Silkworm genomics – progress and prospects

J. Nagaraju*,† and M. R. Goldsmith**

*Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad 500 076, India
**Department of Biological Sciences, University of Rhode Island, Kingston, RI, 02881-0816, USA

The biology and genetics of silkworm, *Bombyx mori*, is the most advanced of any lepidopteran species. Its rich repertoire of genetic resources and potential applications in sericulture and as a model for other Lepidoptera led to the initiation of genomics research. During the past decade much effort has been made in the areas of marker development, and molecular maps have been constructed in standard strains with the use of RFLPs, RAPDs, ISSRs, STSs, and microsatellites. The potential applications of molecular markers and linkage maps include stock identification, Marker Assisted Selection (MAS), identification of Quantitative Trait Loci (QTL), and, ultimately, positional cloning of visible mutations and QTL. To these ends, BAC libraries have been constructed and are being used to make large-scale physical maps, with markers based on ESTs as framework anchors. Altogether this work provides a foundation for identification of gene function, gene and chromosome evolution, and comparative genomics.

THE silkworm, *Bombyx mori*, domesticated for silk production for about 5000 years is the most well-studied lepidopteran model system because of its rich repertoire of well characterized mutations affecting virtually every aspect of the organism's morphology, development, and behaviour and its considerable economic importance.

By virtue of a long history of silkworm rearing for commercial purpose, silkworm genetics has been the subject of considerable research interest resulting in careful collection, cataloguing and maintenance of various silkworm genetic stocks of considerable scientific and economic interest. Today, the opportunities for genetic manipulation and study of the silkworm, *B. mori* include more than 400 visible mutations out of which 200 have been assigned to conventional linkage groups covering 900.2 cM (ref. 1). These mutations affect many fundamental aspects of the insect's life cycle, including egg and egg shell formation, early embryonic development and pattern formation, larval feeding behaviour, molting, embryonic diapause, etc. In addition, a vast array of distinct geographical races and inbred lines that represent variation for a number of qualitative and quantitative traits of basic biological and economic interest including

body size, silk quality, fecundity, pathogen resistance, and heat tolerance are available. In fact, the silkworm *B. mori* is genetically the best known insect next only to the fruitfly, *Drosophila melanogaster*. The haploid genome size of *B. mori* is estimated to be 530 Mb (ref. 2), 3.8 times that of *D. melanogaster* and one-sixth the size of the mammalian genome.

Supported by the infrastructure of data and genetic resources available for research, the silkworm is a key model organism in Lepidoptera which includes more than 160,000 species, of which Bombycoid moths include silkmoths of economic importance and Noctuid moths, the largest group in the order and includes some of the most devastating pests of agriculture, particularly *Heliothis* (*H. virescens*, *H. armigera*, *H. zea*). The genomic information of the model species *B. mori* should be applicable to the most important species in Lepidoptera. Comparative studies on genome sequences between Lepidoptera and other published genome sequences could uncover lepidoptera-specific genes. Products of such novel genes could serve as targets for lepidopteran-specific insecticides.

Bombycoid moths secrete diverse varieties of silk fibres. These species include *B. mori* of family Bombycidae and wild silkmoths that belong to Saturniidae, *Antheraea mylitta* (Indian tropical tasar silkmoth), *A. pernyi* (Chinese oak tasar silkmoth), *A. assama* (Indian golden silkmoth), *A. yamamai* (Japanese oak silkmoth) and *Philosamia cynthia ricini* (Indian castor silkmoth). Silk production based on these moths, specially *B. mori*, *A. pernyi*, *A. mylitta* and *A. assama* plays important role in rural economies of many populous developing nations. Six million people in India alone are involved in sericulture, which involves intensive labour and provides the key to improving local quality of life. In order to make sericulture economically viable, genes affecting growth rate, yield, fibre quality, virus resistance can be tagged with molecular markers for rapid construction of genetically improved strains. Considering the unique experimental advantages of this organism and its economic importance, an International Consortium on Lepidopteran Genomics was recently formed to promote international cooperation to sequence the genome of *B. mori* and to undertake comparative genomics of other economically important Lepidoptera[3]. Such international cooperation is expected to foster knowledge both in the basic and app-

†For correspondence. (e-mail: jnagaraju@www.cdfd.org.in)

    415

lied aspects of insect science. The current efforts of several laboratories contributing to the silkworm genome project have, as their immediate objective, the generation of high density molecular linkage map, development of tools and techniques to 'fingerprint' silkworm varieties, identification of lepidoptera-specific genes, and generation of a clone-based physical map that will allow virtually immediate access to any defined region of the euchromatin. The long-term objective is to identify the position, DNA sequence and function of all the genes in the organism. These are achievable goals considering the revolutionary developments in analytical tools and technical feasibility of large-scale DNA sequencing as demonstrated recently in complex genomes such as *Drosophila* and human[4–6].

In this review, we summarize the results accrued in various genome analysis projects now underway in several laboratories in the area of molecular genetics and genomics of the silkworm.

Genetic markers

In the past decade, several key advances in molecular biology have greatly advanced the impact of molecular genetics in biology. Most important have been: (i) development of the polymerase chain reaction (PCR), which amplifies specific stretches of DNA to usable concentrations, (ii) the application of evolutionarily conserved sequences as PCR primers, (iii) the advent of hypervariable microsatellite loci as genetic markers and (iv) the advent of routine DNA sequencing in biology laboratories. These innovations, coupled with the recent explosion of powerful analysis and relatively user-friendly computer programs, have accelerated the pace at which molecular genetic data can be used for various programmes. These developments led to the detection, analysis and exploitation of naturally occurring DNA sequence polymorphisms in eukaryotes and laid a foundation for contemporary genomics of several organisms. Polymorphic genetic markers have found potential applications in animal and plant improvement programmes as a means for varietal and parentage identification, construction of molecular linkage maps and evaluation of polymorphic genetic loci affecting quantitative economic traits.

A wide array of DNA marker techniques is available for genetic studies. All DNA markers reflect differences in DNA sequences. Typically, in a diploid organism, each individual can have one or two different states (alleles) per character (locus). The choice of a particular genetic marker often depends upon the purpose of the study and is usually a trade-off between practicality and precision of genetic markers. One manifestation of this is the dichotomy between multilocus DNA techniques which include Random Amplified Polymorphic DNAs (RAPDs), Amplified Fragment Length Polymorphisms

