A. gambiae* (≈99,800 ESTs in public databases; ref. 12), with relatively smaller ones for *D. pseudoobscura* (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Dpseudoobscura/EST; ≈34,000), honey bee (≈15,400 ESTs; ref. 13), cat flea, *Ctenocephalides felis* (≈4,800 ESTs; ref. 14), beetles (≈6,700 ESTs; ref. 15), yellow fever mosquito (≈3,500 ESTs in public databases), and the lepidopteran, *Manduca sexta* (≈2,000 ESTs; ref. 16). Unlike unicellular organisms such as yeast (17), the gene expression patterns in multicellular species deeply depend on tissues as well as developmental stages. The cDNAs from which the ESTs are derived are present in libraries in proportion to the levels of mRNA in the tissue from which the library was prepared. Thus, ESTs are subject to ''expression bias'' for multicellular species (18) and many sequences will be either over- or underrepresented. Therefore, taking advantage of the large size of the silkworm, we used the following strategy: We prepared many cDNA libraries of various tissues and different developmental stages and then carried out random sequencing of a large number of cDNA clones from each library. The use of a large number of cDNA libraries is an effective way to cover almost the whole set of *Bombyx* genes. In addition, this approach explicitly represents the tissue- and stage-specific gene expression patterns of all genes identified. Another advantage of this method is to represent all members of related genes and to identify all members participating in the pathway of each biological process that the cells (or tissues) employ. So far, 36 cDNA libraries of various tissues and different developmental stages have been constructed (Table 1).

cDNA Library Construction. The procedure of making the *Bombyx* cDNA catalog was as follows: First, poly(A)$^+$ RNAs were extracted from many tissues, followed by cDNA synthesis with oligo-dT primers. To ensure the greatest effectiveness in gene classification by protein homology search, cDNA libraries were made by the directional cloning method from the 5′ end (Lambda Zap cDNA cloning kit; Stratagene).

EST Sequencing and Analysis. More than 1,000 cDNA clones were picked up randomly from each library, and a sequence of ≈700 nucleotides from the 5′ end of each cDNA was determined. By a comparison of the deduced amino acid sequences with public protein databases such as PIR and Swiss-Prot, a gene classification was assigned by using a criterion of homology of >30% identity in a sequence >100 amino acids as well as an *E* value lower than −10 in a BLAST search (20). This step was followed by a nucleotide homology search in the *Bombyx* EST database to give the expression profile of each gene.

Results and Discussion

Compilation of the EST Database. We have sequenced ≈35,000 cDNA clones from the 36 cDNA libraries to date. All ESTs have been classified into ≈11,000 groups. Grouping was performed by the following two steps: (*i*) collecting identical clones by BLAST search, using a criterion of >95% identity in a sequence >100 nucleotides, and (*ii*) grouping by CLUSTAL W with default parameters (21). Each group was represented by an EST having the longest sequence among the group. This may cover >55% of all of the genes of *Bombyx*, judging from the comparison with FlyBase discussed below (Table 4). All sequenced ESTs are compiled into the *Bombyx* EST database, named SilkBase, which can be accessed at www.ab.a.u-tokyo.ac.jp/silkbase. SilkBase is equipped with functions such as keyword/clone name search and BLAST search, which facilitate searching for homologous *Bombyx* cDNAs with known amino acid sequences of other species. Functions are also available for direct comparisons with FlyBase and WormBase. This capability will provide a powerful tool to screen candidates for Lepidoptera-specific genes. Sequence data from this article have been deposited in DDBJ (DNA Data Bank of Japan) under accession numbers shown in Table 1. Clones can be obtained by contacting K.M.

