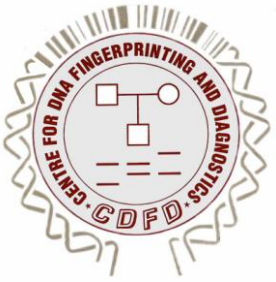


Construction, Characterization and Analysis of Expressed Sequences from Silkmoths

Thesis submitted to

Manipal University

for the Degree of

Doctor of Philosophy

By

K. P. Arun Kumar

(Regd. No. 050100013)

Centre of Excellence for Genetics and Genomics of Silkmoths

Laboratory of Molecular Genetics

Centre for DNA Fingerprinting and Diagnostics

Hyderabad 500076

October 2008

Certificate

This is to certify that this thesis entitled "Construction, characterization and analysis of expressed sequences from silkmoths", submitted by Mr. K. P. Arun Kumar for the degree of Doctor of Philosophy to Manipal University is based on the work carried out by him at the Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad. This work is original and has not been submitted in part or full for any degree or diploma of any other university or institution.

Dr. J. Nagaraju
Thesis Supervisor
Centre of Excellence for Genetics and Genomics of Silkmoths
Laboratory of Molecular Genetics
Centre for DNA Fingerprinting and Diagnostics, Hyderabad.

Dr. Shekhar C Mande
Dean, Academic affairs,
Centre for DNA Fingerprinting and Diagnostics, Hyderabad.

Declaration

The research work embodied in this thesis entitled "Construction, characterization and analysis of expressed sequences from silkmoths", has been carried out by me at the Centre of Excellence for Genetics and Genomics of Silkmoths, Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, Hyderabad, under the guidance of Dr. J. Nagaraju. I hereby declare that this work is original and has not been submitted in part or full for any degree or diploma of any other university or institution.

K. P. Arun Kumar
Regd. No. 050100013
Centre of Excellence for Genetics and Genomics of Silkmoths
Laboratory of Molecular Genetics
Centre for DNA Fingerprinting and Diagnostics
Hyderabad.

Contents

	Title	Page No.
1.	**Synopsis**	1
2.	**Chapter I**	
	Construction, characterization and analysis of expressed sequences from Indian golden silkmoth, *Antheraea assama*	
	Introduction	8
	Materials and methods	10
	Results and discussion	22
	EST analysis	
	Sex determining genes	
	Silk genes	
	Cuticle genes	
	Microsatellites from Antheraea assama	
3.	**Chapter II**	
	Development of WildSilkbase, an EST database of wild silkmoths	
	Introduction	40
	Materials and methods	42
	Results and discussion	48
4.	**Chapter III**	
	***Bombyx mori* testis transcriptome analysis and physical mapping of testis specific genes**	
	Introduction	57
	Materials and methods	59
	Results and discussion	64
	Bombyx mori testis transcriptome analysis	
	Identification of alternative splice forms of intersex gene in silkmoths	
	Physical mapping of testis specific genes onto Bombyx chromosomes	
6.	**Bibliography**	85
7.	**Publications**	94

Acknowledgements

It is a pleasure to thank the many people who made this thesis possible by supporting in matters scientific, and also indirectly by providing such essentials as food, education, money, help, advice, and friendship.

With deep sense of gratitude I thank Dr. J. Nagaraju, my thesis advisor for his constant support and imparting authentic scientific research abilities in me. In addition to Life sciences, I learnt lessons of Life as well. Words limit to express my sincere thanks for the great insights I gained from him, during the 5 years long stay in the lab. Through his Socratic questioning, he brought me closer to the real scientific world, eventually enabling me to grasp its rich complexity. The best training and guidance given, the moral and academic strengths inculcated in me will be honored and remembered.

I could not have wished for better labmates. My seniors in the lab Ramana, Subbu, Satish, Mrinal, Archana and VLN Reddy, batchmate Jayendra and junior Jyothi, were very helpful in all ways from advice to reagents. New entries to the lab, Asha, Chandrapal and Suresh look to be great addition to the lab with their enthusiasm and zeal to learn. Sunil with his exhaustive knowledge taught me many things inside and outside of Science, many thanks to him. I particularly remember the support of Dr. Prashanth during the early years of my PhD. Past members of lab; Padma and Murali were very friendly and helpful. My sincere thanks to other members of lab, Sobhan, Ramesh, Muthulakshmi, Satyavathy, Shekhar, Bhaskar, Kaliappan, Swarna and Sreevidhya, who made a wonderful working atmosphere in the lab and helped in experimentation. Many thanks to Project Assistants, Kifayath, Prem Deepa, Manju Kumari, Vijayasudhakar and Adarsh, and Project Trainees Mayil, Riti, Meghana, Gayathri and Selvi who worked with me during the tenure of my PhD in the lab. I thank Archana Tomar, Eshwar, Vijaya Sarathy and Sravan for help in various computational analyses. I also would like to express my special thanks to Savithri madam, for her kind gesture and moral support in times of great need.

I particularly remember the encouragement received from the Directors of the Centre that facilitated my research sojourn. I would like to thank Dr. Mita san, Dr. Shimada san and Dr. Tamura san for stimulating scientific discussions, research support and hospitality during my visit to Japan on collaborative project. I would like to express my gratitude to the Council for Scientific and Industrial Research for research fellowship and I acknowledge the financial support by Department of Science and Technology, Government of India and Japanese Society for Progress in Science, for my visit to Japan. Additionally, I would like to seize this opportunity to thank the staff of CDFD, APSSRDI, Hindupur and Central Silk Board, Government of India, for their help and support throughout my PhD tenure. My stay at CDFD is memorable especially because of my batchmates, Bibhs, Geetha, Gokul, Nasreena, Noor, Madhav, MRK, Pankaj, Rakesh, Santosh, Smanla and Uma, who are always there to cheer me up. It's because of our batch strength that we enjoyed a lot during and after our graduate course work.

I am forever grateful to my family members, my Mother, Father, Brother, Sister, Brother in law and Sister in law, who kept me away from most of the family responsibilities throughout my PhD tenure. This thesis would not have been a reality without their unconditional love and unlimited support. To them I dedicate this thesis. Thanks to my in laws, Shantha Kumar and Manjula, who understood me and my endeavors, and gave moral support towards the end of my PhD. Finally, I would like to thank Ramya, for her love, affection and encouragement, which made my PhD days a great memory. I can't forget her unconditional support at each turn of the road.

K. P. Arun Kumar *CDFD, Hyderabad*

In the past years from the beginning of the new millennium, genome research has witnessed a tremendous change in terms of sequence data generated and new research resulting from that data. The ultimate goal of genome projects is to produce a complete and accurate sequence of the entire genetic material of a biological species. It was realized at an early stage that the minute part of the genome expressed as mRNA, often referred to as transcriptome, contains much of the information of interest to biologists. Experimentally, a crude snapshot of the transcriptome of a particular tissue or cell type can be obtained by producing a cDNA library of sufficiently high complexity, and sequencing a sufficiently large number of clones, to ensure that most of the information present in the library has been extracted (Jongeneel, 2000). Despite the fact that main aim of any sequencing project is to obtain whole genome sequence and identify a complete set of genes, the eventual utility of the sequence data lies in understanding gene expression pattern of the organism from which the sequence is derived. Once the normal expression profile of a particular gene is known, it will be easy to study what happens in an altered state, such as in disease. Earlier, finding a gene that codes for protein(s) was not easy. However, recent advancements in molecular biology techniques have substantially decreased the time required to locate and fully describe a gene, thanks to the development of, and access to what are called Expressed Sequence Tags (ESTs) (http://www.ncbi.nlm.nih.gov/About/primer/est.html). EST is a short sub-sequence of a transcribed spliced nucleotide sequence (either protein-coding or not). They are useful in identifying gene transcripts, and are instrumental in gene discovery and gene sequence determination (Adams et al., 1991). The identification of ESTs has proceeded rapidly, with approximately 57 million ESTs (dbEST release, 10/2008) now available in public databases. ESTs are generated through single-pass sequencing of a cloned mRNA (i.e., sequencing several hundred base pairs from an end of a cDNA clone taken from a cDNA library). As these clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be present in the database as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand.

ESTs have a variety of applications in genomic science. As ESTs represent expressed portions of genes, they have proven to be powerful tools in the hunt for genes involved in different cellular pathways. ESTs also have a number of practical advantages in that their sequences can be generated rapidly and inexpensively, only one sequencing experiment is needed per each cDNA generated, and they do not have to be checked for sequencing errors because mistakes do not prevent identification of the gene from which the EST was derived (http://www.ncbi.nlm.nih.gov/About/primer/est.html). ESTs come handy to get insights into the nature of genes expressed in a particular tissue of an organism that does not have any previous

sequence information. ESTs are handy tools to predict their protein products, and eventually of their functions. The situation in which the ESTs are obtained gives information on the conditions in which the corresponding gene is acting. In other words, ESTs are suitable resources for identification of tissue, organ and disease specific genes. ESTs contain enough information to permit the design of precise probes for DNA microarrays that then can be used to determine the gene expression pattern. ESTs can be mapped to specific chromosome locations using physical mapping techniques, such as radiation hybrid mapping or FISH. Alternatively, if the genome of the organism that contributed the ESTs has been sequenced one can align the EST sequence to that genome.

The silkworm is a key model organism in the insect order Lepidoptera, which includes more than 160,000 species of which Bombycoid moths consist silkmoths of economic importance. Bombycoid moths secrete diverse varieties of silk fibres. These species include *Bombyx mori* (Mulberry silkmoth) of family Bombycidae and wild silkmoths that belong to Saturniidae, *Antheraea mylitta* (Indian tropical tasar silkmoth), *Antheraea pernyi* (Chinese oak tasar silkmoth), *Antheraea roylei* (Indian oak tasar silkmoth), *Antheraea assama* (Indian golden silkmoth, also known as muga silkmoth), *Antheraea yamamai* (Japanese oak silkmoth), and *Samia cynthia ricini* (Indian castor silkmoth) (Figure 1). India is a major sericulture country in the tropics. It has occupied a place of pride in global sericulture map being the homeland of all the four varieties of natural silks: mulberry, tasar, eri and muga. Six million people in India alone are involved in sericulture, which is very labor intensive and provides a key to improve local quality of life. However, the problems of diseases linked to quality of breeds, unhygienic conditions during rearing, low nutritive quality of leaf fed to silkworms, primitive silk technological conditions have kept the sericulture in India economically unattractive. With the advanced genetic and genomic tools, the genetic loci affecting growth rate, yield and fibre quality, can be tagged with molecular markers for rapid construction of genetically improved strains. Production of heterologous silk and other biomaterials in silkworms would become feasible with the knowledge of the complete genome and biochemical, developmental, and physiological processes that make the silk gland an efficient bioreactor. Thus, in the coming years by combining the traditional breeding and modern biotechnological approaches it may be possible for the geneticists to develop “designer” silkworm varieties that can produce biomolecules, and yield higher and better quality of silk.

Equally exciting is the potential application of the results of silkworm genomics research to other lepidopteras, particularly heliothines and wild silkmoths, which represent the destructive and beneficial extremes of this large and diverse insect order. It is estimated that one-third of the

agriculture production is lost to insect pests, pathogens and weeds. Among these pests, Lepidoptera represent a diverse and important group. The control of agricultural pest populations is achieved mainly by the application of chemical pesticides. Agricultural development is particularly relevant to the food security of developing nations like India. Genetic and genomic research repertoire of the silkworm can spin off genetic information as well as molecular biological tools to be applied in the lepidopteran pest management, particularly to look for new targets for insecticides that are intrinsically selective and therefore potentially safer. Silkworm genetic and genomics resources will thus create new methods of insect pest management and will contribute to sustainable agriculture, protection of the environment and the maintenance of biodiversity.

Functional genomics has particular promise in silkworm biology for identifying genes involved in a variety of biological functions that include: synthesis and secretion of silk, sex determination pathways, insect-pathogen interactions, chorionogenesis, molecular clocks. The members of family Saturniidae, collectively known as saturniids, are among the largest and most spectacular of the Lepidoptera, with an estimated 1,300 to 1,500 different described species distributed worldwide (Grimaldi and Engel, 2005). The muga silkmoth, *A. assama* (n= 15), confined to the North-eastern states of India, is the unique species yet least understood among saturniid moths. The silk proteins of this species have not been investigated so far despite their unique properties of providing golden luster to the silk thread. *S. c. ricini* (n= 13) a multivoltine silkworm commonly referred to as 'eri silkworm' is known for its white or brick-red eri silk. It is distributed in India, China and Japan. Its ecotypes (~16) are distributed across the Palaearctic and Indo-Australian biogeographic regions. The tropical Indian tasar silkmoth, *A. mylitta* (n= 31) represented by more than 20 well-described, genetically distinct ecotypes is a natural saturniid fauna of tropical India. Pursuing the genetics and genomics of saturniids will be of paramount significance for the following reasons: a) Typical of lepidopterans, *B. mori* females are heterogametic, with a ZW chromosome constitution; males are ZZ. Sex chromosomes are considered to be under evolutionary constraints different from those of autosomes. W chromosome is reported to be strongly female determining (Hasimoto, 1933). The sex chromosome system of saturniid silkmoth *A. assama*, on the other hand, is ZZ/ZO as compared to ZZ/ZW observed in other silkmoths. Comparative study of the sex determining genes, would thus reveal diverse sex determination mechanisms in silkmoths, b) Photoperiod plays an important role in the life history traits of saturniid moths and hence it is important to investigate the genes involved in circadian rhythm, c) Silk fibres of different saturniid silkmoths show vast differences in their tenacity, texture, luster and many other biophysical properties. In the light of these, it is

interesting to study the genes encoding silk proteins of wild silkmoths and compare them with those of mulberry silkmoth, and d) Information on immune response genes in these species can throw light on diversity of immune repertoire in these wild silkmoths and may lead to identification of novel immune genes.

Figure 1: Diversity of silkmoth species.

Pictures source: *A. polyphemus*,http://coxon.cesclass.info; *A. semperi* and *A. suraka*, http://www.thebugmaniac.be; *A. oculea*, http://nitro.biosci.arizona.edu; *A. pernyi*, http://www.jpmoth.org

Considering this potential scientific treasure we set out generate large scale ESTs in these silkmoths and carried out extensive computational analyses which led to the identification of several novel genes involved in different cellular pathways.

Chapter 1 describes the construction, characterization and analysis of ESTs from the Indian golden silkmoth, *A. assama*, which is a semi-domesticated, polyvoltine and polyphagous lepidopteran insect, endemic to Northeastern India. Unlike other silkmoths, *A. assama* has low chromosome number (n=15) (Deodikar et al., 1962) and ZZ/ZO sex chromosome system (Gupta and Narang, 1981). *A. assama* is known for its production of quality silk with natural golden color, glossy fine texture and durability. Except for limited EST resources available for *B. mori* (Mita et al., 2003, Xia et al., 2004) and *A. mylitta* (Gandhe et al., 2006) no genomic information is available for any other economically important silkmoths. In view of the above mentioned issues, a gene discovery project was undertaken through large scale sequencing of ESTs from *A. assama*. A total of 35,722 ESTs was generated from 10 different tissues at various developmental stages and *in silico* analysis was carried out to identify genes involved in synthesis of silk, sex determination and formation of cuticle. A number of microsatellite markers were also developed from the characterized ESTs and were tested for polymorphism in natural populations of *A. assama*.

Development of WildSilkbase, an EST database of wild silkmoths has been described in **Chapter 2**. Wild silkmoths have hardly been the subject of detailed scientific investigations, owing largely to non-availability of molecular and genetic data on these species. As a first step, the ESTs generated in three economically important species of wild silkmoths, an Indian golden silkmoth, *A. assama*, an Indian tropical tasar silkmoth, *A. mylitta* and eri silkmoth, *S. c. ricini,* have been catalogued and made accessible to the scientific community through an online database called WildSilkbase. The database is provided with 57,113 ESTs which are clustered and assembled into 4,019 contigs and 10,019 singletons. Several applications such as BLAST query, keywords and Gene Ontology query have been developed for easy access and analysis of data. There are options to carry out searches for species, tissue and developmental stage specific ESTs in BLAST page. Other features of the WildSilkbase include cSNP discovery, GO viewer, homologue finder, SSR finder and links to all other related databases. All the ESTs were annotated for their molecular function, biological process and cellular component using Gene Ontology based annotation (Ashburner et al., 2000). The putative unigenes of each of the three wild silkmoth species were compared among each other and with four model insect species, *Apis mellifera*, *B. mori*, *Drosophila melanogaster* and *Tribolium castaneum* to get insights into conservation and divergence of genes among these species.