(AFLPs) and Inter-Simple Sequence Repeat PCR markers (ISSR-PCR) and single locus techniques which include single copy nuclear DNA regions and microsatellites. Multilocus methods result in simultaneous visualization of many dominant markers and are technically convenient, but have some marked weaknesses and limitations, including dominant inheritance – the DNA fragment can be scored only as present or absent, and in many cases substantial proportion of variation they detect can be non-heritable or not even derived from the target organism. By contrast, single locus markers are quite informative because they provide alleles whose zygosity status can be assayed easily.
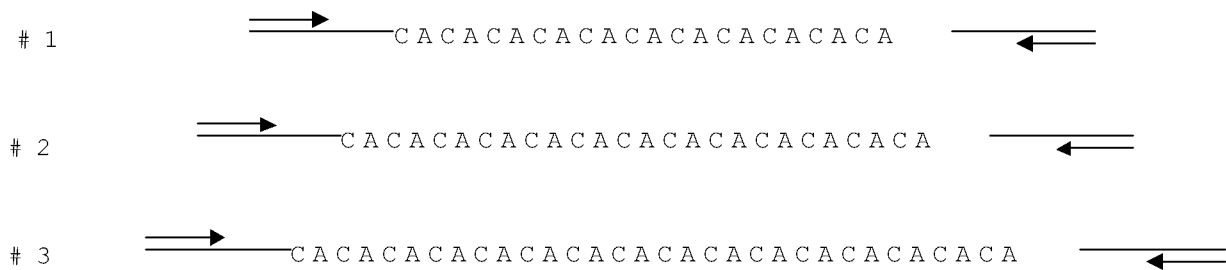
Development of molecular markers is important in the silkworm for construction of linkage map, fingerprinting of strains for breeding, and marker-assisted selection. In the earlier studies, fingerprinting analyses of silkworm strains have been carried out using RAPDs[7], heterologous minisatellite probes, Bkm(2)8 (refs 8, 9), M13 (ref. 10), and ISSR-PCR (ref. 11). Genetic linkage maps of silkworm have been constructed using RAPDs[12–14], RFLPs[15], and AFLPs[16]. Most of these markers, except RFLPs, are dominant and strain-specific.

Considering the genetic informativeness offered by microsatellite loci, which are co-dominant, a microsatellite discovery programme in the silkworm genome was initiated[21]. Microsatellites, also known as simple sequence repeats (SSRs), are short stretches of nucleotides in which a motif of one to six bases is tandemly repeated, have emerged as ideal markers in eukaryotic genetics. They are ubiquitous in prokaryotic and eukaryotic genomes and are randomly distributed, both in protein coding and non-coding regions. The sequences in microsatellite motifs can differ in repeat number among individuals, providing a ready source of polymorphism. With the advent of the PCR, this property of microsatellite DNA was converted into highly versatile genetic markers[17–19]. PCR products of different lengths can be amplified using primers flanking the variable microsatellite region (Figure 1). Due to the advantages offered by microsatellites they have been used to construct linkage maps[20]. In addition to their clear utility for practical applications, information about the distribution and variability of microsatellite sequences in the genome of a given species can help elucidate its genetic history from the standpoint of evolution and artificial selection.

Microsatellite motifs in the silkworm genome

In a previous study[21], analysis of microsatellite loci cloned and characterized from a *B. mori* sub-genomic library was reported. The total fraction of the genome analysed (2810 kb), which accounts for 0.53% of the *Bombyx* genome (530 Mb) yielded 57 $(CA)_n/(GT)_n$ and

Alleles

1   ⟹──────C A C A C A C A C A C A C A C A C A ─────⟸

2   ⟹──────C A C A C A C A C A C A C A C A C A C A ─────⟸

3   ⟹──────C A C A C A C A C A C A C A C A C A C A C A ─────⟸

Genotypes

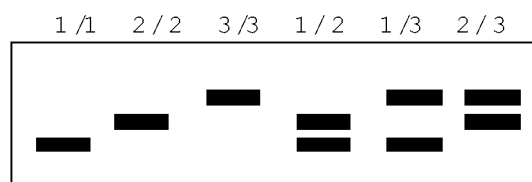1 /1   2 / 2   3 /3   1 / 2   1 /3   2 / 3

Figure 1. Microsatellite alleles and their detection. Three alleles of different sizes at a microsatellite locus composed of a (CA) repeat are shown. The arrows indicate the locus-specific primers designed for the region flanking the (CA) repeat, used for PCR amplification. Illustrations of the various gel patterns that would be observed with different allele combinations are indicated.

Table 1. Percentage of different categories of $(GT)_n$ and $(CT)_n$ microsatellites in silkworm, honey bee (Estoup, A. *et al.*[42]) and human (Weber, 1990). Number in parenthesis is the number of $(GT)_n$ or $(CT)_n$ microsatellites analysed

	Silkworm		Honey bee		Human
Type of motif	GT (21)	CT (7)	GT (23)	CT (52)	GT (57)
Perfect	62	71	48	46	61
Imperfect	29	29	22	31	27
Compound	9	0	30	23	2

27 $(GA)_n/(CT)_n$ harbouring clones. The frequency of occurrence of dinucleotide repeats, was estimated to be, on average, one $(GT)_n$ every 49 kb, and one $(CT)_n$ microsatellite every 104 kb. The frequencies of occurrence of these two types of microsatellite motifs compare well with those reported in honey bee, pig, rat and human. The perfect (no interruption in the run of repeats), imperfect (interruption in the run of repeats) and compound (a mixture of repeats of two or more motifs) repeat microsatellites in the silkworm genome were on the order of 62%, 29% and 7% respectively (Table 1). This is in close agreement with 72% perfect, 28% imperfect, and 11% compound repeats in human[22] (Table 1).

The primer pairs developed for the flanking sequences of the microsatellite loci isolated from the silkworm genome successfully amplified the repeats producing a vari-

ety of co-dominant markers. This indicated that the microsatellite loci were conserved and not generally derived from transposable elements, which are expected to be mobile, giving rise to null alleles. A summary of the microsatellite loci and PCR amplification conditions for sets of representative microsatellite loci are given in Table 2. The polymorphisms, which manifested as size differences in the PCR products, could be easily scored across silkworm strains (Figure 2). For example, the 15 microsatellite loci, 11 $(GT)_n$ and 4 $(CA)_n$, revealed a total of 113 alleles in 13 silkworm strains. The number of alleles scored at each locus in these strains varied from as few as 3 to as many as 17. The $(CA)_n$ motifs showed a tendency to reveal a greater number of alleles compared to $(GA)_n$ motifs. For example, 11 $(GT)_n$ loci revealed a total of 85 alleles, ranging from 3 to 17, with an average of 7.7 alleles per locus whereas the four $(CT)_n$ loci revealed a total of 28 alleles ranging from 5 to 9 with an average of 7 alleles per locus.