cDNA Libraries. Table 1 presents a list of the cDNA libraries analyzed as well as information on *Bombyx* strains from which the libraries were obtained. To study the changes of gene expression patterns during metamorphosis and development, several libraries were made from successive developmental stages of the same tissue, including brain, embryo, fat body, hemocyte, silk gland, testis, and wing disk. For fat body, in addition to fifth-instar larval and prepupal stages (fbS2), the sex difference in gene expression can be observed by a comparison between libraries derived from fifth-instar larval male (fbVm) and female fat body (fbVf). A comparison of the gene patterns between two imaginal discs, compound eye and wing disk, will reveal shared and unique factors responsible for tissue-specific gene expression under the control of ecdysteroid hormone. For imaginal disk development, cDNA libraries of six successive stages of wing disk, wdV1, wdV3, wdV4, wdS0, wdS2 and wdS3, were prepared to analyze the genes related to tissue differentiation during metamorphosis (22, 23). For the compound eye cDNA library, because the tissue is vanishingly small and difficult to prepare enough for each successive stage, we made two cDNA libraries of mixed stages (ce-- and ceN-) and sequenced >5,000 clones. For brain, having a similar situation as the compound eye, we prepared three libraries: a mixture of stages from day 3 of the fifth-larval instar to day 3 after start of spinning (brS-), beginning pupation stage (br--), and pupal stage (brP-). For the silkgland, comparison of posterior silkgland cDNA libraries between fourth molt (ps4M) and the fifth-instar (psV3) can reveal genes active during molting, whereas a comparison between middle (msgV) and posterior (psV3) silkglands of fifth-instar larvae, which produce different classes of silk proteins, may provide information on tissue-specific gene expression. We also prepared cDNA libraries of baculovirus-infected cultured cells at 0, 2, 6, and 12 h postinfection (N---, NV02, NV06, NV12) and pupal fat body (fbpv) 2 h postinfection to investigate the effects of virus infection on the gene expression patterns and mechanism of host cell defense.

Characterization of cDNA Libraries and Approaches to Increase Coverage. Table 2 presents the average size of cDNA clones in each library. The size of the cDNA was estimated from agarose gel electrophoresis of PCR products amplified with vector primers. The average size of the cDNA was 0.92–1.59 kb, which indicates the high quality of the cDNA libraries analyzed. Table 2 also shows the fraction of library-specific ESTs in the complete set of ESTs in 22 of the cDNA libraries arranged in the order of when they were analyzed. The values fluctuate between 17–49%, showing no correlation with the time of analysis. This finding indicates that we have not yet reached saturation with clones that have already been found in other cDNA libraries. However, it is obvious that saturation will become a serious problem as the ESTs accumulate more extensively. Therefore, we tried subtraction methods for compound eye (ce) and cultured cells (BmN) to increase the yield of new sequences. For the ce cDNA library, we dotted ≈100 cDNA clones onto a nylon membrane filter (GeneScreen; NEN) and hybridized with a mixed probe containing 20 house-keeping genes, which are expressed highly and ubiquitously, including cDNA clones encoding several ribosomal proteins, elongation factor 1-α, tubulin, and actin. We then picked up the clones that did not hybridize strongly with the mixed probe, denoted cesb, and sequenced them. In another approach, we used a normalization method for the cultured cell cDNA library, denoted as Nnor (24, 25). The method will reduce the high variation in abundance among the clones of a cDNA library through a reassociation procedure of which a detailed protocol was described by Soares *et al.* (24). Single-stranded circular cDNAs of a directionally cloned (lambda ZAP; Stratagene) cDNA library were primer-extended under controlled conditions to synthesize short complementary strands. Resulting partial duplexes were melted and reannealed to a relatively low Cot, followed by hydroxyapatite column chromatography to recover unassociated circles as a normalized library. A comparison between the original library and subtracted one by using the two methods is presented in Table 3. The fraction of library-specific ESTs was increased by 5–11%, whereas the fraction of house-keeping genes examined was decreased by both subtraction procedures, indicating that these approaches efficiently increased the coverage of genes.