In silico analysis of testis derived ESTs from *A. assama* revealed an unusual phenomenon of conservation of several testis specific genes in diverse species of distant taxa, from insects to mammals. However, due to limited availability of genomic resources for *A. assama*, it was difficult to do in-depth analysis in this species. Hence, a study was initiated in *B. mori* system that has large EST dataset of more than 200,000 (Mita et al., 2003, Xia et al., 2004) and 9X solid coverage of genome sequence (Mita et al., 2004, Xia et al., 2004). In **Chapter 3**, analysis of testis transcriptome of *B. mori* has been presented. Male germ cells in animals differentiate into sperms with extensive morphological and physiological changes. This complex process is orchestrated by the expression of thousands of genes, encoding proteins that play essential roles during specific phases of germ cell development. Such genes were studied in *B. mori* through analysis of ESTs derived from testis. This analysis revealed high gene diversity in testis transcriptome, duplicated tektin genes, transposons expressing only in testis and several testis specific genes conserved from insects to mammals. Further, *in silico* analysis revealed a testis specific splice form of *intersex* (*ix*) gene, an important player in female sex determining pathway in *Drosophila melanogaster*; whereas *ix* is reported to be a single exon gene in *Drosophila*. Finally, to study the distribution of testis specific genes on different chromosomes, they were mapped onto their specific locations on *B. mori* chromosomes. The analysis revealed an interesting feature that Z chromosome harbors significantly higher number of testis specific genes than autosomes. Earlier report (Rice, 1984) suggest that the dominant mutations favoring the homogametic sex have a greater chance to be fixed on the Z chromosome where they are exposed to selection. We speculate that lack of dosage compensation and sexual antagonism have favored accumulation of many male specific genes on *Bombyx* Z chromosome. Initial analysis of testis specific paralogs suggest that male advantageous genes are getting accumulated on Z chromosomes either by translocation from other chromosomes or by tandem duplication of such genes on Z.

Results emanated from this work show that ESTs are powerful tools to study the genome of an organism. They are also useful resources for comparative genomic analysis. This study has identified a number of novel genes in different silkmoth species, which will be useful in further work on comparative and functional genomics of silkmoths.

References

1. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, et al. (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252: 1651-1656.

2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

3. Deodikar GB, Chowdhury SN, Bhuyan BN, Kshirsagar KK (1962) Cytogenetic studies in Indian silkworms. Curr Sci 31: 247-248.

4. Gandhe AS, Arunkumar KP, John SH, Nagaraju J (2006) Analysis of bacteria-challenged wild silkmoth, *Antheraea mylitta* (lepidoptera) transcriptome reveals potential immune genes. BMC Genomics 7: 184.

5. Grimaldi DA, Engel MS (2005) Evolution of the Insects. New York: Cambridge University Press.

6. Gupta ML, Narang RC (1981) Karyotype and meiotic mechanism in Muga silkmoths, *Antheraea compta* Roth. and *A. assamensis* (Helf.) (Lepidoptera: Saturniidae) Genetica 57: 21-27.

7. Hasimoto H (1933) The role of the W-chromosome in the sex. determination of *Bombyx mori*. Jpn J Genet 8: 245-247.

8. Jongeneel CV (2000) Searching the expressed sequence tag (EST) databases: Panning for genes. Briefings in Bioinformatics 1: 76-92.

9. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, et al. (2004) The genome sequence of silkworm, *Bombyx mori*. DNA Res 11: 27-35.

10. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, et al. (2003) The construction of an EST database for *Bombyx mori* and its application. Proc Natl Acad Sci U S A 100: 14121-14126.

11. Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735-742.

12. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, et al. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306: 1937-1940.

Chapter I

Construction, characterization and analysis of expressed sequences from Indian golden silkmoth, *Antheraea assama*

The order Lepidoptera includes >160,000 species of which Bombycoid moths consist silkmoths of economic importance. Bombycoid moths secrete diverse varieties of silk fibres. Silk production based on these moths, especially *Bombyx mori*, *Antheraea mylitta*, *Antheraea assama* and *Samia cynthia ricini* plays important role in rural economies of many developing nations. *A. assama*, one of the economically important wild silkmoths whose genome is among the least understood is unique among saturniid moths. Many recent studies show pointer to it being progenitor species of saturniidae. By virtue of the narrow ecological distribution of host food plant, *A. assama* is confined only to Northeastern part of India. Empirical observations show that the population is declining due to depletion in genetic variability (Jolly *et al.*, 1981).

The ultimate goal of genome projects is to produce a complete and accurate sequence of the entire genetic material of a biological species. The part of the genome expressed as mRNA, often referred to as the transcriptome, contains much of the information of interest to biologists. Expressed sequence tags (ESTs) are short stretches of sequences (usually 200-600 bp) obtained by single-pass sequencing of the 5' and 3' ends of cDNAs. This contains at least partial sequences of most mRNAs present in the various tissues used for library construction. Generally, initial goal of any genome sequencing projects will be to identify complete set of genes in an organism. ESTs offer an inexpensive way of tagging the expressed regions of genomes which are unexplored at molecular level.

Functional genomics has particular promise in silkworm biotechnology for identifying genes involved in synthesis and secretion of silk and for identifying gene/enzyme pathways involved in processes such as insect-pathogen interactions. It is dependent on the availability of relevant genomic or EST database and sophisticated bioinformatics tools for database mining. At present opportunities for functional genomics analysis of silkworm have become available due to the construction of relevant EST databases of different tissues of the domesticated silkworm, *B. mori* (Mita et al., 2003, Xia et al., 2004).

Unlike other silkmoths, *A. assama* has low chromosome number (n=15) (Deodikar *et al.*, 1962), ZZ/ZO sex chromosome system (Gupta and Narang, 1981) and fragmented populations with a narrow habitat range, probably experiencing genetic drift. The silk proteins of this species have not been studied so far despite their unique properties of providing golden luster to the silk thread. Hence, it will be exciting and scientifically rewarding to study this wild silkmoth species at molecular level. No genomic information is available for any of the silkmoths except for *B. mori* (Mita et al., 2003, Xia et al., 2004) and *A. mylitta* (Gandhe *et al.*, 2006). In view of this, it is important to explore the transcriptome of *A. assama* and compare it with the well studied *B. mori* and other insect species, which would throw light on the basic biological differences between

domesticated and wild silkmoths in particular, and insects in general. Particularly, the transcriptome resource would enable us to decipher some important genes involved in antimicrobial activity, silk-fibre quality, sex determination and pattern formation in wild silkmoths. Taking into account all the above mentioned issues, a large number of ESTs was generated for *A. assama*. Computational analysis of these ESTs was carried out to identify several novel genes.

Economic importance and biology of A. assama

Native of Assam and named after Assamese word "Muga" which indicates the amber (brown) color of cocoon, wild silkworm *A. assama* is known for its production of quality silk with natural golden color, glossy fine texture and durability. *A. assama* is an endemic species prevalent in the Brahmaputra valley and adjoining hills. It is polyphagous in nature and feeds on leaves of Som (*Machilus bombycina*), Soalu (*Litsaea monopetala*) and other plants which grow abundantly in Brahmaputra valley. Muga silk production is confined to Assam, border areas of neighboring Northeastern states and Cooch Behar in West Bengal. This luminous golden muga silk has now secured Geographical Indication status, the recognition that confers intellectual property right that it has its origin in Assam region. The most expensive of silks, muga is intrinsically woven into the cultural traditions of the people of Assam. The vibrant Sualkuchi sarees and mekhala-chaddars are the traditional items made from muga silk. In recent times, fashion designers have found exciting prospects in using muga silk for developing new products and designs. Use of muga yarn as a substitute for 'zari' in sarees is finding favor with reputed weavers. For six hundred years, muga silk was worn only by the Ahom kings and noble families of Assam. The motifs that are used on muga silks are largely animals that belong to the Kaziranga - a direct indication of the region to which the muga belongs. The fabric was unknown to the outside world until 1662, when the French explorer Jean Joseph Tavernier travelled through Assam. The gold color and shine of a muga textile increases with every wash, in sharp contrast to the natural law of decay of shine in fabrics with time. Clothes made from muga silk have been known to last for 50 years. Muga silk possesses the highest tensile strength among all the natural textile fibres and is comfortable to wear in both summer and winter seasons. Muga silk is also believed to have medicinal properties and is apparently used as a skin whitener.

A. assama, a multivoltine and polyphagous insect feeds on 15 different host plant species. As in any typical lepidopteran species, life cycle of *A. assama* consists of four distinct stages of development: egg (ova), larva, pupa and adult (imago) (Figure 1). Egg stage has an incubation period of 10-14 days. The egg-shell provides a protective covering for embryonic development. The larval developmental stage includes five instars, each instar having distinct morphological characteristics. The moth has bipectinate antennae. The male moth is smaller than the female (Figure 2). The chromosomal organization of this species is n = 15 (Figure 3) (Deodikar *et al.*, 1962).

Heterogeneity and sexual dimorphism with regard to colour pattern are less pronounced in *A. assama* than in *A. mylitta* because of its limited geographical distribution. The eggs of this species are streakless and brownish. The follicular imprints consist of a single pattern with oval

main cells. The newly hatched larva is characterized by prominent black intersegmental markings over the yellowish body with brown head. After the first molt the body turns green, while the head remains brown. Yellow colour body larvae can also be seen. The pupa is copper brown, weighs about 6 grams. The cocoon is single shelled, light brown, oblong, closed, reelable and slightly flossy with a weak peduncle. The cocoon is golden brown or glossy white. The approximate body length of male moths is 3 cm and of the females 3.5 cm (Jolly *et al.*, 1981) (Figure 2).

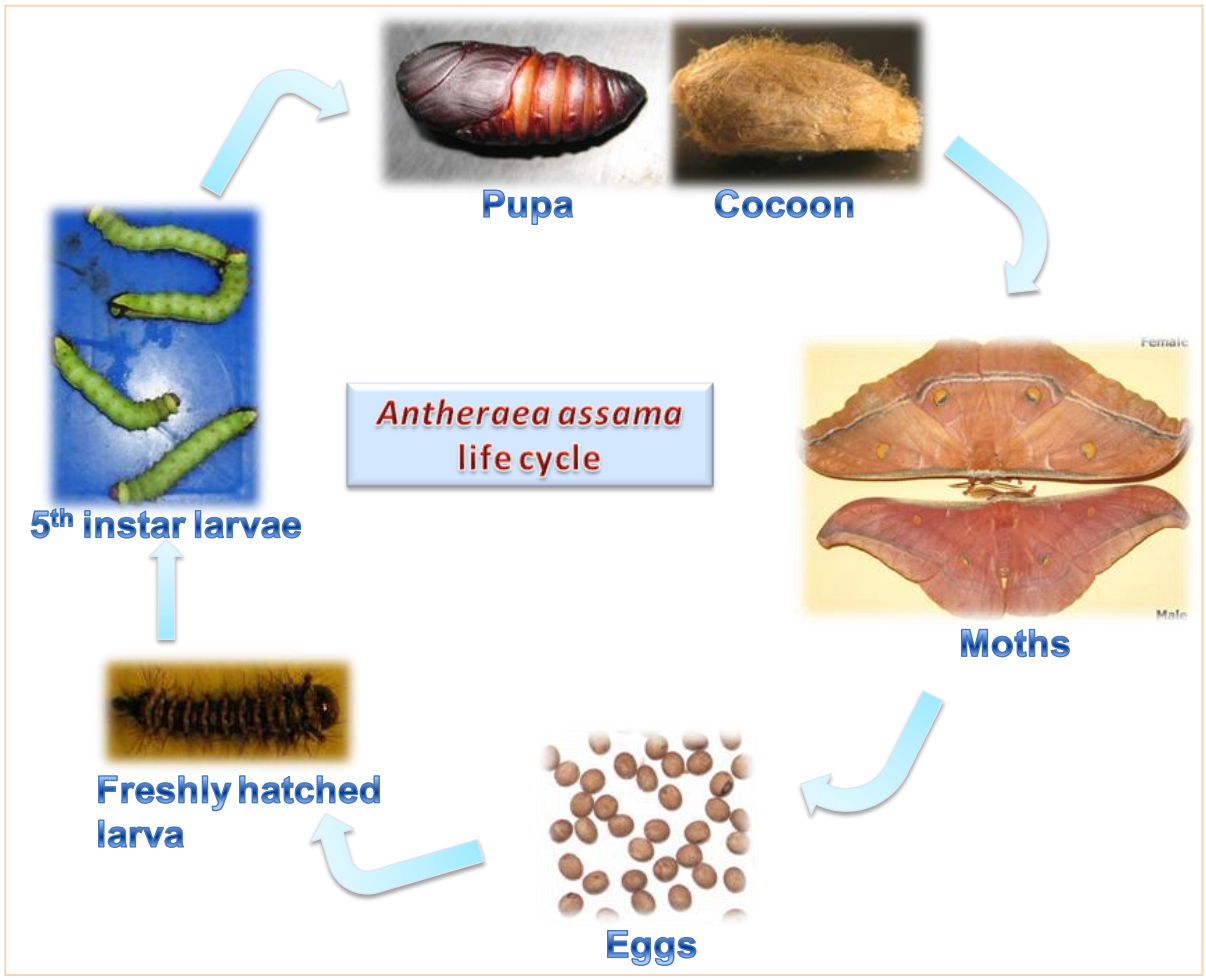


Figure 1: Life cycle of *A. assama* (See the text for details).

Insects and tissue collection

The larvae and eggs of *A. assama* were procured from farmers through the Central Muga and Eri Research and Training Institute, Central Silk Board, Assam, India. The testis, ovary, brain, fatbody, epidermis, midgut, posterior and middle silkgland tissues were dissected from day five of 5th instar

larvae (Table 1). Dissection was carried out in Insect Ringer's solution under dissection stereomicroscope. The dissected tissues were transferred to liquid N_2 and stored in -80^0 C.

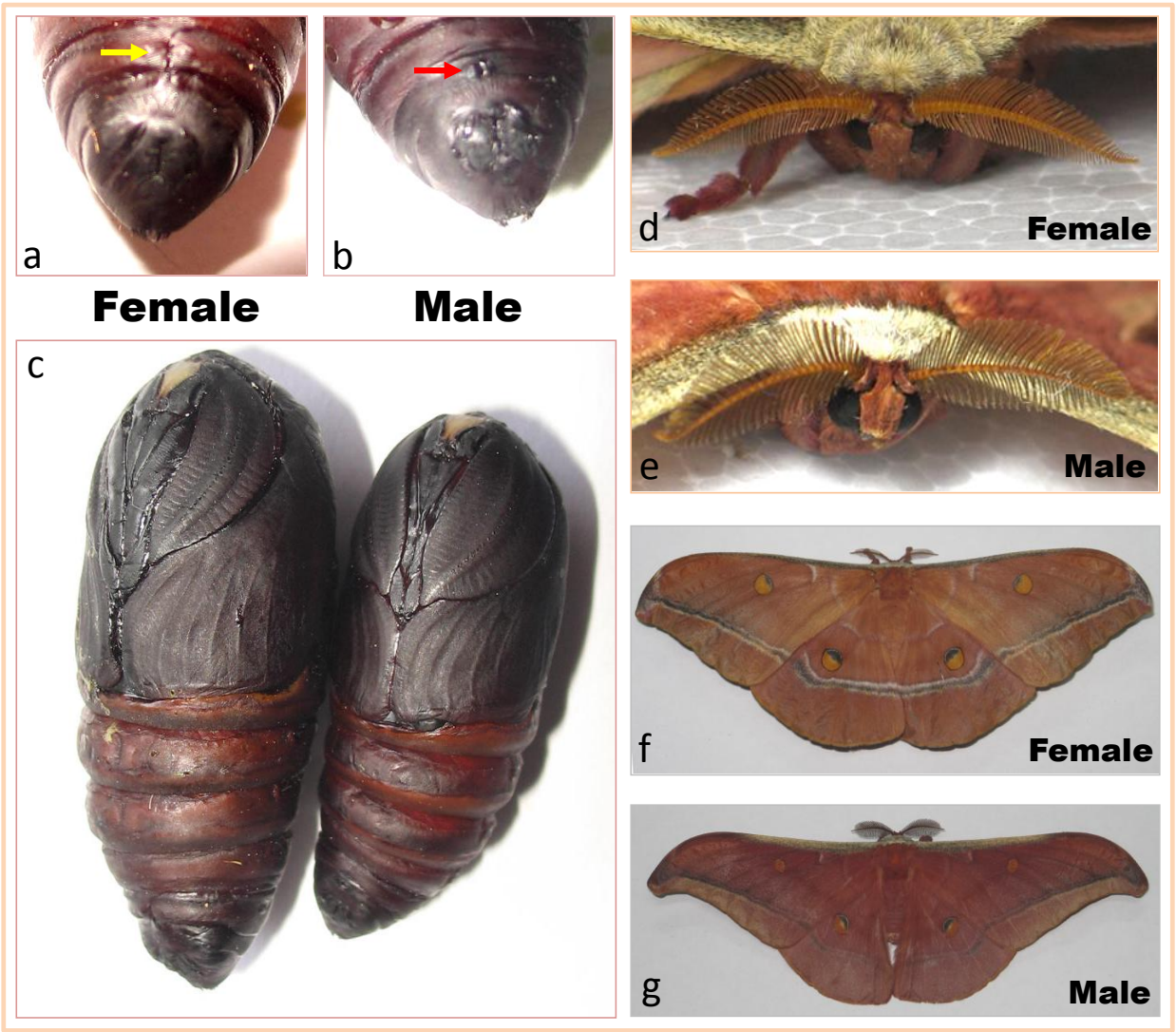


Figure 2: Male and female pupae (a, b, and c) and moths (d, e, f and g) of *A. assama* showing sexual dimorphism. Yellow arrow shows female genital marking (a) and red arrow male genital marking (b) on pupae. Female moth antennae are tapered (d) whereas male antennae are broad (e). Female pupae (c) and female moths (f) are generally bigger than that of males.