Heterozygosity values

All 15 microsatellite loci revealed allelic length polymorphism in the 13 silkworm strains analysed. The average heterozygosity value was 0.79, ranging from 0.66 to 0.90. The most polymorphic locus, *sat2763*, which revealed 17 alleles, also showed the highest heterozygosity

Table 2. Microsatellite loci, repeat pattern, forward and reverse flanking primer sequences, number of alleles, allelic range, heterozygosity values and annealing temperature (T) in *Bombyx mori*

Locus symbol	Repeat motif length	Primer sequence Forward	Primer sequence Reverse	No. alleles	Allele size range (bp)	Hetero-zygosity	T (°C)	MgCl₂ (mM)
Sat158	$(CA)_9(CA)_4$	5'-cttccagacaaccattagat-3'	5'-cagcatccatccttatttat-3'	11	190–246	0.76	49	1
Sat211	$(GT)_{10}$	5'-gatcgaaactacgcattacg-3'	5'-cagcatccatccttatttat-3'	6	106–146	0.89	49	2
Sat256	$(CA)_5(TA)_7$	5'-attgcctcggacaggagtgt-3'	5'-tgtgaggacaggccgtta-3'	7	111–194	0.82	53	3
Sat346	$(GA)_9$	5'-gaagacagagcgaagtgga-3'	5'-atggattcctgctggtagat-3'	6	139–189	0.83	53	3
Sat892	$(GT)_{10}$	5'-caataaatgcttacgagtttaa-3'	5'-tatcggtagttccttgactt-3'	3	175–187	0.66	47	3
Sat951	$(GA)_{23}$	5'-attgtaaccgatttgagaga-3'	5'-attcgcacaataagttcact-3'	5	107–122	0.68	48	1
Sat962	$(GT)_{10}$	5'-cataataaataaataataat-3'	5'-tgtagtgggatagtcagtat-3'	6	108–114	0.78	42	3
Sat1013	$(GT)_9$	5'-aacagatgctgcggactggt-3'	5'-tgccattcacaatacaacat-3'	5	135–162	0.80	50	1
Sat1411	$(GT)_8(GT)_5$	5'-gaatgttctgctggtgg-3'	5'-taatgttttatactttattatatg-3'	8	109–162	0.68	45	3
Sat1423	$(CA)_{11}$	5'-ctttcgatcaccgcgttctc-3'	5'-cgctacgaaataccattatctgaca-3'	9	130–176	0.82	55	2
Sat1893	$(CA)_{10}$	5'-aatgcagaatcgtaatttt-3'	5'-tttgaccacagacaataagc-3'	7	98–158	0.85	45	2
Sat2550	$(GA)_{12}$	5'-ggtcccttgaaactgcgat-3'	5'-cagagacctgccggttgtcgttc-3'	9	121–168	0.87	53	2
Sat2604	$(CT)_9$	5'-gctcgccatatgcaatcctc-3'	5'-cgtcattgccttcatttcagttc-3'	8	143–186	0.86	53	2
Sat2763	$(GT)_{22}(GT)_7$	5'-acgcgctctacaaaataccatta-3'	5'-gatcaccggttctgctctcg-3'	17	105–179	0.90	53	2
Sat3513	$(GT)_5(TGCG)_6(TA)_2$	5'-cgcaattctgtattagataa-3'	5'-taaaggtattattcttattcg-3'	6	134–223	0.84	46	3
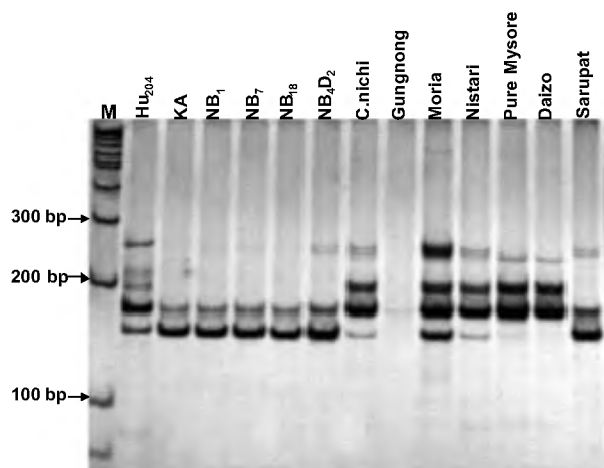
Figure 2. Allelic variations detected by a microsatellite locus (Bmestsat10) in different silkworm strains.

of 0.90. This locus harboured two stretches of (GT) motifs, one with 22 repeats and another with 7 repeats. The locus *sat892*, with a $(GT)_{10}$ motif that revealed only 3 alleles in the 13 silkworm populations, showed the lowest heterozygosity values. There was no clear evidence with regard to the relationship between the repeat length and degree of polymorphisms. For example, locus *sat951* $(GA)_{23}$, revealed 5 alleles with a heterozygosity value of 0.68, and the locus *sat2604* with $(GA)_9$ showed 8 alleles with a heterozygosity value of 0.86. The loci that were similar in length also differed in their number of alleles and heterozygosity. The loci *sat211*, *sat892*, *sat962* and *sat1893*, which carry repeat motifs of $(GT)_{10}$, showed 6, 3, 6, and 7 alleles respectively. These results are in contrast to the reports on human and other species where longer repeats have been reported to generate a higher degree of polymorphism[22,23].

Genetic diversity studies

A variety of molecular techniques have been employed successfully in silkworm to detect genome-wide polymorphisms. These methods include Restriction Fragment Length Polymorphisms and a variety of PCR-based techniques. In a study reported by Nagaraju *et al.*[8] and Sharma *et al.*[9], a banded krait minisatellite, Bkm(2)8 containing a 545 bp DNA sequence consisting of 66 copies of GATA repeats interspersed with a variable number of dinucleotide (TA) repeats in several locations, was used as a probe. The probe revealed intra- and inter-population genetic diversity when tested on 13 divergent silkworm genotypes. The phenetic analysis of RFLP data separated the 13 strains into three groups: one consisted of all dia-pausing strains (except HU₂₀₄); the second consisted of all non-diapausing strains (except Pure Mysore and Nistari); and the third consisted exclusively of Pure Mysore and Nistari (Figure 3). In another study, six anonymous multilocus probes obtained from a *Pst*I subgenomic silkworm library, were used on the same set of 13 silkworm strains[24]. The genetic informativeness expressed in terms of number of loci revealed (effective multiplex ratio, EMR) and amount of polymorphism detected (diversity index, DI) was compared with the three PCR-based techniques, Random Amplified Polymorphic DNA (RAPD), Inter-Simple Sequence Repeat PCR (ISSR-PCR) and simple sequence repeats (SSRs). The six multilocus RFLP probes produced 180 products of which 97% were polymorphic; the 15 SSR loci gave rise to an average of 8 alleles each, of which 86% were polymorphic; and ISSR-PCR produced 39 fragments of which 77% were polymorphic (Table 3). The highest diversity index was observed for ISSR-PCR (0.957) and the lowest (0.744) for RAPDs (Table 4). The RAPDs, ISSR-PCR
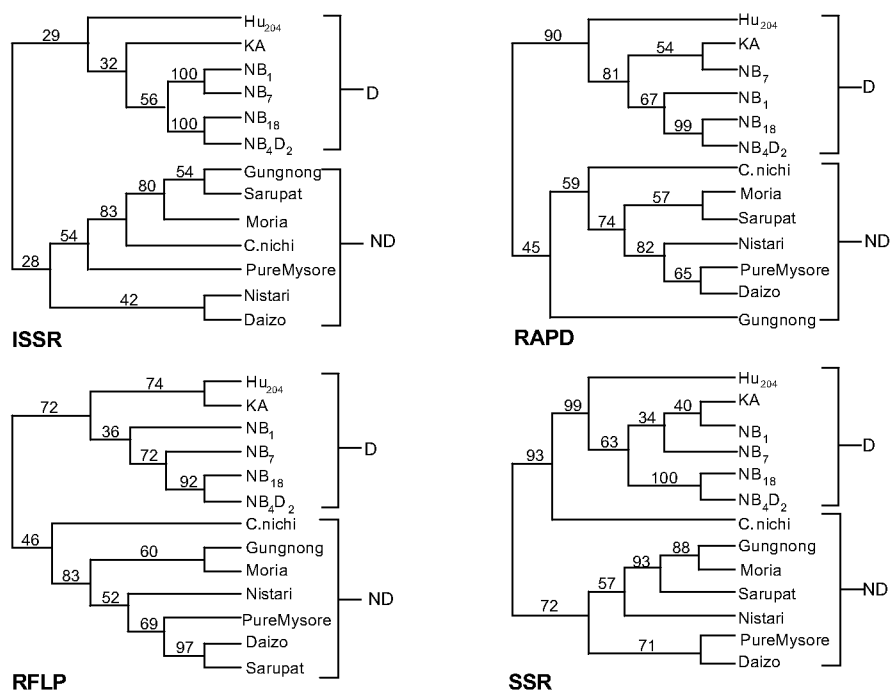
Figure 3. Dendrograms generated from genetic similarity matrices calculated using Nei and Li coefficients for ISSR-PCR, RFLP, RAPD and SSR marker assays. D, diapausing silkworm strains; ND, non-diapausing silkworm strains. Numbers on the nodes indicate the number of times a particular branch was recorded per 100 bootstrap replications following 1000 replications.