Table 2. Characterization of cDNA libraries

Library	ESTs, no.	Average size, kb	Fraction of library-specific ESTs, %
N---	756	1.10	23.0
mg–	604	0.92	31.1
e40h	756	1.58	32.7
e96h	704	1.39	32.2
pg–	468	1.38	24.8
P5PG	811	ND	29.1
wdV4	1,897	1.40	24.4
wdS0	864	1.21	26.9
wdS2	760	1.57	20.1
wdS3	769	1.20	25.7
prgv	779	1.22	49.2
msgV	633	ND	25.3
heS0	828	1.15	17.2
heS3	588	0.95	26.7
fbS2	281	1.44	30.1
fbpv	688	1.21	23.0
ce–	1,578	1.02	33.9
an–	606	1.34	33.3
brP-	1,587	1.40	24.8
brS-	798	ND	24.8
maV3	1,697	ND	21.9
ovS3	2,710	ND	27.9

The cDNA libraries are arranged in the order of the time when the library was analyzed, i.e., top, oldest; bottom, latest. ND, not determined.

Overview of Results Obtained from the Silkworm EST Database. Table 4 summarizes the set of silkworm ESTs identified as highly homologous to known genes and classified by the GO system (26,

GENETICS

Table 3. Effects of normalization and subtraction on cDNA libraries

Library	ESTs, no.	Library-specific ESTs, %	Ribosomal proteins, %	Tubulin, %	Actin, %	Elongation factors, %
pg	468	24.8	5.3	0.8	4.0	1.5
NRPG	1,521	29.8	4.4	0.1	0.7	0.8
N---	752	23.0	9.4	1.1	1.5	2.8
Nnor	221	33.9	5.4	0.9	0.9	0
ce–	1,578	29.4	3.6	3.9	12.4	0.6
cesb	83	38.6	3.1	2.4	1.3	0

Explanations for Nnor and cesb cDNA libraries are given in the text.

27). Most highly expressed house-keeping genes, especially genes related to the translation machinery such as ribosomal proteins, translation initiation factors, and elongation and release factors have already been identified, whereas the coverage seems rather low for the tissue-specific and developmental stage-specific genes with low transcription levels such as transcription factors, developmental proteins, channel-related proteins, and proteins for various transport systems. The comparison with FlyBase shown in Table 4 indicates that the present SilkBase contains sequences homologous to ≈55% of *Drosophila* genes.

Several interesting results, such as the finding of previously unidentified genes and comprehensive cloning of some gene families have already emerged from the SilkBase search. In wing development, previously unobserved ecdysteroid-inducible genes have been identified (28, 29), and, altogether we found nine cuticle protein genes expressed in the prepupal wing disk cDNA libraries (30). A couple of important sex-differentiation factors including the *doublesex* gene were identified (31, 32), which can provide clues to understanding the sex-determination and -differentiation mechanisms in the silkworm. These findings are based on a putative female-determining factor located on the W chromosome (*Bombyx* females are heterogametic, ZW, and males homogametic, ZZ; ref. 33), and critical differences from *Drosophila*, whose sex-determining mechanism is based on the ratio of sex (X) chromosomes to autosomes (34). From the pheromone gland cDNA libraries, we have identified homologs for acyl-CoA desaturase and acyl-CoA-binding proteins, which play a significant role in the production of the sex pheromones regulated by the neurohormone, PBAN (35, 36). Comparative EST analysis of cDNA libraries derived from BmNPV-infected BmN cells revealed that the expression of several genes, including cytochrome C oxidase 1, increases in the late stages of virus infection, although most of the host genes are depressed as the infection progresses (37). In addition, two apoptosis-related genes of the host cells were identified during virus infection. We have also identified what appear to be all members of the *Bombyx* α- and β-tubulin gene families (38), and found a set of proteinase inhibitors associated with *B. mori* cocoons (39). Many other interesting genes have been identified in SilkBase, which have provided tools for homology-based identification of related genes in other insects, as well as phylogenetic analysis. These genes include a bacteria-induced serine proteinase inhibitor serpin found in *Manduca sexta* (40), heat shock protein and related genes in *Spodoptera frugiperda* (41, 42), and a member of the cadherin superfamily associated with resistance to the *Bacillus thuringiensis* δ-endotoxin Cry1Ac in *Heliothis virescens* (43).