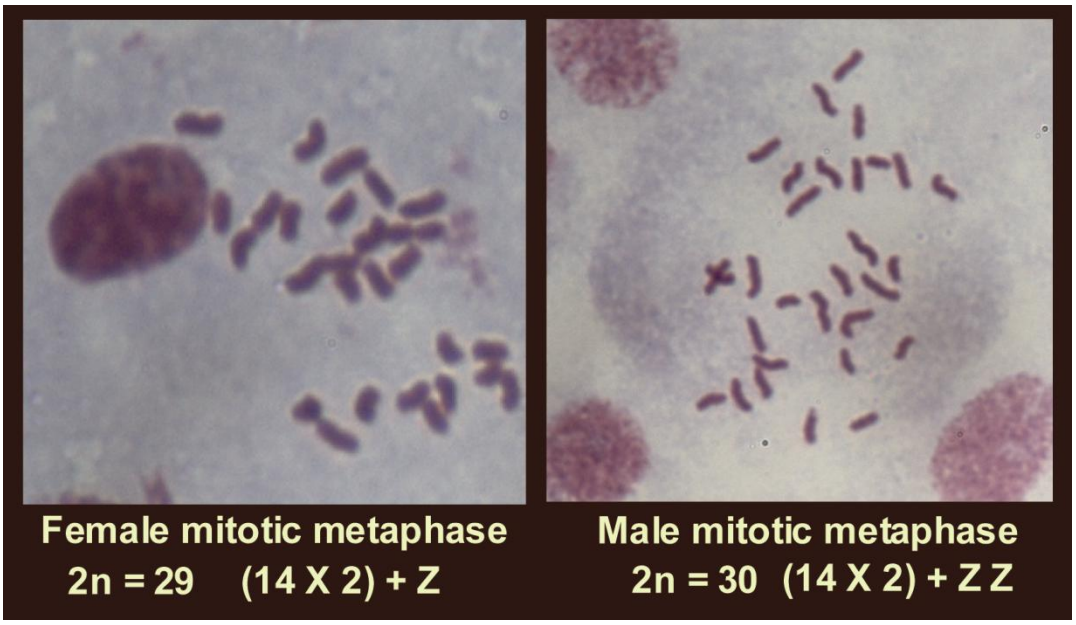


Figure 3: Chromosomes of *A. assama*. Unlike other silkmoths, *A. assama* lacks W chromosome.

Total RNA isolation and cDNA library preparation

Total RNA was extracted from different tissues harvested from several larvae and embryos as mentioned in Table 1, using Trizol reagent (Invitrogen, Carlsbad, CA, U.S.A.) followed by treatment with DNase (Invitrogen, Carlsbad, CA, U.S.A.). The cDNA synthesis was carried out using Stratagene ZAP-cDNA synthesis kit following manufacturer's instructions. ZAP-cDNA synthesis kit uses a hybrid oligo(dT) linker–primer that contains an *Xho* I restriction site. Messenger RNA was primed in the first strand synthesis with the linker–primer and is reverse-transcribed using StrataScript™ reverse transcriptase and 5-methyl dCTP. During second-strand synthesis, RNase H nicks the RNA bound to the first strand cDNA to produce a multitude of fragments, which serve as primers for DNA polymerase I. DNA polymerase I "nick-translates" these RNA fragments into second-strand cDNA. This method allows construction of cDNA ready for directional insertion into lambda vector pBluescript SK(+). Directional cDNA library was constructed by cloning of size fractionated cDNA fragments into pBluescript II SK(+) vector and electroporation into *E. coli* strain DH10B.

Table 1: Complementary DNA libraries prepared from different tissues of *A. assama.*

Library name	Tissue type	Stage	Number of tissues isolated
Aaem	Embryo	96 hours after oviposition	5000
Aamg	Midgut	5th instar larvae (fifth day)	20
Aafb	Fatbody	5th instar larvae (fifth day)	20
Aats	Testis	5th instar larvae (fifth day)	4000
Aaov	Ovary	5th instar larvae (fifth day)	4000
Aabr	Brain	5th instar larvae (fifth day)	4000
Aamsg	Middle silkgland	5th instar larvae (fifth day)	50
Aapsg	Posterior silkgland	5th instar larvae (fifth day)	50
Aaep	Epidermis	5th instar larvae (fifth day)	20
Aace	Compound eyes	5th instar larvae (fifth day)	100

Software used in the EST analysis

Different software have been used through various steps of processing the sequences. Most of them are stand-alone versions for UNIX based systems.

Repeat Masker

Repeat Masker is a program that screens DNA sequences for interspersed repeats known to exist in mammalian genomes as well as for low complexity DNA sequences (http://www.repeatmasker.org/RMDownload.html).

Cross match

'Cross match' was used to clean up the sequences from viral sequences, vector and linker regions (http://www.phrap.org/phredphrapconsed.html). Vector may cause reads to be identified as "possible chimeras", or otherwise interfere with proper assembly. Cross match is a general-purpose utility (based on a "banded" version of SWAT, an efficient implementation of the Smith-Waterman algorithm) for comparing any two sets of (long or short) DNA sequences. It can be used to compare a set of reads to a set of vector sequences and produce vector-masked versions of the reads; a set of cDNA sequences to a set of cosmids; contig sequences found by two alternative

assembly procedures (e.g., phrap and xbap) to each other; or phrap contigs to the final edited cosmid sequence. It is slower because it allows gaps but more sensitive than BLASTN.

TGICL

TGI Clustering tools (TGICL)(http://www.tigr.org/tdb/tgi/software/) are a software system for fast clustering of large EST datasets. The purpose of this package is to efficiently cluster and create assemblies (contigs) from a set of DNA sequences given in a fasta file. The "clustering" phase is intended to partition the input data set into smaller groups of sequences (clusters) that have stringent similarity and potentially coming from the same longer original sequence. However, the clustering phase does not perform any multiple alignment but accomplishes only fast pairwise alignments (using megablast), which are then filtered and used to build subsets of sequences by a transitive closure approach. In the assembly phase each such cluster is sent to the assembly program (CAP3), which attempts the multiple alignment of the sequences in the cluster and creates one or more contigs (consensus sequences).

BLAST

The Basic Local Alignment Search Tool (ftp://ftp.ncbi.nlm.nih.gov/) employs the Smith-waterman algorithm of local alignment in which the word of specified size is first matched between the query and the subject and once after a perfect match is found the sequence is searched out on both sides for the maximum possible extent to give out the nearly perfect alignment. As the name suggests the alignment that is performed is local. So the motifs that are found in the queries with significant similarity to those in the subject are retrieved.

Databases employed in the EST analysis

A number of bioinformatics databases was used in the study to carry out different kinds of data analysis.

UniVec

UniVec is a database (ftp://ftp.ncbi.nih.gov/pub/UniVec/) that is used to quickly identify segments within nucleic acid sequences, which may be of vector origin. Screening using UniVec is efficient because a large number of redundant sub sequences are eliminated to create a database that contains only one copy of every unique sequence segment from a large number of vectors. In addition to vector sequences, UniVec also contains sequences for adapters, linkers, and primers

commonly used in the process of cloning cDNA or genomic DNA. The software enables detection of contamination of these oligonucleotide sequences during vector screen.

Non-redundant protein database

The non-redundant protein database (ftp://ftp.ncbi.nlm.nih.gov/blast/) is compiled from a variety of sources, including SwissProt, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq. As suggested by the name, there are no redundant entries in the database.

PERL 5.0

Perl is a "Practical Extraction and Report Language" freely available for Unix, MVS, VMS, MS/DOS, Macintosh, OS/2, Amiga, and other operating systems (www.perl.com/download.csp). Perl has powerful text-manipulation functions. It eclectically combines features and purposes of many command languages. Perl has enjoyed recent popularity for programming World Wide Web electronic forms and generally serves as glue and gateway between systems, databases, and users. Perl is useful in computational analysis of biological information, especially sequences because of its capability to handle with the strings.

MySQL 6.0

MySQL (www.mysql.com/downloads) is used for creating the database of the resultant annotated EST sequences and to archive the results in various steps, especially in clustering and assembly.

Sequencing and analysis of ESTs

All the manipulations for sequencing were carried out in 96 well microtitre plates and reactions were carried out using a 96 well PCR machine. The cDNA inserts were sequenced from 5' end with RV-M primer (5'GAG CGG ATA ACA ATT TCA CAC AGG 3') using MegaBase 3000 sequencer (Amersham Biosciences Inc, UK). All the clones were archived in glycerol stock and stored at -70 ^{o}C deep freezer.

Raw sequences from sequence chromatograms were processed using several programs. A cut off Phred quality value of $\geq$15 was assigned to extract quality sequences from chromatograms. The quality sequences were screened for the presence of vector**,** low quality and adapter sequences using 'Cross Match' program (Ewing *et al.*, 1998). Further, masked vector sequences were automatically removed by in-house developed trimming tool. Sequences shorter than 50 bases were removed. The resulting high-quality sequences were assembled into sequence contigs with the TGICL program (Pertea *et al.*, 2003), which initially makes clusters using MegaBLAST and

subsequently makes an assembly using CAP3 for each cluster generated in the first step. A cluster is defined as a unique set of sequences which shares common sequence similarity. A cluster containing only one sequence is termed a singleton.

The unique putative gene sequences obtained by clustering and assembly were annotated by running BLAST (Altschul *et al.*, 1990) against non-redundant (*nr*) protein database of NCBI. Further, BLAST output was parsed to classify the putative gene transcripts into different functional classes. For the sequences showing high similarity to hypothetical and/or unknown proteins, and those sequences showing weak similarity to *nr* protein database, domain search was performed using ProDom (Bru *et al.*, 2005) and BLOCKS Search tools (Henikoff *et al.*, 2000). Putative function was assigned based on domain type.

Based on the Gene Ontology (GO) annotation of the closest annotated homolog, ESTs were assigned a molecular function, biological process and cellular component from the GO database (Ashburner *et al.*, 2000). GO annotation produces a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge of gene and protein roles in cells is still accumulating and changing. To this end, the Seqdblite FASTA sequence file from the GO database was downloaded. By running BLAST against Seqdblite, molecular function, biological process and cellular localization were assigned. An E-value of 1e-20 was set as a cutoff value for all EST annotations.

Analysis of EST lengths

The length of ESTs was determined using the in-house program. The mean length of ESTs for each tissue was calculated separately, using Microsoft Office Excel program. The distribution of lengths was plotted on bar graph, against the number of ESTs of each length.

Analysis of cuticle genes

A. assama unigenes were BLAST searched for cuticle genes, to identify genes coding for cuticle proteins and to estimate cuticle transcript content in different tissue transcriptomes. The resultant set of unigenes was subsequently analyzed to get the information on number of ESTs contributed by different tissues. Furthermore, each putative *A. assama* cuticle gene was categorized into RR1 and RR2 type of cuticular proteins using the "Distinguish any sequence between RR1 & RR2" feature of CuticleDB, a relational database of Arthropod cuticular proteins (Magkrioti *et al.*, 2004).

Analysis of silkgenes

ESTs derived from silkglands are good source for identification of genes coding for silk proteins. In the present study, ESTs originated from middle and posterior silkglands were BLAST searched using published silk protein sequences (fibroin, sericins and P25) of silkmoths as query. The resultant sets of ESTs for each gene were clustered and assembled to attempt aligning full length transcript. Number of ESTs contributed by each gene was counted to estimate their proportion in the silkgland transcriptomes.

Attempts were made to get full length transcripts by combining EST data, and *de novo* sequencing of EST clones and other PCR amplicons produced by newly synthesized primers. Several EST clones identified by BLAST search as coding for silk proteins, were cultured and plasmids were extracted. Length of the inserts was determined by restriction digestion and the clones which had insert length of more than 1 kb were sequenced from both ends to get longer sequences. Sub-cloning was done for inserts of more than 1.5 kb and inserts were sequenced. In case the sequencing was difficult through sub-cloning, plasmid walk was done to get complete sequence of the insert. The new sequences thus obtained were aligned with EST sequences and the alignment was manually checked for accuracy. The sequences leading to erroneous alignment were removed. The primers used in amplifying and sequencing fibroin and sericin cDNAs are listed in Table 2.

Fibroin and sericin proteins are composed mainly of repeat motifs of variable lengths. The partial gene sequences of fibroin and sericins were conceptually translated to get corresponding amino acid sequences. These sequences were analyzed for repeat motifs and were compared with those of other silkmoths.

Northern blot: To estimate the length of transcripts and to study the expression pattern of fibroin gene, northern hybridization was carried out using in vitro transcribed fibroin EST RNA. Total RNA was extracted from middle and posterior silkglands separately, using TRIZOL reagent. Quality check and estimation of quantity of total RNA was carried out using Nanodrop spectrophotometer. Twenty micrograms each of RNA was loaded onto agarose gel (1.2%) containing 2.2M formaldehyde. The samples were run at 60V for 5 hours in 1X MOPS electrophoresis buffer. After the run, the gel was rinsed in DEPC water for 20 minutes on a rocker and was then washed in alkaline transfer buffer for 20 minutes. This would aid in efficient removal of formaldehyde which, if present, would hinder transfer. Once separated by denaturing agarose gel electrophoresis, the RNA was transferred to a positively charged nylon membrane and then immobilized by UV cross-linking, for subsequent hybridization. The blot was washed in 2xSSC for

10 minutes and air dried. The blot was hybridized with RNA probe synthesized by *in vitro* transcription. The EST clone H14, which harbored 1.5 kb region of fibroin 3' end, was selected for probe preparation. The radio-labeled fibroin RNA probe was synthesized by *in vitro* transcription of H14 linearised plasmids, using MegaScript *In Vitro* transcription kit (Ambion). Prehybridization, or blocking, is required prior to probe hybridization to prevent the probe from coating the membrane. Good blocking is necessary to minimize background problems. Therefore prehybridization was done in pre-hybridisation solution containing 6x SSC, 0.5% SDS and 100 µg sperm DNA (pre-warmed at 65 ^{o}C) at 65 ^{o}C for 50 minutes. Then denatured (at 100 ^{o}C for 10 minutes and immediately chilled) probe (20 µl) was added to hybridize it for 16 hours in a hybridization chamber at 65 ^{o}C. After hybridization, unhybridized probe was removed by washing in several changes of buffer with initial low stringency buffers (2X SSC) and subsequently in high stringency buffers (0.1X SSC). Immediately, the blot was wrapped in a saran wrap and placed in the Phosphor Imager cassette. The PhosphorImager screen was placed over it. Exposure was done for 1 minute – 10 hours depending on the signal intensity. Then the screen was scanned using the Typhoon scanner (Amersham Biosciences Inc, UK).