Table 3. A summary of each type of marker assay performed on 13 silkworm strains

	Total number of assays	Total no. products	Total poly-morphic products	Mean no. products per assay	Percentage of poly-morphic products
SSR	15 (primer pairs)	122	105	8.10	86.00
Inter-SSR	6 (primers)	239	184	39.83	76.98
RAPD	40 (primers)	216	204	5.40	94.44
RFLP	6 (probes)	180	176	30.00	97.77

Table 4. A summary of mean DI, EMR and MI for different marker assays of silkworms

Marker assay	Mean diversity index (DI)	Mean effective multiplex ratio (EMR)	Marker index (MI)
SSR	0.795	1.0	0.796
Inter-SSR	0.957	23.6	22.570
RAPD	0.744	4.8	3.580
RFLP	0.930	28.7	26.670

$DI = 1 - \sum p i^2$, where pi is the allele frequency of the ith allele.
$EMR = n_p(n_p/n)$, where n_p is the number of polymorphic loci and n is the total number of loci.
$MI = DI \times EMR$.

and multilocus RFLPs clearly separated the diapausing and non-diapausing silkworm varieties, as did the Bkm(2)8 probe (Figure 3).

The analysis of the same set of 13 silkworm strains using 5 different marker systems highlighted the utility of each of the marker systems in genetic analysis of silkworm. These results show that ISSR-PCR and multilocus RFLP markers are powerful techniques for fingerprinting silkworm varieties. Although most ISSR loci are dominant rather than co-dominant, ISSR-PCR markers offer several advantages over RFLPs for genotyping, the major one being the rapid production of a large number of markers in a cost-effective manner. Because of these advantages, the ISSR-PCR technique has potential use in silkworm breeding, germplasm evaluation, and genetic mapping. The recently automated ISSR-PCR technique by including a fluorescent dye in the PCR reaction followed by analysis on an ABI automated sequencer[25] has been shown to be ideal for high-resolution mapping experiments and DNA profiling (see the later section of this article). Although multilocus RFLP probes offer features

such as high reproducibility and easy scoring, the requirement of a large quantity of template DNA, and Southern hybridization render them less suitable for large-scale studies. In the RAPD technique, though widely used, many factors such as concentration of Mg^{2+}, quality of the template, thermal cycler and the source of polymerase have been reported to affect the repeatability of the results between different laboratories. In this aspect, RAPD analysis may not be a method of choice for large-scale genotyping experiments. Considering the reproducibility, genetic informativeness and speed, SSRs are the most useful markers among all the marker systems. Since SSRs are co-dominantly inherited, multiallelic in nature, and can be multiplexed (5–6 loci can be analysed on a single lane) and automated, they are useful for genetic diversity, pedigree evaluation and genetic mapping studies. Although the initial cost to develop locus-specific primers is high, the advantages that they are robust and are amenable to automation ultimately increase the cost-benefit ratio, particularly considering the value of being able to compare data sets directly.

Cross-amplification of *Bombyx* microsatellites in Saturniid silkmoths and Noctuid moths

Advances in lepidopteran genomics have been limited by general lack of genetic information for individual species because of lack of genetic tools with broad applicability across species[26]. An accessible approach to such a problem is to look for chromosome conservation across species by using comparative genome mapping. Recent results[27] (Nagaraju et al., under preparation) show that many *Bombyx* microsatellite loci are conserved across Saturniid silkmoths which include most of the commercially important wild silkmoths such as *Antheraea assama* (Muga silkworm), *A. mylitta* (Tropical tasar silkworm), and *A. pernyi* (Temperate tasar silkworm), as well as a Noctuid moth, *Helicoverpa armigera,* a serious agricultural pest. Considering that the genetics of these moths has hardly been studied, the microsatellite development programme in *B. mori* makes it possible to carry out molecular genetic analyses of these moths without investing in a microsatellite development programme for each species. The use of heterologous primers will significantly reduce the cost of developing markers for the wild silkmoths and agricultural pest, *Heliothis* spp. Besides, as mapping is already underway in *Heliothis*[28,29], a comparative mapping approach will enable the prediction of the presence of microsatellites associated with genes of interest. Such conserved loci could be potentially used for assessing the evolutionary relationships within and across species with a reasonable degree of accuracy.

Linkage maps

The first attempt to construct molecular linkage maps for the silkworm utilized RFLPs generated by 15 character-
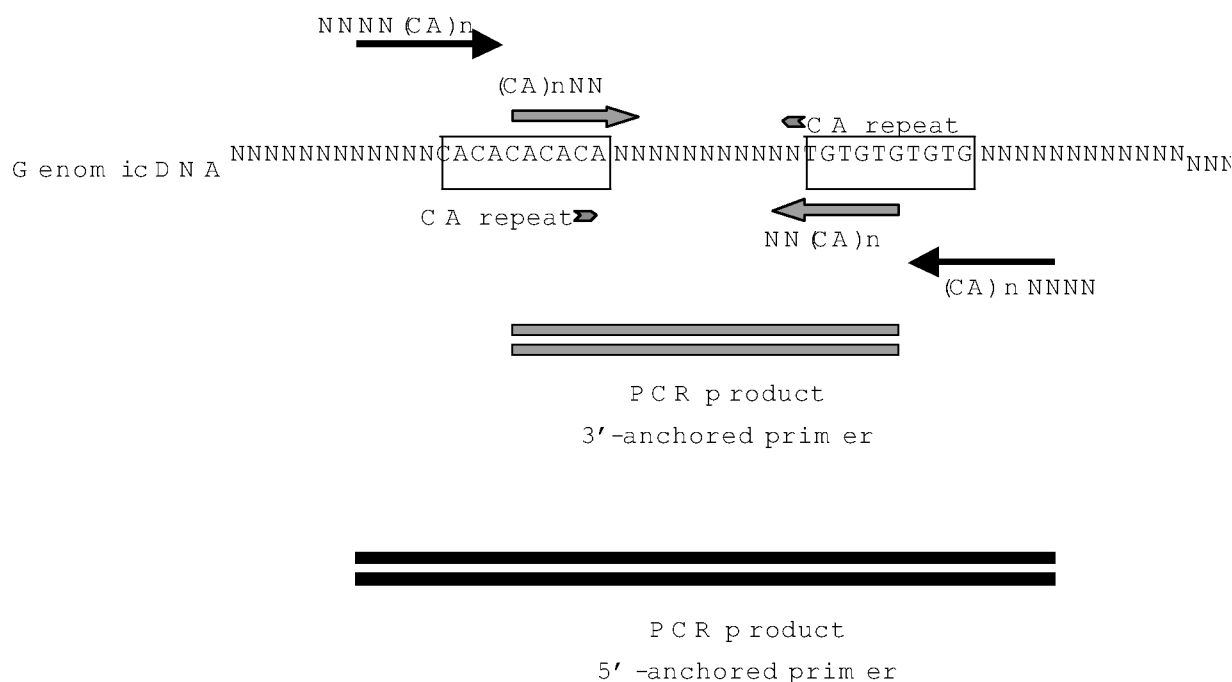