Table 4. Silkworm ESTs categorized by gene ontology terms

Gene ontology term	No. of categorized genes in SilkBase	No. of categorized genes in FlyBase	Ratio of SilkBase vs. FlyBase, %
Ribosomal protein	97	104	93
Translation initiation factor	20	21	95
Elongation factor	9	9	100
Actin binding	109	127	86
Tubulin binding	28	39	72
RNA binding*	274	359	76
Chaperone/heat shock	160	148	95
Proteasome	57	143	40
Transcription factor	280	485	58
Developmental protein	71	~150	47
Cell cycle	110	193	57
Cell adhesion	37	47	79
Axon/ neurotransmitter	98	163	60
Signal transduction	104	246	42
Channel	40	136	29
Protein kinase/ phosphatase	222	370	60
Enzyme†	1,270	2,595	49
Motor	63	98	64
Transport	391	803	49
Total	3,440	6,236	55.2

*Includes RNA processing and small nuclear ribonucleoprotein complex.
†Does not include protein kinase/phosphatase.

Applications of the *Bombyx* EST Database. ***Construction of* B. mori *BAC contigs and physical map.*** We have constructed a high-quality *B. mori* BAC library in collaboration with P. de Jong's group (Children's Hospital Oakland Research Institute, Oakland, CA). Its average insertion size was estimated to be 168 kb with 11-fold redundancy. High molecular weight genomic DNA of *B. mori* (p50 strain) was extracted from isolated nuclei of posterior silkglands of day 2 of fifth-instar larvae (44). Construction of the *Bombyx* BAC library was followed exactly by the protocol of Osoegawa *et al.* (45). We have begun to construct BAC contigs by filter hybridization, using nonredundant ESTs as probes. Based on the genome size of 530 Mb (46, 47) and even distribution of protein-coding sequences, a single BAC clone will have, on average, 3–4 EST markers if 11,000 nonredundant ESTs are available, indicating that more will be needed to construct the complete set of BAC contigs. An important advantage of using this approach is that the physical map made by this method is colinear to the gene map because ESTs are used as markers. This result offers the major advantage that the ESTs can serve as anchors to the genetic map for positional cloning of mutations and for analysis of comparative genome organization of other Lepidoptera and other insects.

EST microarray. Another major application of the EST database is microarrays, which can be used for many functional studies as well as for genome analysis, because they can provide quantitative expression profiles of a large number of genes at one time. From the hybridization experiments using high-density replica filters with EST probes to make BAC contigs, we found that >10% of *B. mori*