Fibroin protein extraction from cocoon: Fibroin protein was extracted from *A. assama* cocoons to get information on protein size and number of proteins. Hundred milligrams of cocoon flakes were degummed by treating in 8M Urea+1% SDS+2% Mercapto-ethanol for 1 hour. Then the mixture was centrifuged and the supernatant was removed. Cocoon flakes were dried in vacuum drier for 4 hours and then treated with saturated Lithium Thiocyanate (LiSCN) for 2 hours at 65 ^{o}C. Clear supernatant was obtained by centrifugation of the treated cocoon flakes solution, at 13000rpm for 15-30 minutes at 4 ^{o}C. The supernatant was subjected to dialysis against 20mM Tris-HCl (pH 8.0) containing 2M urea at 4 ^{o}C, overnight. The concentrated fibroin protein thus obtained was estimated using Bradford reagent and resolved on 8% SDS Polyacrylamide gel.

Table 2: Details of primers designed for *Fibroin* and *Sericin* genes of *A. assama*. Ta: annealing temperature.

Gene	Primer code	Primer sequence	Ta
Fibroin	AaFibH_F	TTTGAACGATTCACAACACG	58.16
Fibroin	AaFibH_R	TCGAAAGAAGTGGATGACCTC	59.27
Fibroin	Aa_fib_1	GATCTTGTGCTGCGTTTTGCAGTA	65.68
Fibroin	Aa_fib_2	AGAGCCTGGACCATAACCACCATC	66.58
Fibroin	Aa_fib_3	GATGGTGGTTATGGATCAGGCTCTTC	66.96
Fibroin	Aa_fib_4	AATTAGTGGACGGAAATTCTGGAAGC	65.55
Fibroin	Aa_fib_5	ATATCCACCGCCATATCCTCCAGCAC	70.46
Fibroin	acpsg0030_F	TGAACCTGAACCATATCCGTCGT	64.85
Fibroin	acpsg0030_R	GCTGGAGGATATGACGGTGCTT	64.77
AaSer1	Aa_Sericin1_F	GCGGCAGCTATGGTGGTACT	62.49
AaSer1	Aa_Sericin1_R	CTTCGACTTTGGCCACTTTT	59.35
AaSer2	Aa_Sericin2_F	GCGATGGTAGAACATACAGCAA	60.16
AaSer2	Aa_Sericin2_R	CTTTGACTTCGGCCATTTTTAC	60
AaSer2	AaSer5	TATTCTCATGCCGATGGTGA	48
AaSer2	AaSer6	GTAGTGGCAGAACAGCACAGTC	56

Amplification of complete fibroin gene using long range PCR: Analysis of silkgland ESTs led to identification of 5' and 3 ends of fibroin gene. These ends have non-repetitive DNA stretches, which were exploited for designing specific primers. These primers (Aa_fib_1 and Aa_fib_4, Table 2) were used to amplify complete gene using genomic DNA as template and following long range PCR protocol (Long Range PCR Enzyme, MBI fermentas).

Development of microsatellite markers

To develop EST-SSR markers, 8,197 unique sequences derived from 35,722 *A. assama* ESTs were screened using WebTROLL (Tandem Repeat Occurrence Locator) software (Castelo *et al.*, 2002) for simple sequence repeats (SSRs) harboring di, tri, tetra and pentanucleotide repeats.

For the construction of repeat enriched genomic library, DNA was isolated as described earlier (Prasad *et al.*, 2002) from *A. assama* pupae. Genomic DNA was enriched using an oligonucleotide mix of $(ATT)_8$, $(GAGT)_2$, $(CA)_{10}$, $(GA)_{10}$, $(GATA)_{10}$, $(CAC)_7$, and $(AGC)_7$ following previously reported protocol (Glenn and Schable, 2005). In short, DNA was digested with *Rsa*I and *Xmn*I restriction enzymes (New England Biolabs Inc., USA), ligated to double stranded superSNX linkers, hybridized with biotinylated microsatellite oligonucleotides and captured on streptavidin coated magnetic beads. Unhybridized DNA was washed away and the captured DNA was recovered by PCR using single stranded superSNX-F as a primer. The PCR products were ligated into the pCR 4-TOPO TA vector (Invitrogen, Carlsbad, USA) and transformed into XL1-Blue competent cells. Colony PCR was performed to select the amplicons of 500-1000 bp and purified using the AMPure PCR purification systems (Agencourt Bioscience, USA). The inserts were amplified using M13 primers and sequenced on an ABI Prism 3100 Genetic Analyzer (Applied Biosystems, USA), with BigDye Terminator (version 3.1) chemistry.

Primer 3 program (Rozen and Skaletsky, 2000) was used to design primers flanking SSRs in the EST clusters and microsatellite enriched sequences. PCR was carried out on a Mastercycler Gradient (Eppendorf, Germany) in a 10 μl reaction containing 1X PCR buffer (10mM Tris-HCl, pH 8.8, 50mM KCl) (MBI Fermentas, USA), 100 μM dNTPs, 1.5 mM $MgCl_2$, 5 picomole of each primers, 10 ng of genomic DNA as template and 0.5 U *Taq* polymerase (MBI Fermentas, USA). The PCR conditions were: initial denaturation of 94°C for 3 minutes, 35 cycles of 94°C denaturation for 20 seconds, appropriate annealing temperature (established empirically) for 10 seconds and 72°C extension for 45 seconds, and 72°C for 10 minutes as final extension. The amplified products were then analyzed on 3.0% Agarose gel (Hi Resolution, Sigma-Aldrich, St. Louis, USA) along with 50bp size standard (MBI Fermentas, USA) at 90V for 3.5 hours and visualized by Ethidium Bromide staining. The alleles were scored using the BIO-RAD Quantity One Software (Bio-Rad Laboratories, USA). The scored alleles were analyzed using genetic analysis software. GenAlex version 6 (Peakall and Smouse, 2006) was used to calculate allele frequency, observed mean and expected heterozygosity. Pairwise tests for linkage disequilibrium were performed using web based program GENEPOP version 3.4 (Raymond and Rousset, 1995). The presence of null alleles at each locus was tested using micro-checker version 2.2.3 (van Oosterhout *et al.*, 2004).

To get insight into expressed portion of the *A. assama* genome, a total of 35,722 ESTs was generated from ten tissues collected at larval and embryonic stages. An EST processing pipeline was developed to characterize and annotate the ESTs. Upon redundancy removal 8,197 unique putative genes were obtained. Functional annotation based on BLAST against NCBI protein nr database revealed the presence of several genes that are involved in silk production, circadian rhythm, sex determination, immune response and also several novel tissue specific genes. ESTs generated in the present study represent fairly the gene content of this insect species as we have selected 10 important tissues from 2 important developmental stages where diversity of genes expressed is higher.

EST analysis

Sequencing, assembly and analysis of A. assama ESTs

Standard unidirectional cDNA libraries were generated from nine different tissues collected at day five of 5th instar larvae and 96 hours embryos (Table 1). From each cDNA library 1,500-11,500 clones were sequenced. A total of 41,500 clones were processed and sequenced, generating ~39,000 5' end ESTs. After trimming low quality and vector sequences, and removing contaminant bacterial sequences and sequences <50 bp the resulting dataset contained 35,722 high quality ESTs with a Phred quality score of $\geq$15 and an average length of 472 bp. Mitochondrial sequences were not filtered and they accounted for ~2% of total ESTs (~700 ESTs). Insert amplification of clones from each library showed inserts ranging between 500 bp to 2000 bp with an average insert size of 1000 bp.

The initial data set of 35,722 ESTs yielded 8,197 non-redundant putative gene transcripts after clustering and assembly. There was varied reduction in sequence data due to redundancy removal through clustering and assembly in all the 10 tissue libraries. Embryo ESTs showed highest reduction in sequence data (88%) followed by midgut (87%), posterior silkgland (82%), middle silkgland (79%), fatbody (70%) and epidermis (65%), respectively. A very high degree of EST diversity was found in testis ESTs with a lowest percent reduction (27%) in sequence data after clustering and assembly, followed by ovary (40%), compound eye (46%) and brain (53%), respectively (Table 3). Higher diversity of genes in *A. assama* testis as compared to other tissues suggests that the complex process of spermatogenesis requires expression of a variety of genes. This is consistent with the observations recorded in previous studies. Analyses of ESTs and microarrays indicated that an unusually diverse set of mRNAs is expressed in mouse, human, and

Drosophila testes and that a large number of these mRNAs are expressed only in spermatogenic cells (Andrews et al., 2000, Kerr et al., 1994, Pawlak et al., 1995). In a recent study in *B. mori* such a phenomenon was observed through microarray analysis of several tissues (Xia *et al.*, 2004). The analysis revealed striking differences in gene expression between testis and ovary of *B. mori* fifth instar larvae. The study found 1,104 genes specifically expressed in the testis, and were associated with spermatogenesis, reproduction, mitosis and fertilization.

Table 3: Total number of *A. assama* ESTs generated from each EST library, number of quality sequences obtained after trimming contaminants and number of non-redundant sequences (contigs and singletons) obtained after clustering and assembly.

Library name	No. of raw sequences	No. of quality ESTs obtained	No. of contigs	No. of ESTs in contigs	No. of singletons	Total non-redundant sequences	% reduction in sequence data
Aaem	12,500	11,502	458	10,688	921	1,379	88.00
Aabr	6,000	5,299	536	3,411	1,941	2,477	53.26
Aats	5,000	4,235	497	1,701	2,572	3,069	27.53
Aaov	4,500	3,891	461	2,076	1,846	2,303	40.81
Aamg	3,000	2,439	116	2,286	185	301	87.66
Aafb	3,000	2,318	231	1,884	460	691	70.19
Aapsg	3,000	2,543	125	2,227	316	441	82.66
Aamsg	1,500	1,031	65	905	150	215	79.15
Aaep	1,500	1,386	115	1,040	359	474	65.80
Aace	1,500	1,078	118	628	461	579	46.29
Total	**41,500**	**35,722**	**2,260**	**30,204**	**5,937**	**8,197**	**77.05**

Distribution of read length of ESTs

One of the many ways to assess the quality of ESTs is by their read lengths. An average read length of 450-500 bp, is considered to be a good measure of EST library. The length of ESTs was analyzed and their mean lengths were calculated in all the 10 tissue ESTs. The distribution of lengths was plotted against the number of ESTs. Average length was 436, 496, 488, 463, 495, 468, 448, 433, 499 and 496 bp in the ESTs derived from embryo, brain, testis, ovary, midgut, fatbody, middle silkgland, posterior silkgland, epidermis and compound eye, respectively. Among these, average length of embryo derived ESTs was less than 450 bps. All other library ESTs had an average read length of 472 bp (Figure 4).

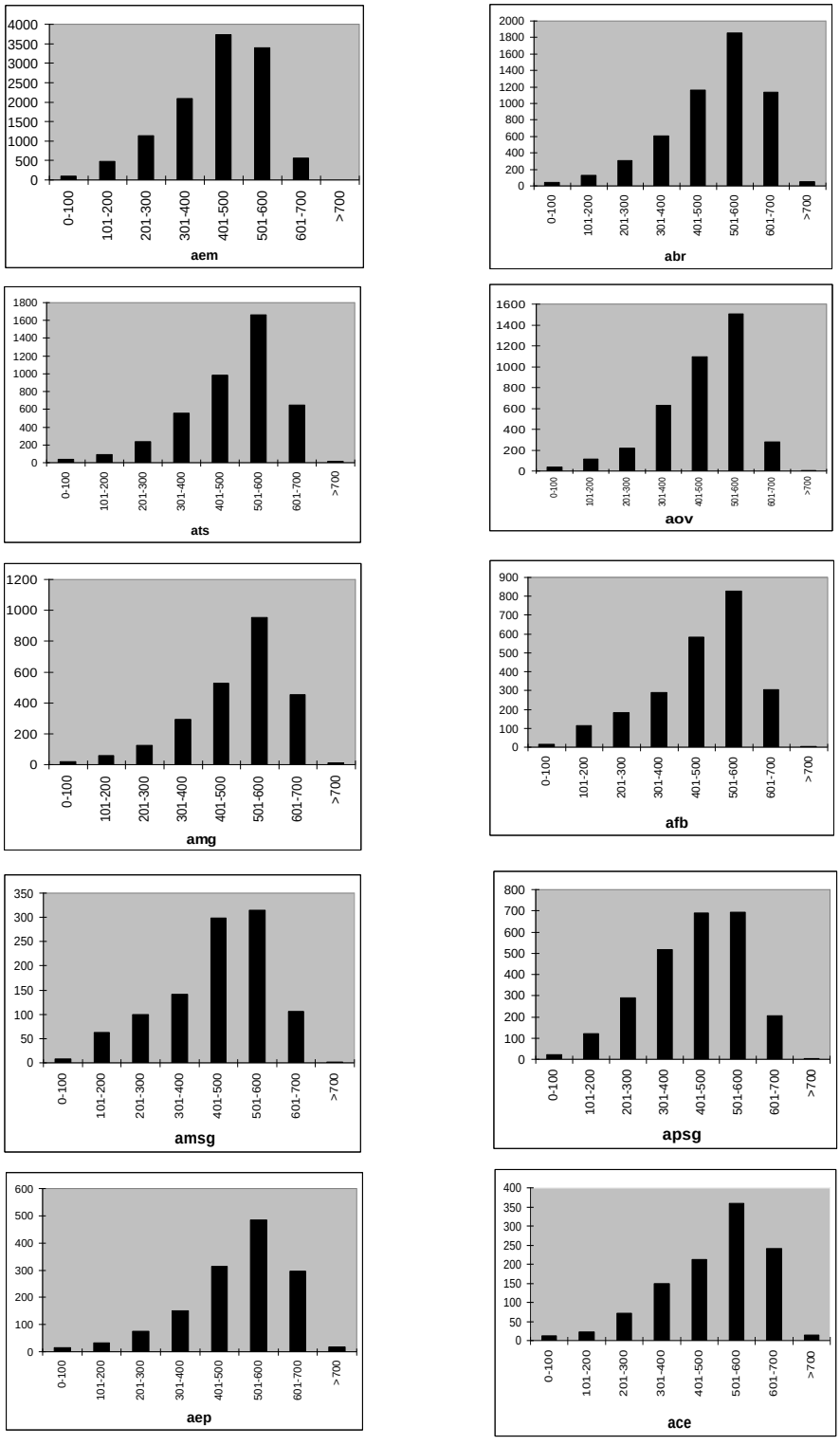


Figure 4: Distribution of read lengths of sequenced *A. assama* ESTs.

X axis- aem: embryo, abr: brain, ats: testis, aov: ovary, amg: midgut, afb: fatbody, amsg: middle silkgland, apsg: posterior silkgland, aep: epidermis and ace: compound eyes ESTs. Y axis represents number of ESTs falling into particular length category.

BLAST analysis of unigenes

Functional annotation based on BLAST against NCBI protein database revealed the presence of several genes that are involved in silk production, circadian rhythm, sex determination, immune response and also several novel tissue specific genes. We identified twelve testis specific putative transcripts based on similarity to NCBI protein database. We were able to align ~1900 bp length of *fibroin* gene by clustering and assembling silkgland ESTs. A homologue of *intersex*, a gene regulating sexual differentiation in *Drosophila melanogaster* females, was also identified among ESTs from ovaries. *Timeless*, a gene involved in regulating circadian rhythm was also detected from the brain.

Sex determining genes

Sexual dimorphism is established by regulatory hierarchies that switch the initially ambiguous embryo to the appropriate path of sexual development. The initial cues that trigger this process, which can be either genetic or environmental, and the regulatory pathways that respond to these cues, vary remarkably between species. Sex-determining mechanisms evolve rapidly, possibly more so than other important developmental processes. Molecular similarities between sexual regulatory genes in different phyla have begun to suggest that sex determination might be ancient but highly diverged, and seem to support evolutionary models in which sex determination pathways are formed by the sequential addition of upstream regulators (Zarkower, 2001). Even within the class of insects alone, sex can be determined by amazing variety of mechanisms: by Y(W) chromosomal factors, by autosomal factors, by the number of X chromosomes, by haploidy versus diploidy, or by infection of microorganisms.