Figure 4. Inter-Simple Sequence Repeat-PCR (ISSR-PCR). A single primer targeting a $(CA)_n$ repeat anchored either at the 3′ (teal arrows) or at the 5′ end (yellow arrows) of the repeat, is used to amplify genomic sequence flanked by two inversely oriented $(CA)_n$ repeats (adapted from Zietkiewicz et al.[41]).

Figure 5. The Mendelian segregation of Fluorescent-ISSR-PCR (FISSR-PCR) markers in silkworm. The FISSR-PCR markers were generated using a primer 5' RAY RAT RC (GA)₇ 3' on two parental strains, p50 and C108, and their F₁ and F₂ offspring. The arrows and arrowheads indicate markers specific to p50 and C108, respectively.

ized single copy sequences, 36 anonymous sequences derived from a follicular cDNA library, and 10 loci corresponding to a low copy number retrotransposon, *mag*[15]. Recently, this work was extended using as probes partially sequenced cDNAs (ESTs) derived from an embryonic and a maternal mRNA library. By using the classic backcross strategy to detect linkage (see below), a small number of segregants (e.g., 15 animals) from backcrosses of single heterozygous females derived from strains p50 and J02 to inbred J02-derived males have been used to assign more than 200 codominant markers to 27 linkage groups and one independent marker. Integration of these linkage groups with the conventional genetic maps of morphological and biochemical markers is underway, together with the reciprocal backcrosses designed to assign recombination distances within each linkage group (W. Hara, H. Banno, H. Fuji, pers. commun.).

An RAPD linkage map of silkworm has been constructed using 320 random primers, which produced 243 polymorphic products between two parental silkworm strains, C108 and P50. In the F₂ population of 101 individuals used for these maps, segregation ratios of 168 bands were nearly 3 : 1 in chi square tests, showing Mendelian inheritance[12]. The MAPMAKER program sorted 168 bands into 29 linkage groups. The sum of map distances is approximately 900 cM. Recently, Nagaraja[14] has constructed a Z-chromosome RAPD linkage map using a backcross population (BC) derived from crossing an F₁ (male) heterozygous for a recessive sex-linked marker, transluscent larval skin (*od*) × a homozygous translucent (*od*) (female).

In silkworm, both F₂ (Figure 6) and backcross populations (Figure 7) have been used by different groups. Four groups have used an F₂ population derived from a cross of the same two reference parental strains, C108 and P50. Although both strains are of Chinese origin, they exhibit high phenotypic diversity for such complex characters such as size, growth rate, diapause, nutritional requirements, general vigour and cocoon properties such as cocoon weight, cocoon shell weight, and silk fibre length,

suggesting that considerable polymorphism exists at DNA level. Different groups have used a different number of F₂ offspring for mapping, affecting the ability to detect linkage between distal markers. Shi *et al.*[15] have used 52 F₂ progeny for RFLP maps, whereas Promboon *et al.*[12] and Yasukochi[13] have used 101 and 166 F₂ progeny, respectively, for maps based on RAPDs. The SSR and Fluorescent ISSR (FISSR) markers have been mapped recently using the same 101 F₂ offspring used by Promboon *et al.*[12]. (Muthulakshmi, Kathirvel, Selvendran and Nagaraju, under preparation). The FISSR markers are particularly advantageous for high throughput genotyping in mapping of genomes[25] (Figures 4 and 5). For constructing high-density linkage maps, a large number of F₂ individuals need to be analysed using a number of markers. The template requirement in the case of other conventional PCR methods is relatively high. The FISSR assay provides a large number of DNA markers per primer and allows detection of markers with as little as 2–5 ng of template DNA, on an automated sequencer obviating the necessity for using radioactive isotopes.

Backcross populations have been used for mapping by three other groups[14,16]. Nagaraja[14] has used a polyvoltine strain, Pure Mysore, whose larval skin is normal white (+/+), and a bivoltine strain (*od/od*), whose larval skin is translucent, controlled by a recessive gene, *od*, located on the Z chromosome. Fifty five progeny were used for mapping. Tan *et al.*[16] have used a similar back cross with 47 offspring. Similarly, Hara and coworkers have used a backcross strategy to make initial linkage assignments of codominant markers using a small number of segregants (15 offspring), to be followed up with a much larger number of progeny (on the order of 100) to develop a recombination map (W. Hara, pers. commun.).

The use of backcross populations for mapping is well-established in *Bombyx* genetics and recently in agricultural pests. It is advantageous since it avoids complications arising from the use of dominant markers and facilitates a biphasic approach to study linkage due to the occurrence of achiasmatic oogenesis (Figure 7). This

approach was elegantly demonstrated recently by Heckel *et al.* in the mapping of a major gene conferring resistance to *Bt* toxin in the diamondback moth, *Plutella xylostella*[28], and in the tobacco budworm, *Heliothis virescens*[29], using AFLP markers. In these studies two backcross populations were created, one by crossing an F_1 female (derived from the initial cross between a susceptible strain and a resistant strain) with its susceptible parent and a second backcross population by crossing an F_1 male with the susceptible female parent. From the first, population information on the linkage group contributing to the trait for resistance was derived. This is possible because in the absence of crossing over in the F_1 females, all of the genes present on the linkage group harbouring the gene for *Bt* toxin resistance would show absolute linkage. The second population was used to determine the location of the trait of interest in the corresponding linkage group because crossing over during gametogenesis in the F_1 males results in recombination of markers and thus allows one to determine map distance between markers. A similar strategy based on the molecular map could be used to track genes of economic interest in silkworm.

Quantitative genetics

The concerted action of two or more individual genes as well as their interactions with each other and with the environment determine many important phenotypic characteristics. Such characteristics are referred to as polygenic or quantitative traits, and the individual gene locations as quantitative trait loci (QTLs). In silkworm almost all of the important economic traits such as fecundity, silk cocoon weight, silk fibre length, silk fibre thickness, and resistance to baculovirus are quantitative in nature. Very little is known about the number of genes involved, their chromosomal locations, or their gene products.