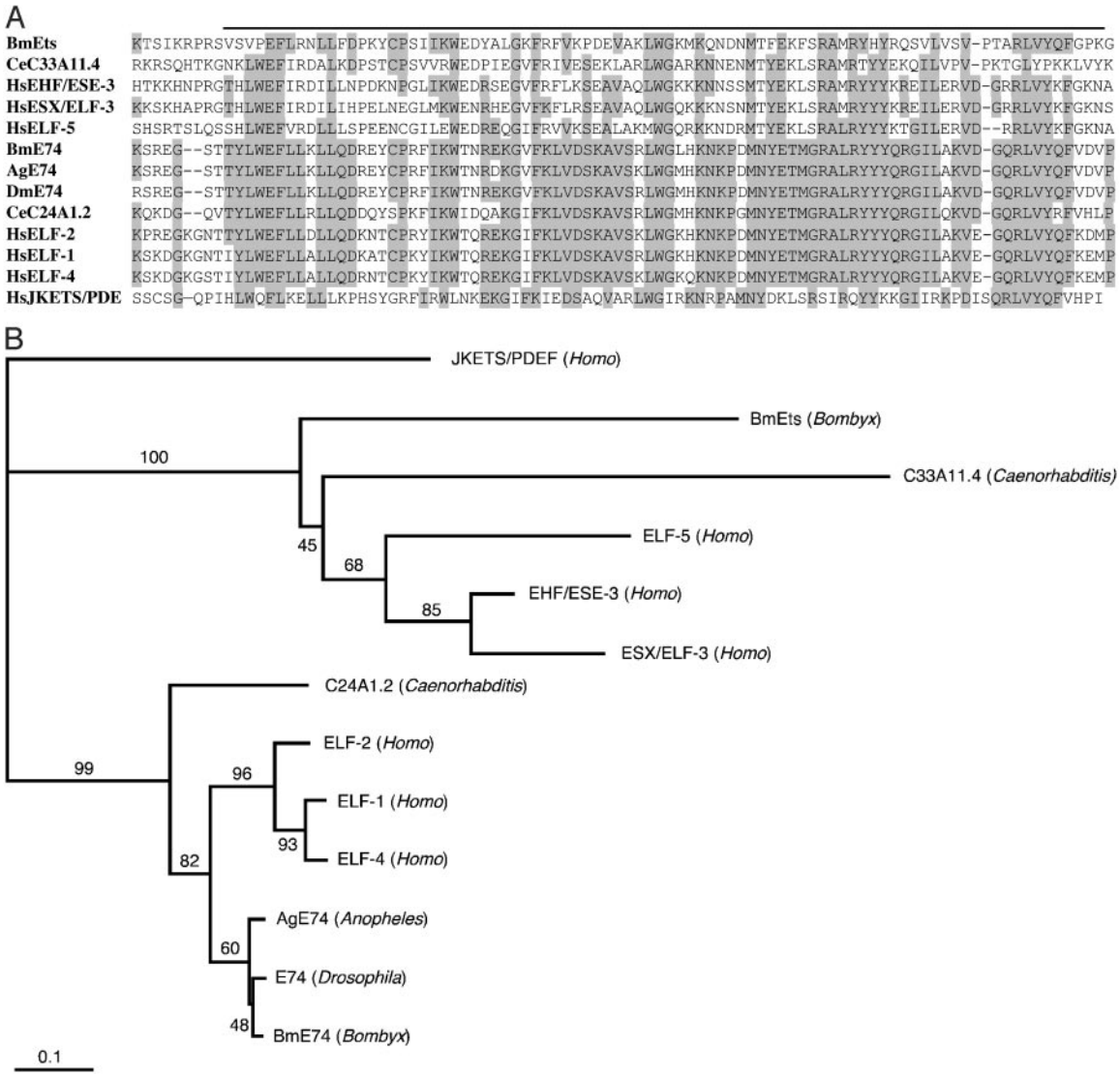

Fig. 1. (*A*) BmEts and its close relatives are aligned with the members of the E74 family. The bar indicates the ETS DNA-binding domain. The amino acid sequences were derived from the following: BtEts, SilkBase:wdV30217; CeC33A11.4, TrEMBL:Q93320; HsEHF/ESE-3, TrEMBL:Q9H509; HsEST/ELF-3, TrEMBL-new:CAD97611; HsELF-5, TrEMBL:O95175; BmE74, this work, DDBJ:AB114625; AgE74, GenBank:AAAB01008807; DmE74, TrEMBL:Q86NS8; CeC24A1.2, TrEMBL:O17057; HsELF-2, TrEMBL:Q15724; HsELF-1, TrEMBL:Q9EQY2; HsELF-4. TrEMBL:Q9Z2U4; and HsJKETS/PDE, PRF:2606391A. (*B*) A neighbor-joining tree of BmETS and its homologs. The branch lengths show the accepted point mutation distances calculated from the alignment in *A*. The numbers above the branches indicate the bootstrap support values in 100 replicates.

genes include repetitive sequences such as Bm1 (48) in their 3′ UTRs, which would interfere with measuring accurate levels of mRNAs by hybridization. Therefore, we designed and synthesized 6,000 specific primers located ≈500 base pairs downstream from the 5′ end of each cDNA to remove repetitive sequences from DNAs to be used for the microarray. Six thousand DNAs were amplified by PCR with a T3 vector primer and a specific primer, followed by spotting on glass slides. By using this procedure, we were able to obtain almost equal sizes and amounts of DNA from PCR, resulting in an even efficiency of DNA fixation.