Lepidoptera, unlike Diptera, have female heterogamety wherein females carry ZW and males with ZZ sex chromosomes. In Lepidoptera, the W chromosome has been found only in suborder Ditrysia to which most moths and butterflies belong, including silkmoths (Traut and Marec, 1996). It is speculated that the W chromosome has been acquired after the divergence of Ditrysia from other suborders of Lepidoptera (Traut and Marec, 1996). In spite of the fact that the W chromosome is recently acquired, it has assumed a very important role in *Bombyx* sex determination. The W chromosome possesses a strong ability to determine the female sex in *Bombyx* (Hasimoto, 1933). However, *A. assama* is unique in that it lacks W chromosome. Therefore it will be interesting to study the sex determination mechanism in this species to identify both upstream and downstream genes involved in sex determination in ZZ/ZO background. Initial search for sex determining genes in EST dataset revealed six ESTs (Ac. Nos.

Aaov1642, Aaov0427, Aaov0387, Aaov1877 and Aaov2957 from ovary ESTs, and Aabr4525 from brain ESTs), that showed similarity to *intersex*, a gene involved in female sex differentiation in *Drosophila*. Surprisingly no ESTs were found matching to *doublesex* (*dsx*), an ubiquitously expressed downstream gene in the sex determining pathway known to be highly conserved across insect taxa (Raymond et al., 1998). However, the other downstream *dsx* target genes in sex determination pathway, such as *vitellogenin* (Ac. No. Aabr2281) and *storage protein* (Ac. No. Unigene_Aa00283) were identified. We also obtained an EST matching to *fruitless* (Ac. No. Aabr4886), a gene required for proper development of several anatomical structures necessary for courtship, including motor neurons which innervate muscles needed for fly sexual behaviors.

Analysis of silk genes

Silk fibers are produced from various types of ectodermal glands in mites, spiders and several groups of insects (Craig, 1997). The mechanism of silk production in the paired labial gland of caterpillars and the structure of spun-out filament have drawn the attention of biologists and mechanical engineers but neither of these two processes is fully understood to date. The silk gland, essentially a modified salivary gland, is a highly specialized organ whose function is to synthesize silk proteins. Several hormone-processing enzymes are active in silk gland, which are of interest because these hormones participate in regulation of silk protein genes. So far most of the work on silk genes comes from the silkworm, *B. mori*. The silk of *B. mori* is composed of two groups of proteins: the fibroin and linked subunits which constitute the silk thread and are synthesized exclusively in the posterior silk gland, and the sericins, a family of proteins which ensure the cohesion of the cocoon by sticking the silk threads together are produced only in the middle silk gland. Fibroin is a very large insoluble polypeptide and is difficult to purify from the cocoon without degradation. The sericin fraction of silk is composed of different but similar peptides, characterized by their solubility in hot water which enables the industrial degumming of the silk threads and by their unusual amino acid composition, where serine represents about one residue out of three (Gamo et al., 1977, Lucas and Rudall, 1968).

Silk genes of *A. assama* have not been studied so far due to non-availability of gene or protein sequence information. To identify the genes coding for different silk proteins in *A. assama*, more than 3600 ESTs were generated from middle and posterior silkglands of fifth instar larvae (Table 3). After clustering and assembly 441 and 215 unique sequences were obtained from posterior and middle silkglands, respectively. These sequences were searched for silk genes, which resulted in the identification of a contig of 1.8 kb, showing similarity to carboxy terminus of *A.*

pernyi fibroin. A few more unique sequences showing similarity to fibroin were also obtained, one of which was mapped to N-terminus region. Further mining for other silk proteins resulted in identification of 2 ESTs (Acc. nos. Aamsg0714 and Aapsg0515) homologous to repetitive region of *B. mori* sericins and 9 ESTs showing similarity to non-repetitive regions. However, homologues of light chain fibroin and P25 proteins were not detected in silkgland ESTs. Therefore a global search was carried out in the complete EST dataset of *A. assama*, which revealed the presence of a brain transcriptome derived EST, homolgous to P25. None of the ESTs showed similarity to fibroin light chain protein.

***Fibroin* gene**

All known H-fibroins include a conserved non-repetitive amino-terminus, a large internal region composed of repeats, and a non-repetitive carboxy terminus. The repetitive region consists of the hierarchical arrangement of species specific motifs. The non-repetitive N-terminus begins with a 17-mer or 18-mer signal peptide, which is nearly identical in all examined species (Sehnal and Zurovec, 2004). To date no fibroin genes have been fully structurally characterized because of its instability when cloned (Sezutsu and Yukuhiro, 2000).

Genes that encode proteins with repetitive structures are often called coding-minisatellite sequences (Paulsson *et al.*, 1992). They seem to have properties similar to minisatellite sequences, which are tandemly repeated sequences of 10–100 bp units (Jeffreys *et al.*, 1985). Minisatellite sequences are highly unstable components of genomes. Getting full length gene sequence of fibroin is a difficult task as major chunk of the transcript has many tandemly repeated blocks of nucleotide sequences and it is difficult to maintain the plasmid clones harboring such inserts. Recently, the entire sequence of a fibroin gene for two of the saturniid species, *A. pernyi* (Sezutsu and Yukuhiro, 2000) and *A. yamamai* was determined, through genomic library screening and sequencing.

Attempts were made in the present study, to align full length fibroin transcript using EST and other *de novo* sequenced DNA fragments. For this purpose 16 fibroin EST clones were cultured and plasmids were checked for insert sizes. Out of these, two clones, H14 (1.5 kb) and K24 (1.2 kb) were selected. Other clones were excluded as they contained very short inserts. From H14 clone, 250 bases and 593 bases were sequenced with M13 forward and reverse primers respectively. These sequences were aligned with the other fibroin ESTs to get 816 bp and 679 bp

sequences respectively. The alignment was further extended manually using a few more fibroin EST sequences to get a contig of 1825 bp mapping to 3' end of fibroin.

The clone K24 was found to contain a gene insert that mapped to the 5'end of the *fibroin* gene. After sequencing the clone with M13 forward and reverse primers, 944 bp sequence was obtained. This sequence was used to pull out overlapping ESTs from silkgland EST dataset, which resulted in extension of sequence of length 1,674bp. Based on the conceptual translation using Expasy (http://www.expasy.ch/tools/dna.html) and BLASTX results, the position 766-768 of this 1,674bp contig was predicted to be translation start site of the *A. assama* fibroin gene.

Northern blot analysis carried out to study the expression pattern and transcript size showed the fibroin gene expressing exclusively in posterior silkgland and length was calculated to be between 8-8.4kb (Figure 5). Fibroin was found to be expressed only in posterior silkgland in other silkmoth species. Transcript size of *A. assama* fibroin was similar to that of other saturniid species, *A. yamamai* and *A. pernyi* (Sezutsu and Yukuhiro, 2000).

To estimate the fibroin protein, size the *A. assama* fibroin was extracted from the degummed cocoon using LiSCN. After resolving on SDS Polyacrylamide gel, its size was found to be around 230 KDa. When compared to fibroin proteins size of other silkmoths species the protein size of *A. assama* was found to similar to that of other saturniid silkmoths but not of *B. mori* (Figure 5). Complete fibroin gene was amplified using long PCR protocol and it gave a single amplicon of size around 9Kb. This size range was also similar to that of other satuniids. However, the gene size did not match with that of *B. mori* (Figure 5).

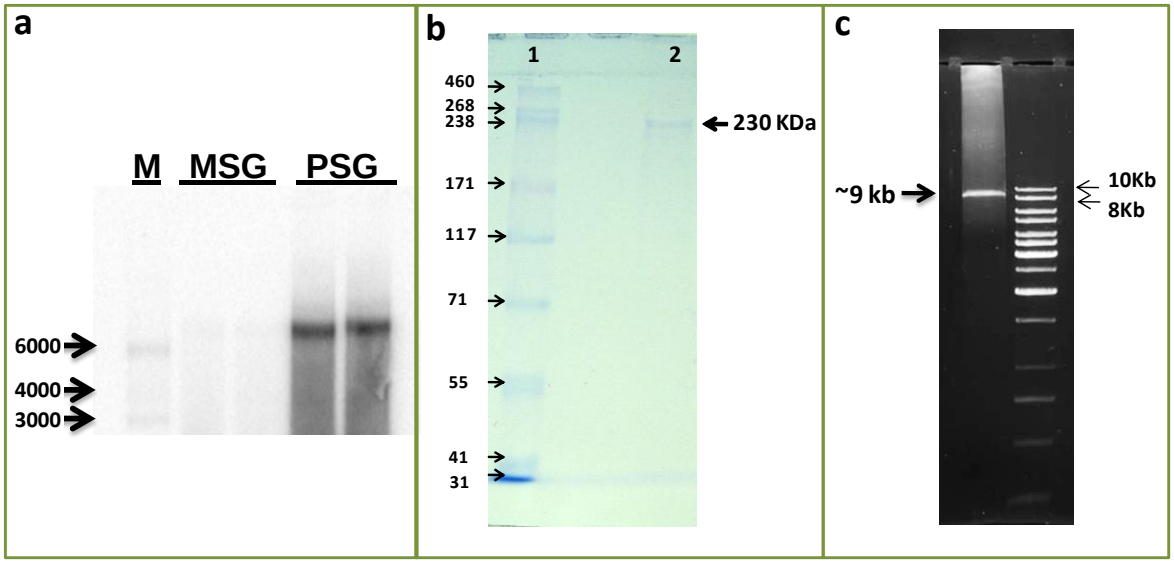
Repeat block analysis: In saturniid silkmoths, the basic repeat units of H-fibroin are 21-40 residues long. Each has a track of 4-14 alanines followed by one of the 6 different types of non-crystalline motifs that are rich in G, Y, and S (Sezutsu and Yukuhiro, 2000). Using the amino acid sequence obtained after conceptual translation of *A. assama* fibroin contigs, repeat block analysis was done and was compared with that of other silkmoths. Like in other Saturniids, it has got a repeat motif consisting of a polyalanine block (PAB) followed by a non-polyalanine block (NPAB). In *A. assama* fibroin, $(A)_{8-14}$ followed by NPAB was the pattern observed. The first segment of the NPAB was mostly GSGAGG and the last segment was mostly GYGSGS/GYGSGSS with a dissimilar middle segment. Totally 14 motifs were found in the 608 amino acids stretch present at the 3' end. In case of *A. pernyi*, 78 motifs were found (Fedic *et al.*, 2002), with first segment having mostly GSGAGG/SGAGG and the last segment GGYGSDS. Most of the PABs contained 12 or 13 'A' residues. In *A. yamamai*, 77 motifs were found (Fedic *et al.*, 2002) with first segment having

mostly GSGAGG/SGSGG and the last segment GGYGSDS (Figure 6). This shows that the first segment may share similarity in all the three species. However, in the last segment there is similarity only between *A. pernyi* and *A. yamamai*. The repeat pattern seen in *B. mori* and *G. mellonella* are different. The *B. mori* H- fibroin crystalline domain consists largely of Glycine-X dipeptide repeats in which X is Alanine in 65% of cases, Serine in 23%, and Tyrosine in 9% (Zhou *et al.*, 2001). In the pyralid moth *G. mellonella*, the repetitive region has 3 repeats designated as A (63 residues), B1 (43 residues), and B2 (18 residues). These 3 domains are assembled into higher order repetitions AB1 and AB2 which in turn are reiterated into 11 large domains $[AB1AB1AB1AB2(AB2)AB2]_{11}$ that constitutes more than 95% of the H-fibroin molecule (Fedic *et al.*, 2002). Last two NPABs of fibroins of the three saturniid silkmoths were found to be highly conserved with only a few differences. The sequence conservation of this region indicates its functional significance (Sezutsu and Yukuhiro, 2000).

Sericins

Sericin is a group of serine-rich cocoon proteins produced by the silkworms. It is known as glue protein, which binds fibroin fibers. Since the 1930s, several studies have been conducted on the complexity of sericin. Sericins are well studied in *B. mori*. Three genes coding for sericin proteins have been identified and characterized in *B. mori* (Garel et al., 1997, Okamoto et al., 1982, Takasu et al., 2007). The most abundant silk sericins of *B. mori* are encoded by the single gene *Bombyx sericin 1* (*Serl*) which gives rise to mature mRNAs through differential splicing of the primary transcript (Couble et al., 1987, Garel et al., 1997). This maturation is a tissue- and developmentally- regulated process and the corresponding sericin proteins can be visualized as distinct layers piling up around fibroin in the middle silkgland. The structure of this gene is characterized by the presence of a large central alternative exon which encodes an internally repetitive sequence.

Not much is known about sericin genes of saturniid silkmoths except for two studies on sericin protein extraction from cocoons (Ahmad et al., 2004, Dash et al., 2006). In the present study two sericin transcripts possibly originating from 2 genes were identified through BLAST search of *A. assama* silkgland ESTs. These were named *AaSer1* (Acc. no. Aamsg0714) and *AaSer2* (Acc. no. Aapsg0515). Both had 38 amino acid repeat motifs, a characteristic feature of *Ser1*. There was no significant similarity between repeat sequences of *AaSer1* and *AaSer2*. However, both the sequences were rich in serine residues.



d

Scientific name	Acc. No.	Gene Size (bps)	Transcript size (bps)	No. of amino acids	Protein molecular weight (KDa)
Bombyx mori	AF226688	80009	15792	5263	391
Antheraea pernyi	AF083334	9282	7920	2639	220
Antheraea yamamai	AF325500	9921	7968	2655	250
Antheraea assama	This study	9000	8000	2600	230

Figure 5: a) Northern blot showing the expression of *A. assama* fibroin gene only in posterior silk glands. M: RNA marker (numbers are in bases), MSG: middle silkgland, PSG: posterior silkgland. b) SDS-PAGE profile showing the presence of single fibroin protein with size 230KDa. 1 - 460 KDa High molecular weight protein size marker, 2 - fibroin from cocoon. c) Amplification of complete fibroin gene using genomic DNA as template and primers binding to 5' and 3' end of fibroin gene. d) Table shows the gene size, transcript size, no. of amino acids in the protein and protein molecular weight of *A. assama,* in comparison with other silkmoths.

To get full length gene sequences of sericins the EST clones harboring the sericin cDNA inserts were checked for the insert size and clones with longer insert were sequenced to get complete sequence of the insert. The EST clone containing *AaSer2* cDNA (clone ID. I02) was sequenced through primer walk to get complete sequence of the insert. Initial sequence of 673bp was obtained by sequencing the plasmid using M13 reverse primer. Then primer walking was done by designing primers appropriately to the ends of available sequence. Two such walks were done using primers AaSer5 and AaSer6. AaSer5 was designed towards the 3' end of the 673bp contig for primer walking, which yielded an additional sequence of 338bp. Another primer AaSer6 was designed to continue the primer walking, which gave additional 374bp. The sequences were corrected for errors and aligned using CAP3 program to get 1384bp 3' end of *AaSer2*.