The development of high density genetic linkage maps that are based on DNA marker loci provides a useful tool for the resolution of these complex traits into their individual genetic components[30]. Briefly, individuals from genetic crosses such as F_2 or backcross populations, or recombinant inbred lines or near isogenic lines obtained by repeated crossing of strains with contrasting phenotypes, are evaluated for the phenotype of interest and for their genotype classes at DNA-based marker loci at regu-
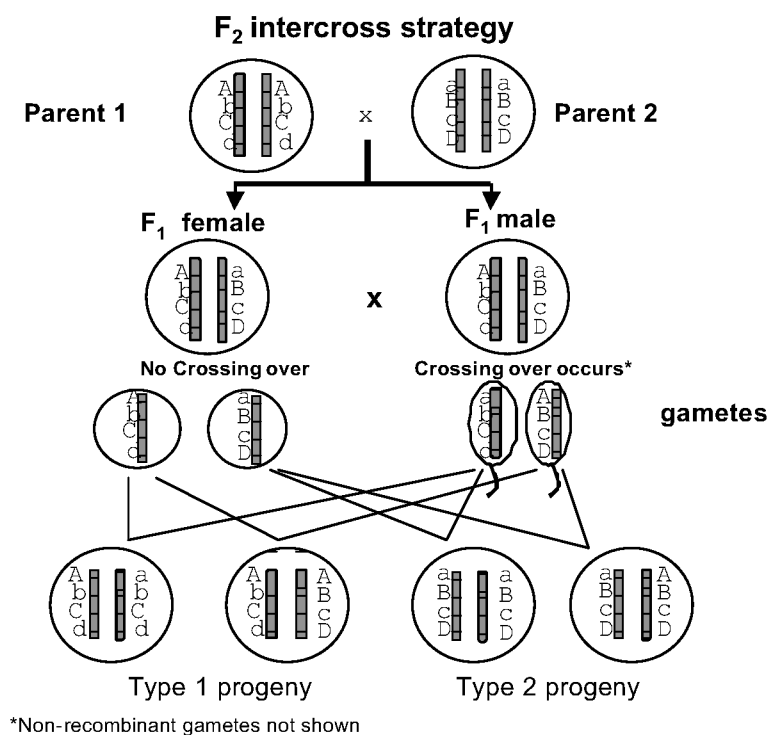

Figure 6. Schematic representation of the behaviour of markers in an F_2 intercross. Achiasmatic oogenesis makes it possible to test whether linkage groups are really independent or not. As shown in the figure, any F_2 individual cannot be homozygous for both maternal and paternal dominant markers on the same autosome. This is specially true for dominant markers where it is not possible to distinguish homozygotes from heterozygotes. The absence of a maternal marker in a certain F_2 progeny means that the nonrecombinant autosome of the progeny is paternal (Type 2 progeny), and vice versa (Type 1 progeny). As a result, each F_2 individual can be typed for each linkage group as to whether its non-recombinant chromosome is paternal or maternal in origin.
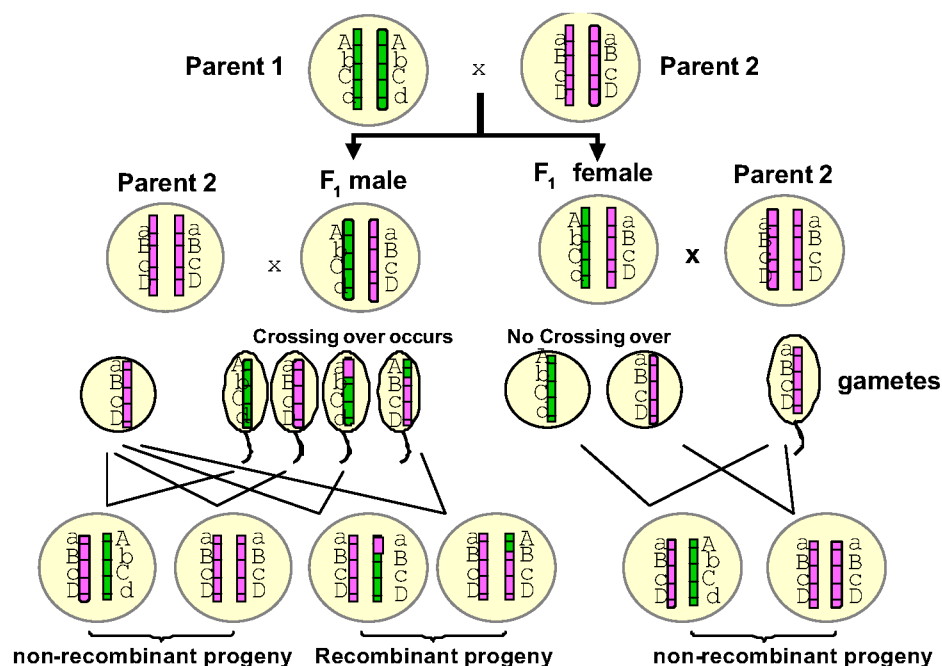
Figure 7. Biphasic nature of genetic linkage in silkworm. Any two loci on the same chromosome are absolutely linked in females as there is no crossing over during oogenesis and the loci cannot recombine. Linkage in males, on the other hand, depends on the recombination distance separating the two loci, because crossing over occurs during spermatogenesis. This biphasic nature of linkage facilitates a sequential approach to mapping. In the first phase, absolute linkage in offspring of female-informative crosses is used to identify the linkage groups, each of which corresponds to a chromosome, contributing to the trait of interest. In the second phase, the location of the gene contributing to the trait of interest within these groups is narrowed down further by analysing recombinant offspring of male-informative matings.

lar 10- to 20-cM intervals across the genome. The differences between the mean phenotype and marker genotype classes are then used to determine the genome positions and the effects of QTLs that contribute to the phenotype.

The DNA markers that are closely linked to the traits of interest in silkworm could be used to select the trait in question in a segregating population of an F_2 cross or using bulked segregant analysis[31], or with recombinant inbred lines[32] or near isogenic lines produced in silkworm improvement programmes. Of course, the effectiveness of such DNA marker-assisted selection (MAS) will depend on the accuracy of the phenotypic classification of the trait of interest and the degree of linkage between a marker(s) and traits of interest in silkworm.