EST microarray technology was successfully used to identify and isolate ecdysone-responsive cuticle protein genes in wing discs (49). We are also using microarrays to detect the changes of gene expression in wing discs during metamorphosis as a model to describe the gene cascade triggered by ecdysteroid hormone, and to extend our earlier studies (37) to elucidate the mechanism of baculovirus infection on host gene expression.

Lepidoptera-specific genes. Lepidopteran insects like the silkworm have taxonomically specific biological phenomena including sex determination, pheromone-dependent sexual communication, silk production, diapause, and interactions with plants and microbes. As highly destructive agricultural pests, Lepidoptera are among the most urgent targets for the design of insecticides that act effectively and specifically, but are harmless to other species and the environment. This result can be achieved most effectively by a better understanding of Lepidoptera-specific genes and their functions. Comparative studies on gene sequences between Lepidoptera, *Drosophila*, *Caenorhabditis elegans*, and other species for which genome analysis is well advanced can point to Lepidoptera-specific genes. From a comparison with Release 2 of the *Drosophila* genome annotation in FlyBase and WormBase Release wormpep56, we found that amino acid sequences encoded in several genes in *Bombyx* are more similar to those of bacteria than to *Drosophila* and other eukaryotes. One of the chitinase genes (*BmChi-h*) that has been mapped on *Bombyx* chromosome 7 was found to be orthologous to bacterial genes by phylogenetic analysis, which indicates that

BmChi-h was derived from the ancestral chitinase gene of a *Serratia*-like bacterium or a baculovirus (50). We found similar candidates for horizontal gene transfer from bacteria for a sucrase, a glucose-1-phosphatase and a glycerophosphodiesterase (GlpQ). For example, *Bombyx* GlpQ shares high similarities with eubacteria and a small number of plants in contrast with very low similarities to those of other eukaryotes, suggesting that *BmGlpQ* was acquired from a *Pseudomonas*-like bacterium by horizontal gene transfer (see Fig. 2, which is published as supporting information on the PNAS web site, www.pnas.org).

Gene comparison with FlyBase and WormBase can also lead to the discovery of the loss of an ortholog from *Drosophila* (13). For example, we found that the Ets transcription factor (BmEts) plays a critical role in embryonic diapause, which is a specific feature of silkworm development (51). Fig. 1 shows that apparently the *Drosophila* genome does not contain the ortholog of *BmEts*, but contains a paralog, *E74*, whereas the *Bombyx* and *C. elegans* genomes have both *Ets* and *E74* orthologs, indicating that *Drosophila* likely lost the ortholog of *BmEts* during its evolution.

Conclusion

An EST database in a model lepidopteran species will help to find gene sequences and gene functions not only in the model species itself but also in nonmodel lepidopterans. Although no complete lepidopteran genome has yet been sequenced, the silkworm EST database already contains ≈11,000 independent cDNAs, which, at the present rate of progress, could expand to cover >80% of total *Bombyx* genes within a couple of years. Plans to establish a full-length cDNA database that will provide essential data for accurate comparative genome analyses are warranted. The direct comparison of the silkworm EST database with FlyBase and WormBase, which has been added to SilkBase as an informatics tool, is a valuable and efficient approach for revealing, at the molecular level, what makes Lepidoptera different from other insects, and providing potential candidates for targets of Lepidoptera-selective insecticides.

We thank Shun-ichi Sasanuma, Yoshie Ishihara, and Izumi Matsumoto for technical assistance. This work was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences and the Animal Genome Research Program in the Ministry of Agriculture, Forestry, and Fisheries of Japan. It was also supported in part by Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research 13202074 (to K.O.) and 15011207 (to T.S. and M.G.S.), and The Japan Society for the Promotion of Science "Research for the Future" Program Life Science Grant 12-1 (to T.S.).

1. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. *et al.* (2000) *Science* **287,** 2185–2195.
2. Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., Wincker, P., Clark, A. G., Ribeiro, J. M., Wides, R., *et al.* (2002) *Scienc*e **298,** 129–149.
3. Promboon, A., Shimada, T., Fujiwara, H. & Kobayashi, M. (1995) *Genet. Res.* **66,** 1–7.
4. Yasukochi, Y. (1998) *Genetics* **150,** 1513–1525.
5. Shi, J., Heckel, D. G. & Goldsmith, M. R. (1995) *Genet. Res.* **66,** 109–126.
6. Kadono-Okuda, K., Kosegawa, E., Mase, K. & Hara, W. (2002) *Insect Mol. Biol.* **11,** 443–451.
7. Tan, Y. D., Wan, C., Zhu, Y., Lu, C., Xiang, Z. & Deng, H. W. (2001) *Genetics* **157,** 1277–1284.
8. Reddy, K. D., Abraham, E. G. & Nagaraju, J. (1999) *Genome* **42,** 1057–1065.
9. Heckel, D. H. (2003) *Annu. Rev. Entomol.* **48,** 236–260.
10. Rubin, G. M., Hong, L., Brokstein, P., Evans-Holm, M., Frise, E., Stapleton, M. & Harvey, D. A. (2000) *Science* **287,** 2222–2224.
11. Stapleton, M., Liao, G., Brokstein, P., Hong, L., Carninci, P., Shiraki, T., Hayashizaki, Y., Champe, M., Pacleb, J., Wan, K., *et al.* (2002) *Genome Res.* **12,** 1294–1300.
12. Dimopoulos, G., Casavant, T. L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., Schultz, J., Benes, V., Bork, P., Ansorge, W., *et al.* (2000) *Proc. Natl. Acad. Sci. USA* **97,** 6619–6624.
13. Whitfield, C. W., Band, M. R., Bonaldo, M. F., Kumar, C. G., Liu, L., Pardinas, J. R., Robertson, H. M., Soares, M. B. & Robinson, G. E. (2002) *Genome Res.* **12,** 555–566.
14. Gaines, P. J., Brandt, K. S., Eisele, A. M., Wagner, W. P., Bozic, C. M. & Wisnewski, N. (2002) *Insect Mol. Biol.* **11,** 299–306.
15. Theodorides, K., De Riva, A., Gomez-Zurita, J., Foster, P. G. & Vogler, A. P. (2002) *Insect Mol. Biol.* **11,** 467–475.
16. Robertson, H. M., Martos, R., Sears, C. R., Todres, E. Z., Walden, K. K. O. & Nardi, J. B. (2000) *Insect Mol. Biol.* **8,** 501–518.
17. Morimyo, M., Mita, K., Hongo, E., Higashi, T., Sugaya, K., Ajimura, M., Yamauchi, M., Tsuji, S., Park, W.-Y., Sasanuma, S., *et al.* (1997) in *Biodefence Mechanisms Against Environmental Stresses*, eds. Ozawa, T., Hori, T. & Tatsumi, K. (Kodansha Scientific, Springer, Tokyo), pp. 115–123.
18. Marra, M. A., Hillier, L. & Waterston, R. H. (1998) *Trends Genet.* **14,** 4–7.
19. Grace, T. D. (1967) *Nature* **216,** 613.
20. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
21. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994) *Nucleic Acids Res.* **22,** 4673–4680.
22. Chareyre, P., Guillet, C., Besson, M. T., Fourche, J. & Bosquet, G. (1993) *Insect Mol. Biol.* **2,** 239–246.
23. Fletcher, J. C. & Thummel, C. S. (1995) *Development (Cambridge, U.K.)* **121,** 1411–1421.
24. Soares, M. B., Bonaldo, M. F., Jelene, P., Su, L., Lawton, L. & Efstratiadis, A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 9228–9232.