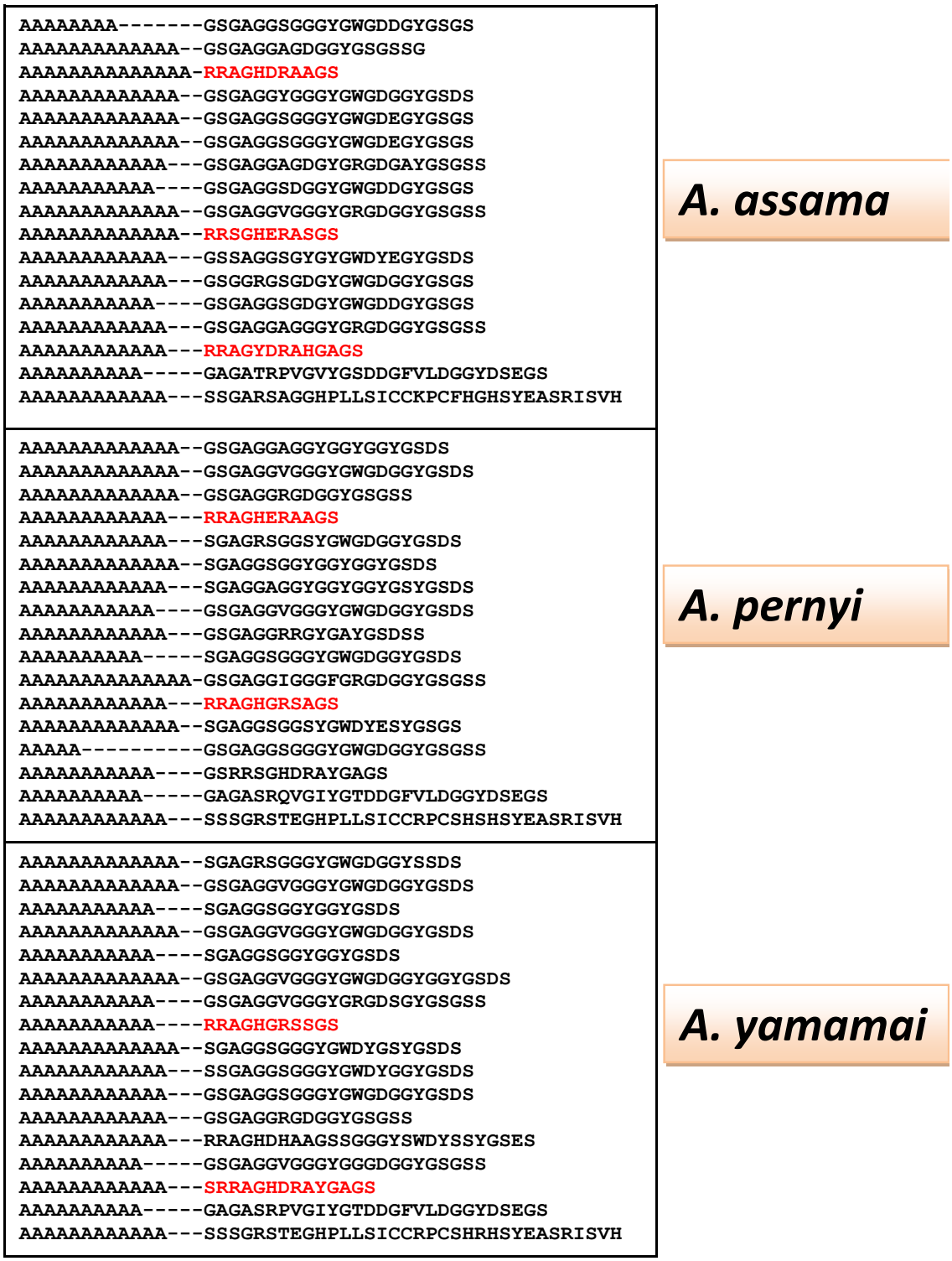


Figure 6: Carboxy termini of fibroin proteins of the three saturniid silkmoth species, *A. assama*, *A. pernyi* and *A. yamamai*.

B06 (containing *AaSer1* cDNA) clone harbored insert of ~500bp. On sequencing this plasmid, 483 bp was obtained. In order to get the remaining sequence of *Sericin* mRNA, the RT product of MSG was subjected to PCR using gene specific primers, AaSer1F and AaSer1R. Multiple bands were generated out of which, four products were cloned into TA vector, transformed and screened for appropriate clones. The inserts were then sequenced and aligned to get longer sequence of *AaSer1*.

Amino acid repeat motifs of AaSer1 were compared with that of *A. mylitta* (AmSer1) and *B. mori* (Ser1). The comparison showed the conservation of four serine residues at a particular position in sericin of all the three species. This shows that serine residues at these positions are indispensible for the structure and function of sericin proteins in silkmoths. A 'TSG' motif was found to be conserved in all the three sericins. Twenty out of 38 amino acid residues of repeat motif were found to be conserved between the two saturniids, *A. assama* and *A. mylitta*. Among all the amino acids, Threonine residues were the highest number to be conserved between the two sericins. Only 7 residues were conserved among the repeat motifs of all the three sericins, AaSer1, AmSer1 and Ser1 (Figure 7).

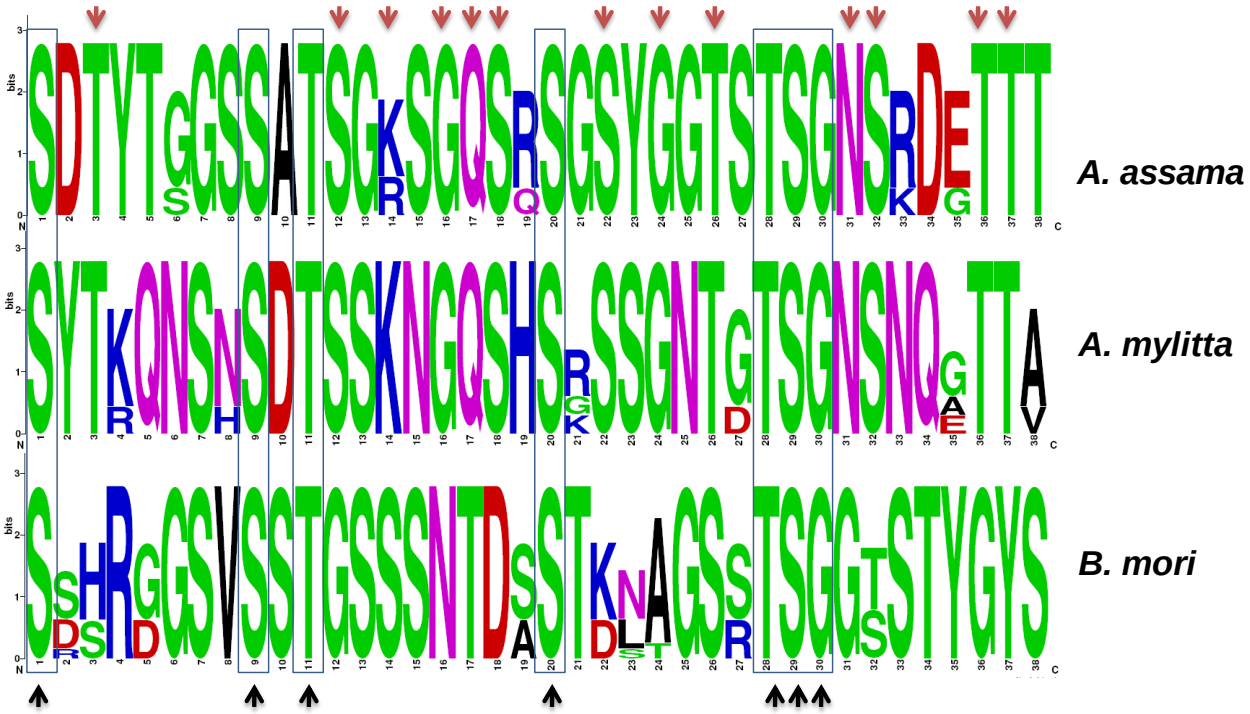


Figure 7: Comparison of 38 amino acid tandem repeat motifs of *A. assama* sericin 1 (AaSer1), *A. mylitta* sericin 1 (AmSer1) and *Bombyx* sericin 1 (Ser1). Sequence of AmSer1 was obtained by BLAST search of *A. mylitta* ESTs in WildSilkbase. Ser1 repeat motif was downloaded from Genbank. Black arrows indicate the amino acids conserved in all the three silkmoths and the red arrows indicate the amino acids conserved only between *A. assama* and *A. mylitta*.

Conservation of silk genes in lepidopteras

Previous reports on identification of homologous silk genes in *G. melonella* suggested that the formation of silk filament from H-fibroin, L-fibroin and P25 is widespread in Lepidoptera. The orthologs of *B. mori* H-fibroin gene were found in three saturniid species and in pyralids, *G. melonella*, *Ephestia kuehniella* and *Plodia interpunctella*. The H-fibroin was also found to be conserved in *A. assama*, however, with respect to N and carboxy terminus, though repeat structure differs. The L-fibroin and P25 genes were detected in *G. melonella*, *Dendrolimus spectabilis* and *Papilio xuthus*. However, all efforts to detect L-fibroin and P25 homologues failed in the silkmoths belonging to saturniidae. It seems likely that the core of saturniid silk filament consists of H-fibroin homodimers.

The N-terminus of fibroin from several silkmoths and spidroins from *Naphila clavipes* were compared through phylogenetic analysis. A dendrogram was generated to study the pattern of evolution of fibroin in these species (Figure 8). *A. assama* was found to represent basal lineage of saturniid species. *G. mellonella* and *Y. evonymellus* were close to saturniids than Bombycoid species. However, they formed a separate cluster. Pairwise distance matrix revealed that Spidroin 2 of *N. clavipes* was more close to *Bombyx* fibroins, whereas Spidroin 1 was more close to fibroins of saturniid silkmoths.

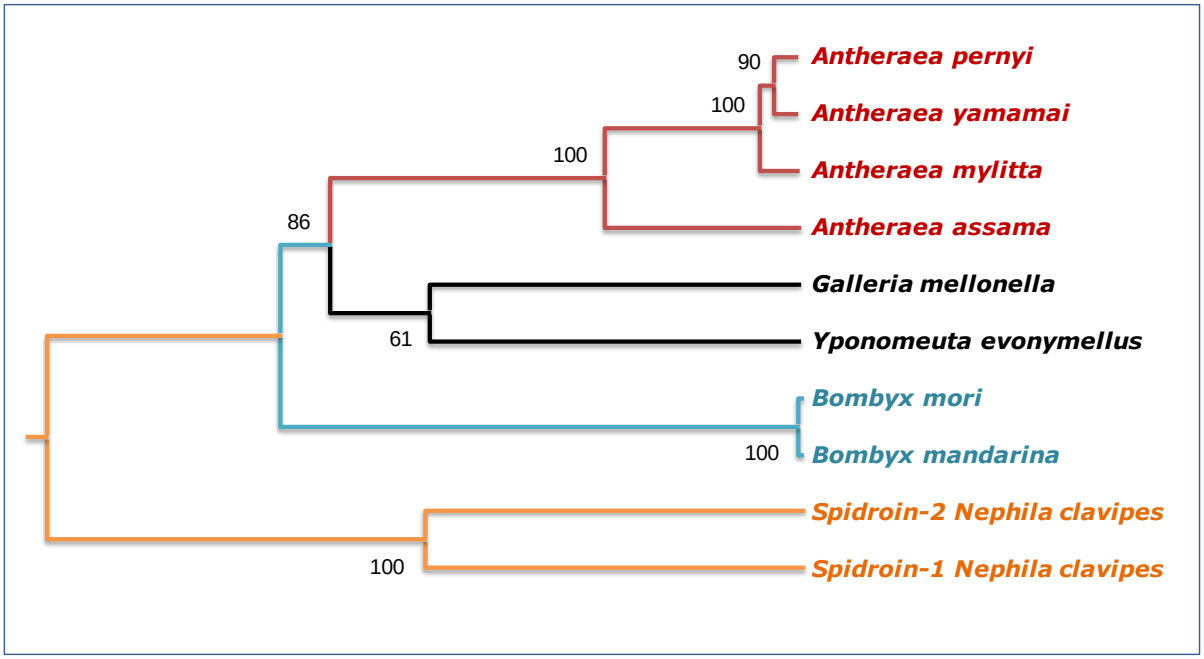


Figure 8: Bootstrap UPGMA tree generated using N-terminal region of *fibroin* and *spidroin* genes.

Cuticle genes of *A. assama*

Cuticle is the protective shield of Arthropods. Its basic constituents are chitin (the polymer of N-acetylglucosamine) and proteins (Cohen, 1987). These two macromolecules are combined to build a complex structure, not only strong, but flexible as well. The architecture of cuticle is helicoidal and most probably this building plan is responsible for its extraordinary mechanical, thermal and physiological properties (Neville, 1975). In this helicoidal structure, chitin is in the form of crystalline filaments and proteins play the role of the matrix. Although chitin is a simple polysaccharide, the second constituent of cuticle i.e., proteins, present a great variety. Dozens of different proteins may appear in the cuticle of a single organism on a single moment. These proteins may differ between developmental stages or, most commonly, their quantitative distribution may change. It seems that the combination of specific proteins at specific quantities is unique for each region of cuticle, for each developmental stage and for each species (Willis, 1996).

Insect cuticle is constructed from many cuticle proteins with various combinations of different temporal and spatial patterns (Willis, 1996). The amino acid sequences of more than 100 cuticular proteins have been determined in insects. These cuticle protein sequences have been shown to have a much conserved hydrophilic region (Andersen et al., 1995, Rebers and Riddiford, 1988). Epidermal cells produce larval, pupal, and adult cuticle proteins from the same cells in lepidopteran insects.

Manual inspection of BLAST results of 96 hours embryo derived *A. assama* ESTs showed high occurrence of genes coding for cuticular proteins. It is known earlier that insects produce many different cuticle proteins at one stage from the same tissue (Andersen *et al.*, 1995). Therefore, a global search for cuticle protein coding genes in *A. assama* was carried out, which resulted in identification of 114 putative unigenes showing similarity to cuticle genes. Upon retrieving and analyzing all the sequences forming these putative unigenes it was found that a total of 4063 ESTs code for cuticle proteins. These ESTs originated from all the 10 tissue transcriptomes of *A. assama* but with varied proportion. Surprisingly 3427 ESTs of them were embryo derived, which accounted for 84% of the total cuticular gene derived ESTs in the total dataset. The fifth instar larval epidermis transcriptome showed presence of 303 ESTs (7% of the cuticular gene derived ESTs) coding for cuticle proteins. The remaining tissue transcriptomes however showed less number of cuticular protein coding ESTs.

Patterning of the embryonic epidermis in Drosophila has been studied extensively by examining the cuticular pattern elements deposited by epidermal cells late in development. These

structures serve as indelible markers of cell fates within the epidermis and have proven invaluable in genetic screens designed to identify mutations that disrupt cell fate decisions (Nusslein-Volhard and Wieschaus, 1980). Results from the present study show that in the late embryonic development there is high expression of cuticular proteins in *A. assama*. Among all the tissues, highest number (11500 ESTs) of ESTs was generated from 96 hours embryos, around 30% (3427 ESTs) of which is composed of ESTs coding for cuticular proteins. The analysis also revealed that most (75 out of 114) of the putative unigenes were specific to embryo. This shows that cuticular genes are highly active in 96 hours embryonic tissues and many are specific to embryo.

All structural cuticular proteins are small (with an average length of 100-200 aminoacid residues), they lack cysteine and share several characteristic motifs. These motifs may be small and repetitive or large occurring only once (Andersen *et al.*, 1995). The most well known motif was recognized by Rebers and Riddiford, and named “R&R motif” (Rebers and Riddiford, 1988). This 68 amino acid region, named the "extended R&R consensus" is what is recognized by PF00379, the Pfam motif for chitin binding of Arthropod cuticle (Andersen et al., 1995, Rebers and Willis, 2001). Three types of the "extended R&R consensus" have been found: RR1, RR2 and RR3 (Andersen, 1998, 2000). The two main types are RR1 and RR2, which presumably appear in proteins from soft and hard cuticles, respectively (Andersen, 1998). Very few cuticular protein sequences show the existence of RR3 (Andersen, 2000).

All the cuticle genes identified in the present study were further classified into RR1 and RR2 type cuticle proteins. A total of 33 and 52 putative unigenes were grouped under RR1 and RR2 types respectively. The remaining 29 did not fall into any of the categories. Some of them may fall into RR3 but could not be grouped with certainty, into any category possibly because of their short sequence length. Results of the RR1 and RR2 predictions can be used as a guide for identifying a certain protein as coming from either soft or from hard regions of the cuticle. Most importantly, the information about the RR1 and RR2 distinction can be used for studies of cuticle's mechanical properties. As RR1 and RR2 proteins appear in soft and hard cuticles respectively, the former interact with chitin more loosely than the latter (Magkrioti *et al.*, 2004).

In *A. assama*, RR2 type putative unigenes were represented more than RR1, though EST composition of both the types was almost equal. Thirty three RR1 type proteins were resulted from 1883 ESTs, of which 1368 were derived from embryo. A total of 1793 ESTs constituted 52 RR2 type proteins, which included 1683 from embryo. A small portion of ESTs (387) could not be categorized into any group and they formed 29 putative unigenes. The analysis shows that RR2

type proteins are expressed highly in late embryo stage, whereas RR1 types are expressed more in other tissues when compared with RR2 type proteins.