Cross-breeding strategies have been extensively used as a means of harnessing heterosis in the silkworm[33]. In fact hybrid silkworms have contributed significantly to the dramatic increase in silk production of the world[34]. An enormous amount of breeding effort has been invested in the development of silkworm hybrids, and this has resulted in the release of many hybrid combinations in Japan, China, India and other countries. While considerable success has been achieved in silkworm hybrid programmes, experimental data pertaining to the genetic basis of heterosis in silkworm have remained scarce. The recent advances in genome research have generated con-

siderable interest in predicting hybrid performance using molecular markers in crop breeding programmes. Most of the studies have been reported in corn and rice. A large number of studies conducted in maize have produced variable results: high correlations between molecular marker distances of the parents and hybrid performance were detected in some studies[35], while low correlations were observed in others[36,37]. In rice also, there are extensive studies on the relationship between molecular marker heterozygosity and hybrid performance. As in maize, results are variable[38]. Selection of genetically pure and divergent parental strains is critical to the success of a hybridization programme in silkworm. The use of DNA markers to relate the genetic homozygosity and genetic distance of the parental strains to the manifestation of hybrid vigour, and understanding the genetic basis of heterosis, will be quite useful to select suitable parental lines for hybridization programmes. Recently, molecular markers have been genotyped on a set of 22 silkworm strains crossed in a full-diallel programme (Rao, Chandrashekariah and Nagaraju, under preparation) to elicit the relationship between within and between strain genetic variability and the basis of mid- and better-parent heterosis. The results indicate a positive correlation between genetic distance and mid-parental heterosis; there is also evidence to show that parental homozygosity has a dis-

tinct influence on the degree of manifestation of hete-rosis. Most important, the studies enabled them to iden-tify the silkworm inbred lines which produced highly heterotic hybrids. Considering that a large number of silkworm inbred lines are maintained in the silkworm breeding stations in India and other silk-producing coun-tries, the available molecular markers should be used to identify the parental lines that are most suitable for pro-duction of heterotic hybrids. The silkworm strain-specific markers could also be used for authentication of varie-ties, protection of breeders' rights, and periodical moni-toring of the purity of varieties.

Physical mapping of the silkworm genome

Isolation and characterization of genes of interest using map-based cloning methods and systematic sequencing of the Bombyx genome by the proposed International consortium requires a high resolution physical map. A clone-based physical map consists of a set of ordered, overlapping inserts of cloned genomic DNA. Such a map affords a unique resource for studying the structure and function of the genome. The map facilitates the molecu-lar identification of genes of interest, and the clones in the map provide ready access to substrates for genome sequencing. A clone-based map also opens up new opp-ortunities for studies of genome evolution. Clone-based maps have been assembled for several model organisms and agriculturally important species. In most organisms, sets of overlapping clones covering uninterrupted stre-tches of the genome (contigs) are assembled by detecting overlaps by means of shared restriction fragments in fin-gerprints or shared sequence tagged sites (STSs).

The construction of comprehensive and stable DNA libraries cloned in appropriate large-scale vectors is essen-tial to obtain high-resolution physical maps of complex eukaryotic genomes such as silkworm. That the accuracy and efficiency of physical mapping are in direct relation to the mean size of cloned fragments in genomic libraries encouraged the development of several large-scale vector systems in recent years. Yeast artificial chromosomes (YAC) were the first widely used systems, but many problems associated with clone stability, chimerism, and low yield led to the development of another large-scale cloning system based on the E. coli F factor which has largely superceded YAC-based cloning. These vectors are referred to as Bacterial Artificial Chromosomes (BAC)[39]. BAC libraries for silkworm have been con-structed in three laboratories[40] (Mita et al., under prepara-tion) from two B. mori reference strains, p50 and C108, using nuclear DNA from posterior silkglands p50. Alto-gether these libraries contain more than 30 genome equiva-lents, with average insert sizes ranging from 120–170 kb.

The construction of BAC contigs using shared STSs, chromosome walking, and the fingerprinting methods is

underway in the National Institute of Agrobiological Sci-ences, Tsukuba, Japan (Mita, K., pers. commun.). In early studies, the previously reported close linkage bet-ween the Bombyx homologs of the invected (Bm in) and engrailed (Bm en) genes was confirmed by construction of a BAC contig that contained both sequences[40]. Initial screening of BAC libraries has also revealed information about the chromosomal organization of two genes, sericin-1, which codes for one of the major cocoon pro-teins, and the gene encoding the precursor protein of dia-pause hormone and pheromone biosynthesis-activating neuropeptide (the DH-PBAN gene), which are located within a 22-kb interval and are divergently oriented. In addition to their use to orient and organize known genes in fine-structure mapping of silkworm chromosomes, BAC contigs will serve as primary tools for large-scale sequencing of the silkworm genome. Initial efforts are underway to target selected portions of the genome, in-cluding chromosomes 1 (Z), 2, and W, with long-term plans to carry out a complete genome sequence (Mita, K., pers. commun.).

Concluding remarks

The tools and tricks are now in place for large-scale gen-omics of silkworm. The near future will bring an explo-sive growth in functional gene information from the dip-teran paradigm Drosophila, of which much will be useful for silkworm. Foresighted efforts are already underway to initiate the International Silkworm Genome Consor-tium to sequence the Bombyx genome. Technology trans-fer from other genome projects will make important contributions in this endeavour. The next few years should be exciting for silkworm genomics, with great strides being made in the EST sequencing project, BAC library preparation, construction of linkage maps, con-struction of transgenic silkworms for refractoriness to viral diseases, and analysis of QTL. Equally exciting is the potential application of the results of silkworm genomics research to other lepidopterans, particularly Heliothines and wild silkmoths, which represent the des-tructive and beneficial extremes of this large and diverse insect order.

1. Doira, H., Fujii, H., Kawaguchi, Y., Kihara, H. and Banno, Y., Genetic Stocks and Mutations of Bombyx mori, Institute of Genetic Resources, Kyushu University, Japan, 1992.
2. Gage, L. P., Chromosoma, 1974, 45, 27–42.
3. International Lepidopteran Genome Sequencing Consortium (www.ab.a.u-tokyo.ac.jp/lep-genome/proposal).
4. Adams, M. D. et al., Science, 2000, 287, 2185–2195.
5. International Human Genome Sequencing Consortium, Nature, 2001, 409, 860–921.
6. Venter, J. C. et al., Science, 2001, 291, 1304–1351.
7. Nagaraja, G. M. and Nagaraju, J., Electrophoresis, 1995, 16, 1633–1638.