25. Bonaldo, M. B., Lennon, G. & Soares, M. B. (1996) *Genome Res.* **6,** 791–806.
26. Gene Ontology Consortium (2000) *Nat. Genet.* **25,** 25–39.
27. Gene Ontology Consortium (2001) *Genome Res.* **11,** 1425–1433.
28. Quan, G.-X., Mita, K., Okano, K., Shimada, T., Ugajin, N., Zhao, X., Goto, N., Kanke, E. & Kawasaki, H. (2000) *Insect Biochem. Mol. Biol.* **31,** 97–103.
29. Zhao, X., Mita, K., Shimada, T., Okano, K., Kanke, E. & Kawasaki, H. (2001) *Mol. Biol.* **31,** 1213–1219.
30. Takeda, M., Mita, K., Quan, G.-X., Shimada, T., Okano, K., Kanke, E. & Kawasaki, H. (2001) *Insect Biochem. Mol. Biol.* **31,** 1019–1028.
31. Obayashi, F., Suzuki, M. G., Mita, K., Okano, K. & Shimada, T. (2001) *Comp. Biochem. Physiol. B Biochem. Mol. Biol.* **128,** 145–158.
32. Suzuki, M. G., Ohbayashi, F., Mita, K. & Shimada, T. (2001) *Insect Biochem. Mol. Biol.* **31,** 1201–1211.
33. Hashimoto, H. (1933) *Jpn. J. Genet.* **8,** 245–258.
34. Cline, T. W. (1993) *Trends Genet.* **9,** 385–390.
35. Yoshiga, T., Okano, K., Mita, K., Shimada, T. & Matsumoto, S. (2000) *Gene* **246,** 339–345.
36. Matsumoto, S., Yoshiga, T., Yokoyama, N., Iwanaga, M., Koshiba, S., Kigawa, T., Hirota, H., Yokoyama, S., Okano, K., Mita, K., *et al.* (2001) *Insect Biochem. Mol. Biol.* **31,** 603–609.
37. Okano, K., Shimada, T., Mita, K. & Maeda, S. (2001) *Virology* **282,** 348–356.
38. Kawasaki, H., Sugaya, K., Quan, G.-X., Nohata, J. & Mita, K. (2003) *Insect Biochem. Mol. Biol.* **33,** 131–137.
39. Nirmala, X., Mita, K., Vanisree, V., Zurovec, M. & Sehnal, F. (2001) *Insect Mol. Biol.* **10,** 437–445.
40. Gan, H., Wang, Y., Jiang, H., Mita, K. & Kanost, M. R. (2001) *Insect Biochem. Mol. Biol.* **31,** 887–898.
41. Landais, I., Pommet, J.-M., Mita, K., Nohata, J., Gilmenez, S., Fournier, P., Devauchelle, G., Duonor-Cerutti, M. & Ogliastro, M. (2001) *Gene* **271,** 348–356.
42. Lee, J., Hahn, Y., Yun, J. H., Mita, K. & Chung, J. H. (2000) *Biochim. Biophys. Acta* **1491,** 355–363.
43. Gahan, L. T., Gould, F. & Heckel, D. G. (2001) *Science* **293,** 857–860.
44. Ichimura, S., Mita, K., Zama, M. & Numata, M. (1985) *Insect Biochem.* **15,** 277–283.
45. Osoegawa, K., Woon, P. Y., Zhao, B., Frengen, E., Tateno, M., Catanese, J. J. & deJong, P. J. (1998) *Genomics* **52,** 1–8.
46. Gage, L. P. (1974) *Chromosoma* **45,** 27–42.
47. Rasch, E. M. (1985) in *Advances in Microscopy*, eds. Cowden, R. R. & Harrison, S. H. (Liss, New York), pp. 137–166.
48. Adams, D. S., Eickbush, T. H., Herrera, R. J. & Lizardi, P. M. (1985) *J. Mol. Biol.* **187,** 465–478.
49. Noji, T., Ote, M., Takeda, M., Mita, K., Shimada, T. & Kawasaki, H. (2003) *Insect Biochem. Mol. Biol.* **33,** 671–679.
50. Daimon, T., Hamada, K., Mita, K., Okano, K., Suzuki, M. G., Kobayashi, M. & Shimada, T. (2003) *Insect Biochem. Mol. Biol.* **33,** 749–759.
51. Suzuki, M. G., Terada, T., Kobayashi, M. & Shimada, T. (1999) *Insect Biochem. Mol. Biol.* **29,** 339–347.