As *A. assama* silkmoths complete whole of their life cycle from egg to adult, outdoors in forests, they require strong protection from the damages caused due to environmental changes. They are amenable to several kinds of pathogen attack. To protect themselves from these, it seems they have evolved a system wherein they will produce a battery of cuticle proteins in later stages of embryo development, so that when larvae hatch out they will have strong protective cover. This may be one of the reasons for higher and specific expression of several cuticle genes in late embryonic stages. The fact that many of the expressed cuticle genes are RR2 type shows that, epidermis needs to be hard and less flexible to guard from environmental hazards. Studies in *D. melanogaster* showed that the embryonic cuticle is deposited by the epidermal epithelium during stage 16 of development. This tough, waterproof layer is essential for maintaining the structural integrity of the larval body (Ostrowski *et al.*, 2002). In the same study investigators have identified a set of genes that are required for proper development of the *Drosophila* embryonic cuticle. Mutations in these genes are zygotic lethal and result in flaccid embryos with very elastic cuticles that stretch to a remarkable degree when flattened beneath a coverslip (Ostrowski *et al.*, 2002).

All 114 putative unigenes may not be derived from different individual genes. It is possible that two or more putative unigenes may be partial sequences of a single gene. If there is no overlapping sequence between them, they will be represented as different putative unigenes. Therefore the number of cuticle genes will be less than the total number of putative unigenes coding for cuticle proteins. In CuticleDB, there are 39 entries of cuticle genes derived from lepdiopteran insects (Magkrioti *et al.*, 2004). The cuticular gene sequences identified in the present study will add to existing sequence data and would be useful in further studying the architecture of cuticle in lepidopteran insects.

Microsatellite markers from *A. assama*

Over the past two decades there have been reports that the natural populations of *A. assama* have experienced declines that are attributed to continued deforestation activities, population fragmentation and inbreeding. The decline in muga silkworm population has raised concern among silkworm farmers and conservationists. Therefore in this study informative microsatellite markers were developed, which serve as useful markers for analyzing genetic structure and phylogenetic status of this species. With the knowledge of information and importance of

microsatellites that reside in the exonic parts of the genome of various organisms, EST-SSR markers were developed for *A. assama.* In addition, microsatellite markers were also developed from repeat enriched genomic library constructed from *A. assama* DNA.

A total of 63 sequences was selected from unique ESTs and of which 35 were di-, 12 tri-, 4 tetra- and 2 pentanucleotide repeat containing sequences. Ten ESTs harbored compound repeat motifs. From repeat enriched genomic library, 134 of the 185 colonies screened contained inserts of more than 500bp. Only 24 of the 134 inserts were harboring microsatellite repeats.

Of the 63 EST-SSRs and 24 genomic SSRs, 51 and 19 gave amplification respectively, as assessed by 1.5 % agarose gel electrophoresis. These 70 markers were then tested on 40 *A. assama* individuals collected from Tura region (20 Nos.) and West Garo Hills (20 Nos.) located 60 Km apart in Northeast India. The sampling locations are 60 km apart. Only 10 of the 51 EST-SSRs werefound to be polymorphic. Twenty eight loci were monomorphic with single band pattern. Thirteen markers gave bands with size range more than expected and were monomorphic. The possible reasons for unexpected size range of EST-SSR loci are, one or both primers of the EST-SSR extend across a splice site and presence of introns and insertions-deletions (indels) in the corresponding genomic sequence (Varshney *et al.*, 2005). Of the 19 genomic SSRs, only 1 locus was polymorphic with 3 alleles across 40 individuals. The observed and expected heterozygosities ranged from 0 to 0.737 and 0.095 to 0.825 respectively with the allele numbers of 2 to 8. In Tura population, we observed a significant linkage disequilibrium for 3 of the 35 pairwise comparisons between loci after sequential Bonferroni correction, particularly involving loci AaSat002, AaSat020, AaSat044 and AaSat065. However, only one pair (AaSat001 & AaSat053) out of 27 combinations showed significant deviation from linkage disequilibrium in West Garo Hills population. The linkage disequilibrium observed at certain loci may be due to substructure of population or bottleneck. Only three of the 9 loci polymorphic in Tura population, showed no significant deviation from Hardy–Weinberg equilibrium (HWE). The remaining 6 deviated significantly from HWE and three of them (AaSat002, AaSat020 and AaSat053) showed large heterozygote deficiencies. Eight microsatellite markers were found to be polymorphic in West Garo Hills population, of which only two were in HWE (Table 4). The three loci AaSat002, AaSat020 and AaSat053 that deviated from HWE also exhibited overall significant excess of homozygotes with null allele frequency of 0.435, 0.366 and 0.198 respectively, possibly indicating the presence of null alleles in Tura population. In the West Garo Hills population four loci AaSat008, AaSat14 AaSat044 and AaSat040 that also deviated from HWE showed homozygote excess with null allele frequency of 0.2401, 0.3714, 0.4527 and 0.4438 respectively. Further, in the West Garo Hills

population, only less than half of the samples was successfully genotyped with markers AaSat040 and AaSat044 suggesting a high frequency of null alleles in this population. Taken together, these results suggest that the deviation from HWE of many loci may be due to Wahlund's effect caused by subpopulation structure (Wahlund, 1928) which has to be studied further. The loci which showed significant linkage disequilibrium and deviation from HWE in Tura region population, did not exhibit same features in West Garo Hills population, indicating that the 11 microsatellite markers developed in the present study are independent and are useful for population genetic studies.

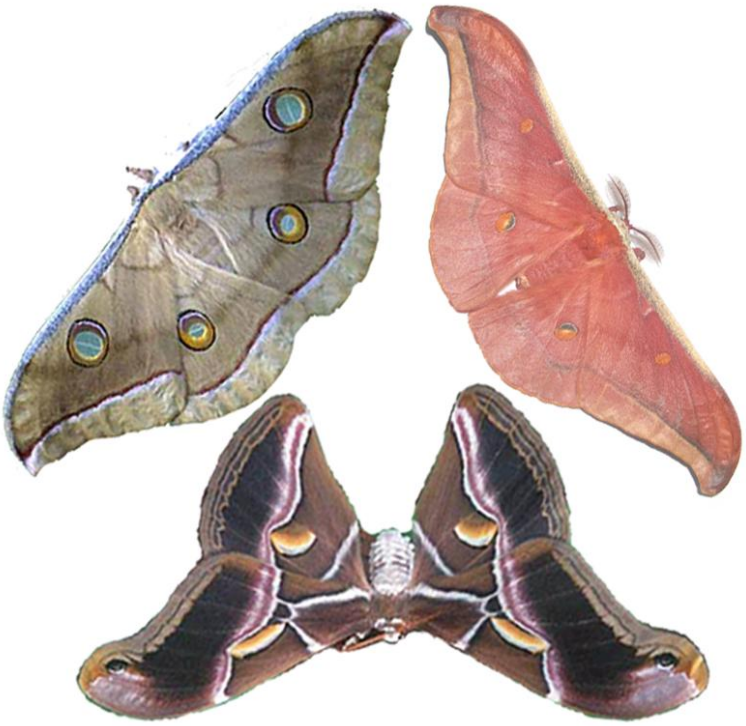
The polymorphic EST and genomic-SSRs developed in the present study will be particularly useful to study inbreeding effects, population structure and founder effects. As there is no microsatellite markers developed in other saturniid moths, these markers can also be tested for their cross-species amplification.

Table 4: Characteristics of 10 EST- derived and one genomic SSR markers of *Antheraea assama*, Ta, annealing temperature, N, number of individuals scored, Na, number of alleles, Ho, observed and HE, expected heterozygosity. *Significant deviation from HWE (P<0.001). Data obtained upon screening 40 individuals (20 individuals each from Tura and West Garo Hills populations).

Locus (Genbank)	Source	Forward/reverse primers (5'-3')	Repeat motif	Size range (bp)	*Ta*	Tura population			West Garo Hills population		
						N/ *Na*	H_E	*Ho*	*N*/ *Na*	H_E	*Ho*
AaSat001 (EU597692)	EST	GTGTTCATTTCACGGAACATT CATTCGCTGTTCGTCTGAGAT	(TA)7	182-200	53	20/2	0.455	0.700*	16/3	0.521	0.438*
AaSat002 (EU597693)	EST	TCTGGACAAATTGTAAAAGCTGTAG ACAAAACGAAAATCGCGTGT	(GTCT)5	130-149	51	19/3	0.609	0.000*	19/6	0.770	0.632
AaSat008 (EU597695)	EST	CACGAAATGCCTCTGTCGTA GGTGTCTGTGGATGATGTGC	(TA)6 Nn (CA)9 Nn (CA)5	199-249	51	Mono	-	-	19/4	0.432	0.211*
AaSat014 (EU597696)	EST	ATCTCTACCTACGCCGACGA AATTCGGCACGAGGAGTTC	(GAT)5 (TAA)4	260-264	48	Mono	-	-	19/2	0.388	0.000*
AaSat020 (EU597697)	EST	TTTCTTCGGTTCGTTTGGTT GACACGCGTTGCTTTGAGTA	(TCGTG)5	164-226	53	20/2	0.375	0.000*	19/8	0.825	0.737
AaSat040 (EU597698)	EST	CGGACGTAACATTTGTCTGG CCACATGACTCTCATCAGCA	(AT)17	133-221	60	16/2	0.430	0.625*	7/4	0.694	0.000*
AaSat044 (EU597699)	EST	CACCAGCTTCCAAAGAATTG CTAAAGCCCACGGGTTCATA	(AT)20	194-226	51	17/3	0.657	0.588	9/3	0.642	0.000*
AaSat053 (EU597700)	EST	GAGTTCGGGTCGGACGTAAT TCTCTACCTACGCCGACGAC	(ATT)4 N7 (ATC)6	204-225	50	20/2	0.095	0.000*	18/5	0.648	0.333*
AaSat059 (EU597701)	EST	CGAATAGCCGATTCCCTTTG TGCAAGCACGCACGTATC	(TGCG)16	103-187	55	17/4	0.715	0.529*	Mono	-	-
AaSat065 (EU597702)	EST	GTCGAGCTGTCATAATTCCT AGTCTGACGTCGCTATAACC	(AT)11 Nn (AT)11 (TA)5	100-168	54	18/2	0.375	0.500	Mono	-	-
AaGSat019 (EU597703)	Genomic	GATGGACTGGACCTCAATCG CCTGAGGAGAGAGGCGATG	(TGA)2 (AGA)3	151-172	55	18/3	0.248	0.278	Mono	-	-

Chapter II

Development of WildSilkbase, an EST database of wild silkmoths



Sudden spurt in sequencing projects in recent years has resulted in exponential increase in the genome sequence repertoire of species that are close relatives of many model organisms. Availability of the sequence resources has accelerated comparative genomic analysis and has thus added to our understanding of organismal biology of these species. In fact new insights into human genome have come only after the sequences of related species such as chimpanzee, monkey and ape were published. However, in the insect order Lepidoptera, which consists of many economically important insects such as silkmoths, agriculture pests and beautiful butterflies, only the domesticated silkworm, *Bombyx mori* has achieved the distinction of the most well studied insect next only to Drosophila. Therefore, comparative genomics in this order is still in its infancy.

Functional genomics has particular promise in silkworm biology for identifying genes involved in synthesis and secretion of silk, pathways involved in processes such as insect-pathogen interactions and sex determination mechanisms. Among lepidopterans, several databases such as Silkbase (Mita et al., 2003) for ESTs, Kaikobase (Mita et al., 2004) for whole genome sequence, Silkdb (Wang et al., 2005) for ESTs and whole genome sequence, and Silksatdb (Prasad et al., 2005) for microsatellites, have been developed for *B. mori*. Apart from *B. mori*, more than 13,000 ESTs have been made available for butterflies through Butterflybase (Papanicolaou et al., 2008) and 32,217 ESTs for the pest species, *Spodoptera frugiperda* through Spodobase (Negre et al., 2006). However, wild silkmoths are least represented owing largely to non-availability of genomic resources in these species. Hence, generation and utilization of genomic information from wild lepidopterans will be extremely useful in understanding these species at molecular level.

The members of family Saturniidae, collectively known as saturniids, are among the largest and most spectacular of the Lepidoptera, with an estimated 1,300 to 1,500 different described species distributed worldwide (Grimaldi and Engel, 2005). The Saturniidae family includes the giant silkmoths, royal moths and emperor moths. The muga silkmoth, *Antheraea assama* (n= 15), confined to the North-eastern states of India, is the least understood and unique species among saturniid moths. The silk proteins of this species have not been studied so far despite their unique properties of providing golden luster to the silk thread. *Samia cynthia ricini* (n= 13) a multivoltine silkworm commonly called as 'eri silkworm' is known for its white or brick-red eri silk. It is distributed in India, China and Japan. Its ecoraces (~16) are distributed across the Palaearctic and Indo-Australian biogeographic regions. The tropical Indian tasar silkmoth, *Antheraea mylitta* (n= 31) is a natural fauna of tropical India, represented by more than 20 well-described, genetically distinct ecoraces.

The rapid and ever increasing deposition of ESTs of various organisms in dbEST database of NCBI shows the role of EST projects in advancing genomic science in eukaryotes. Considering the potential scientific benefits a total of 57,113 ESTs was generated in three wild silkmoth species, *A. assama*, *S. c. ricini* and *A. mylitta* from different tissues at various developmental stages (Table 1). These ESTs and resulted unigenes were annotated for Gene Ontology (GO) terms, simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs). To get insight into conservation and divergence of genes among silkmoths and model insects, we compared the putative unigenes of the three wild silkmoth species among themselves and with those of four insect species, *B. mori*, *Drosophila melanogaster*, *Apis mellifera* and *Tribolium castaneum*. In order to make the EST resources of wild silkmoths and their annotations available for the use of scientific community an EST database called 'WildSilkbase' was developed.

Wild silkmoth species

Non-mulberry sericigenous fauna belonging to the family Saturniidae (superfamily Bombycoidea) are mostly wild Silkmoths. These are used as important tools in basic entomological and biotechnological research in various countries. These are medium to very large in size, and this family includes the largest moths. Adults have a wingspan of 3 to 15 centimeters, relatively small heads, and densely hairy bodies. Larvae are usually very fleshy, with clumps of raised bristles. Caterpillars mostly feed on leaves of trees and shrubs; some cause severe damage. Pupa develops in silken cocoons. Wild silkmoths are reared on wild trees but can also be raised and bred under complete human control. They complete their life cycle of four different metamorphosing phases, egg, larva, pupa and adult (moths). The range of food selection of these insects is wide. Their cocoons are bigger than those of the domesticated silkworm.

In the present study three economically important silkmoth species of family saturniidae have been studied. Biology and economic importance of *A. assama* is already explained in chapter 1.

The Indian tasar silkworm, A. mylitta

The word tasar is derived from the Sanskrit word trasara (Shuttle). Tasar silk is mentioned in literature dating back to 1590 B.C. The Indian tasar silkworm, *A. mylitta* is a natural fauna of tropical India. Wide distribution and polyphagy of this insect species has resulted in extensive variation in the population. As many as nineteen well defined ecoraces have been reported in this species which feed primarily on *Terminalia* species and *Shorea robusta*, and also on number of secondary food plants. The ecoraces are uni, bi or trivoltine depending upon the geo-ecological conditions and differ from each other in qualitative and quantitative traits. Tasar cocoons are reported to be largest among all the silk-producing insects in the world (Akai, 2000). Tasar silk fiber has its own distinctive color, is coarse to feel, but has higher tensile strength, elongation, and stress-relaxation values than the mulberry silk fiber secreted by *B. mori* (Iizuka, 2000, Rajkhowa, 2000). These properties have made tasar silk as quite distinct and unique from other silks.