8. Nagaraju, J., Sharma, A., Sethuraman, B. N., Rao, G. V. and Singh, L., *ibid*, 1639–1642.
9. Sharma, A., Niphadkar, M. P., Kathirvelu, P., Nagaraju, J. and Singh, L., *J. Hered.*, 1999, **90**, 319.
10. Tretjk, A. P., Ryskov, A. P., Sevastyanova, G. A., Fillipovich, Y. B. and Strunnikov, V. A., *FEBS Lett.*, 1992, **303**, 258–260.
11. Reddy, K. D., Nagaraju, J. and Abraham, E. G., *Heredity (UK)*, 1999, **83**, 681–687.
12. Promboon, A., Shimada, T., Fujiwara, F. and Kobayashi, M., *Genet. Res.*, 1995, **66**, 1–7.
13. Yasukochi, Y., *Genetics*, 1998, **150**, 1513–1525.
14. Nagaraja, G. M., Ph D thesis, Univ. of Mysore, 2000.
15. Shi, J., Heckel, D. G. and Goldsmith, M. R., *Genet. Res.*, 1995, **66**, 1–7.
16. Tan, Y-D., Wan, C., Zhu, Y., Lu, C., Xiang, Z. and Deng, H. W., *Genetics*, 2001, **157**, 1277–1284.
17. Litt, M. and Luty, J. A., *Am. J. Hum. Genet.*, 1989, **44**, 397–401.
18. Tautz, D., *Nucleic Acids Res.*, 1989, **17**, 6463–6471.
19. Weber, J. L. and May, P. E., *Am. J. Hum. Genet.*, 1989, **44**, 388–396.
20. Dietrich, W. F. *et al.*, *Nature*, 1996, **380**, 149–152.
21. Reddy, K. D., Abraham, E. G. and Nagaraju, J., *Genome*, 1999, **42**, 1057–1065.
22. Weber, J. L., *Genomics*, 1990, **7**, 524–530.
23. Wu, K. S. and Tanksley, S. D., *Mol. Gen. Genet.*, 1993, **241**, 225–235.
24. Nagaraju, J., Reddy, K. D., Nagaraja, G. M. and Sethuraman, B. N., *Heredity (UK)*, 2001, **86**, 588–597.
25. Nagaraju, J., Kathirvel, M., Subbaiah, E. V., Muthulakshmi, M. and Kumar, L. D., *Mol. Cell. Probes*, 2002, **16**, 67–72.
26. Heckel, D. G., *Annu. Rev. Entomol*, 1993, **38**, 381–408.
27. Nagaraju, J. and Muthulakshmi, M., *V Intl. Cong. Ento. Brazil*, 2000.
28. Heckel, D. G., Gahan, L. J., Liu, Y. B. and Tabashnik, E. B., *Proc. Natl. Acad. Sci. USA*, 1999, **96**, 8373–8377.
29. Gahan, L. J., Gould, F. and Heckel, D. G., *Science*, 2001, **293**, 857–860.
30. Lander, E. S. and Botstein, D., *Genetics*, 1989, **121**, 185–199.
31. Michelmore, R. W., Paran, I., Kesseli, R. V., *Proc. Natl. Acad. Sci. USA*, 1991, **88**, 9828–9832.
32. Reiter, R. S., Williams, J. G. K., Feldman, K. A., Rafalski, J. A., Tingey, S. V. and Scolnik, *Proc. Natl. Acad. Sci. USA*, 1992, **89**, 1477–1481.
33. Nagaraju, J., Urs, R. and Datta, R. K., *Sericologia*, 1996, **36**, 1–20.
34. Nagaraju, J., Klimenko, V. and Couble, P., in *Encyclopedia of Genetics* (ed. Reeve, E. C.), Fitzroy Dearborn Press, London, 2001, pp. 219–229.
35. Smith, O. S., Smith, J. S. C., Bowen, S. L., Brown, S. L., Tenborg, R. A. and Wall, S. J., *Theor. Appl. Genet.*, 1990, **80**, 833–840.
36. Godshalk, E. B., Lee, M. and Lankey, K. R., *ibid*, 1990, **80**, 273–280.
37. Dudley, J. W., Saghai Maroof, M. A. and Rufener, G. K., *Crop Sci.*, 1991, **31**, 718–723.
38. Saghai Maroof, M. A., Yang, G. P., Zhang, Q. and Gravois, K. A., *ibid*, 1997, **37**, 145–150.
39. Shizuya, H., Birren, B., Kim, U. J., Mancino, V., Slepak, T., Tachiiri, Y. and Simon, M., *Proc. Natl. Acad. Sci. USA*, 1992, **29**, 8794–8797.
40. Wu, C., Kawasaki, S. and Yasukochi, Y., *Mol. Gen. Genet.*, 1999, **261**, 698–706.
41. Zietkiewicz, E., Rafalski, A. and Labuda, D., *Genomics*, 1994, **20**, 176–183.
42. Estoup, A., Solignae, M., Harry, M. and Cornuet, S. M., *Nucl. Acids. Res.*, 1993, **21**, 1427–1431.

Construction of an EST database for *Bombyx mori* and its applications

Kazuei Mita[*,†], Mitsuoki Morimyo[‡], Kazuhiro Okano[§], Yoshiko Koike[#],
Junko Nohata[†], Masataka G. Suzuki[#] and Toru Shimada[#]

[†]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

To build a foundation for the genome analysis of *Bombyx mori*, we have constructed an EST database. Since gene expression patterns deeply depend on tissues as well as developmental stages, we analysed many cDNA libraries prepared from various tissues and different developmental stages to cover the whole set of *Bombyx* genes. So far, the *Bombyx* EST db contains 26,000 ESTs from 32 cDNA libraries, which are grouped into about 9,500 non-redundant ESTs. Based on the EST db, we prepared a microarray containing 6,000 non-redundant ESTs for functional studies as well as for genome analysis. In addition, we have constructed a high quality BAC library for *Bombyx mori* characterized by an average insertion size of about 170 kb with 11-fold redundancy. We are constructing *Bombyx* BAC contigs by filter hybridization using non-redundant ESTs as probes.

IN eukaryotes, the genome projects of various species have been vigorously pushed forward; genome sequences have been completed in yeast, *Caenorhabditis elegans*, *Arabidopsis* and *Drosophila*, while those of human and mouse will be finished shortly. Genoscope and Celera are carrying out *Anopheles* genome analysis. In addition, whole genome sequencing is planned in many other species. Genome information provides powerful tools to understand biological mechanisms and functions and is essential for biology, medical science and agriculture.

The Lepidoptera to which the silkworm belongs, include the most highly destructive agricultural pests; hundreds of species of caterpillars cause widespread economic damage on food and fibre crop plants, fruit trees, forests and stored grains. Lepidopteran genome information must make a strong impact on insect science and industries such as insecticide, pest control, and silk production (see International Lepidopteran Genome Project Proposal, 2001)[1]. In Lepidoptera, however, genome information is quite limited so far. As *Drosophila* is fairly distant from Lepidoptera evolutionarily (3 mya), the gen-

ome 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, which 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 using a variety of markers, with the aim of providing a framework for positional cloning of specific genes and mutations, large-scale physical map construction, analysis of quantitative trait loci (QTLs) and comparative genomics. To date, about 1,500 markers based on RAPDs, RFLPs, SSRs and ISSRs (Inter Simple Sequence Repeats) are available for the construction of molecular linkage maps, which now cover all 28 *Bombyx* chromosomes at an average spacing of 2 cM, equivalent to about 500 kb (refs 2–4). The robust genetic resources of *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. This will also help basic research leading towards new genome-based approaches for the control of pest species. Therefore, the analysis of the *Bombyx* genome is one of the most urgent requirements in insect science today.

Aiming for the complete analysis of the *Bombyx* genome, we are taking the following strategies: (1) the construction of an Expressed Sequence Tags database (EST db), (2) the construction of a *Bombyx* BAC library, (3) making BAC contigs based on EST markers, and (4) genomic sequencing by BAC shot-gun sequencing. In this paper, we report on the progress of the cDNA project as the first step to the *Bombyx* genome analysis and the applications of the EST db.

Strategy for preparing a comprehensive *Bombyx* cDNA catalog

A cDNA catalog is the comprehensive identification of all expressed genes by large-scale cDNA sequencing.