The tasar larvae are stout and smooth, and have rudimentor scoli. The egg is oval, dorsoventrally symmetrical along the anteroposterior axis. About 3mm in length and 2.5mm in diameter, it weighs approximately 10mg. At oviposition it is dark brown owing to the gummy coating of meconium. Two brownish parallel lines along the equatorial plane of the egg divide the

surface into three zones; disk, streak and edge. The larva is typically cruciform and has a hypognathous head with biting and chewing mouthparts. On hatching it is dull brownish yellow with black head. The body normally turns green and the head brown after about 48 hours, but also yellow, blue and almond-colored larvae are also found occasionally in nature. Body coloration is retained throughout the larval period. The pupa is obtect adectious, having a well-defined segmented body. The dark brown pupa weighs about 10g. The ventral genital markings are on eighth and ninth abdominal segments. The cocoon is single-shelled, pendent, oval, closed and reelable, having a hard non-flossy shell with fine grains. At the anterior end there is a well formed dark brown peduncle with a ring at the distal end. The cocoons are generally yellow or grey. The females spin larger cocoons than those of males. The moths exhibit distinct sexual dimorphism. The females are bigger, with a distended abdomen and narrow bipectinate antennae. The females exhibit polymorphic forms of grey and yellow, whereas the males are brown. Yellow and grey males and brown females are rare (Figure 1).

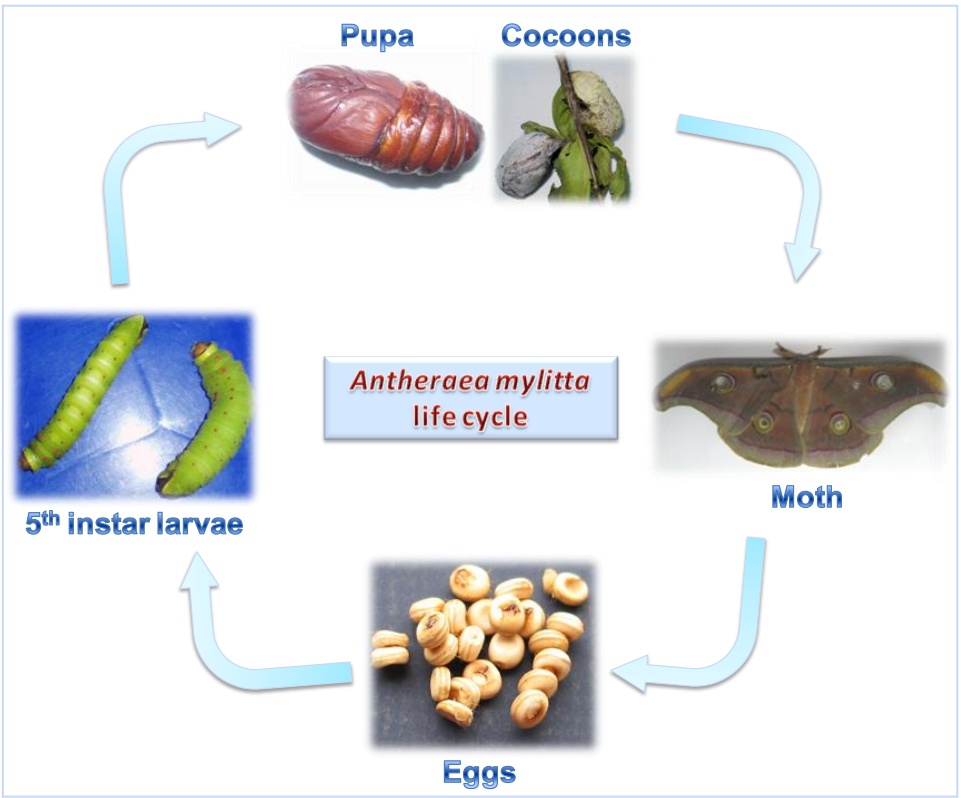


Figure 1: Life cycle of *A. mylitta*

Eri silkmoth, S. c. ricini

The name eri derives from the Assamese word 'era', which means castor-oil plant, the main food plant of this silkworm. *S. c. ricini* (n= 13) a multivoltine silkworm commonly called as 'eri silkworm' is known for its white or brick-red eri silk. It is distributed in North-Eastern part of India, China and Japan. Its other ecoraces are distributed across the Palearctic and Indo-Australian biogeographic regions. The primary food plant of this polyphagous insect is castor (*Ricinus communis* L.), but it also feeds on a wide range of food plants such as *Heteropanax fragrans*, *Manihot utilissima*, *Evodia flaxinifolia* and *Ailenthus gradulosa*. The wild *S. c. ricini* silkworm completes one to three generations per year depending on geographical position and climatic conditions of the region. Populations of *S. c. ricini*, that have been commercially exploited and are present in different regions of north-east India show wide morphological and quantitative variations in characters such as silk content, larval weight, cocoon weight, cocoon shell weight and silk ratio. Eri silkworms were successfully acclimatized in America and Europe, but could not take firm hold.

S. c. ricini shares basic characteristics of the saturniidae, but because of its different generic origin it has many characteristics which distinguish it from the species of *Antheraea*. The eggs are ovoid, candid white. The larva on hatching is greenish yellow. The body color changes gradually to pure yellow by the end of the third day. From the third instar onward the body color segregates into yellow, cream, green, blue or white. The fully mature larva is translucent and covered with a white powdery substance. Both spotted and unspotted larvae are found. The spots are of various types; single, double, zebra and semi-zebra. The obtect adectious pupa does not depart from the basic saturniidae pattern. The cocoon can be easily distinguished from those of *Antheraea* as they are elongated, soft wooly, peduncle-less, open mouthed and un-reelable. They exhibit color polymorphism, being brick red and creamy white. The length of the male moth is about 2.3 cm and female is 3 cm (Figure 2).

Construction and content

EST sequences of *A. assama* and *A. mylitta* were generated in CDFD, Hyderabad (Dr. Nagaraju's Lab) and ESTs from S. c. ricini were generated in University of Tokyo (Dr. Toru Shimada's Lab), by sequencing the cDNA clones amplified from mRNA isolated from several tissues at different developmental stages. All the protocols employed for RNA extraction, cDNA library preparation and EST processing pipeline are available online in the database and can be accessed from 'Protocols' section. The ESTs were further processed with Phred program (Ewing et al., 1998) for base calling DNA sequence traces. A cut off Phred score of 15 was assigned to extract quality sequences from chromatograms. In order to enhance the quality of sequences, ESTs were screened for presence of vector sequences and subsequently detected vector sequences were then removed using 'Cross Match' program. EST reads with length less than 100 bp were discarded. Majority of ESTs included in the database were having the lengths ranging between 400-600 bp.

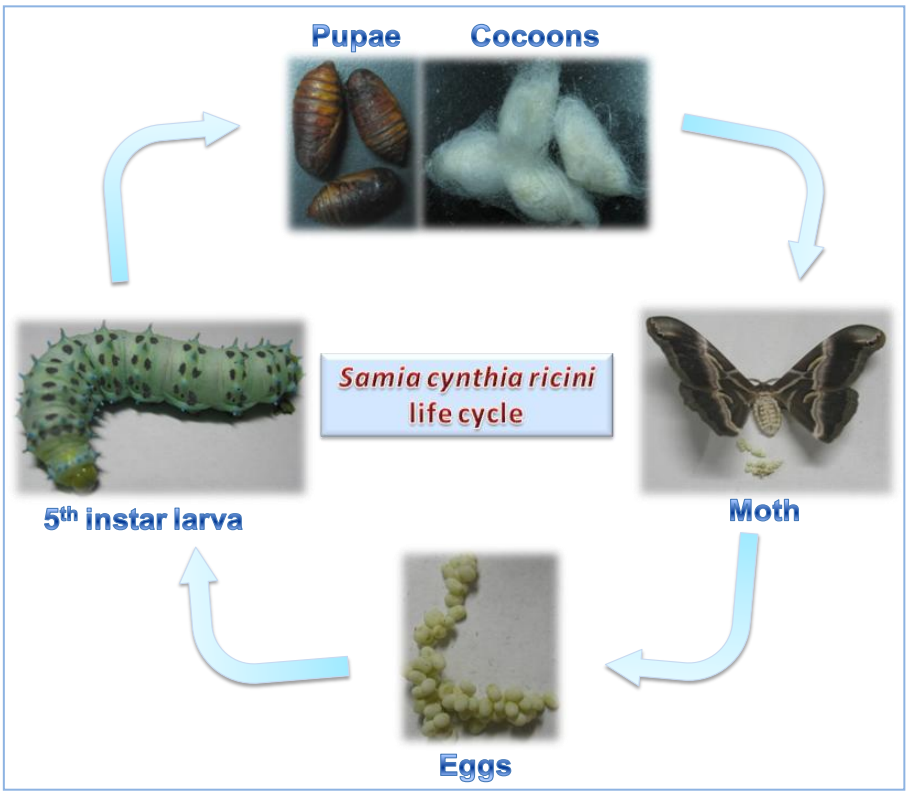


Figure 2: Life cycle of *S. c. ricini*

Table 1: Details of cDNA libraries, number of ESTs generated in each wild silkmoth species and results of EST analysis.

		Total no. of ESTs	No. of contigs	No. of singletons	No. of unigenes
Antheraea assama	**10 libraries**	**35,722**	**2,260**	**5,937**	**8,197**
Tissue type	***Developmental stage***				
Embryo	96 hours after oviposition	11,502			
Brain	Fifth instar	5,299			
Testis	Fifth instar	4,235			
Ovary	Fifth instar	3,891			
Fatbody	Fifth instar	2,318			
Midgut	Fifth instar	2,439			
Posterior silkgland	Fifth instar	2,543			
Middle silkgland	Fifth instar	1,031			
Epidermis	Fifth instar	1,386			
Compound eyes	Fifth instar	1,078			
Samia cynthia ricini	**3 libraries**	**19,979**	**1,593**	**3,528**	**5,121**
Tissue type	***Developmental stage***				
Embryo	96 hours after oviposition	6,647			
Fatbody I	12-24 hours after injection of *E. coli* into hemocoel of fifth instar	6,681			
Fatbody II	12-24 hours after injection of *Candida albicans* into hemocoel of fifth instar larvae	6,651			
Antheraea mylitta	**1 library**	**1,412**	**166**	**554**	**720**
Tissue type	***Developmental stage***				
Fatbody	24 hours after injection of *E. coli* into hemocoel of fifth instar	1,412			
	14 libraries	**57,113**	**4,019**	**10,019**	**14,038**

To produce non redundant EST dataset for further functional annotation and comparative analysis, 57,113 ESTs were clustered and assembled through TGICL package (Pertea et al., 2003) with the CAP3 (Huang and Madan, 1999) default options. Based on regions of similarity, EST sequences were merged into contigs. A total of 14,038 EST clusters consisting of 4,019 contigs and 10,019 singletons, putatively regarded as unigenes, was generated (Table 1). From these unigene sequences, poly-A tails were trimmed using TrimEST program of EMBOSS (Rice et al., 2000). Trimmed unigene sequences thus obtained were annotated for GO (Consortium, 2006). The GO annotation is based on the closest homologues identified by BLAST search against Seqdblite FASTA sequence flat file(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000)(Sezutsu and Yukuhiro, 2000). All the unigenes were assigned a biological process, molecular function and cellular component using GO database. ESTs are potential resources for SSR and SNP marker discovery and hence were screened for SSRs by using Tandem Repeats Finder (TRF) (Benson, 1999). For extraction of repeats, we assigned the following TRF parameters: match = 2, mismatch = 3, indel = 5, match probability = 0.8, indel probability = 0.1, minimum score = 25 and maximum period = 10. *A. mylitta* EST sequence dataset was further analyzed for potential SNPs in cDNA sequences (cSNPs) using SEAN SNP Prediction Program with default settings (Huntley et al., 2006). A total of 118 cSNPs was predicted in 1412 EST sequences. These predicted cSNPs, after experimental validation, will be useful for the analysis of genetic variation and population structure of *A. mylitta* populations.

Utility

WildSilkbase is designed to provide a platform for storage of sequence data and access to annotated information on ESTs of wild silkmoths. The database has been developed using MySQL relational database system, and its web interface has been generated using PHP scripts. It operates under an Apache web server on a Fedora Linux system. The data model for raw and processed data of WildSilkbase is shown in Figure 3.

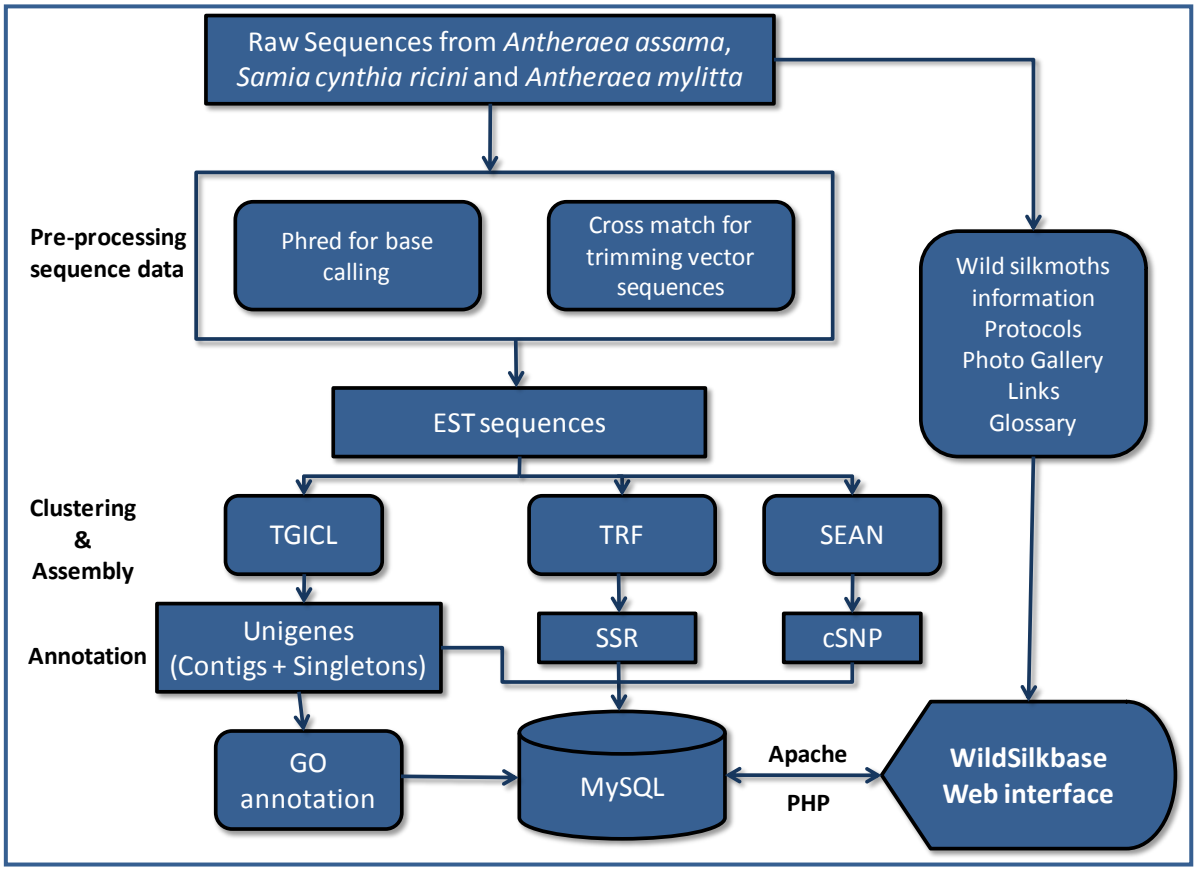


Figure 3: WildSilkbase organization and implementation.

Database can be inspected via four different HTML pages to allow distinctive queries. The 'Search Options' page provides three different options namely, Keyword Search, Homolog Finder and SSR Finder, for searching the database. The 'Keyword Search' gives access to search the database by GO terms, EST clone ID and Unigene ID. 'Homolog Finder' allows end user to search for the homologue of the query sequence against 6 different insect species (*Aedes aegypti*, *Anopheles gambiae*, *Apis mellifera*, *B. mori*, *Drosophila melanogaster* and *Tribolium castaneum*) based on BLAST search. 'SSR Finder' provides access to data on SSRs of wild silkmoth ESTs included in the database. Search results can be limited by the selection of SSRs of specific type belonging to

a particular species. The output generated is displayed in a tabulated format (Figure 4).

Figure 4: Web-shots of the database showing WildSilkbase query options, keyword search, homologue finder, SSR finder, GO viewer, BLAST search, cSNP data and download option.

To categorise transcripts by function, we utilized the GO classification. The 'GO Viewer' interface is designed to browse GO terminologies as a tree of terms. The number next to GO term represents the number of gene products annotated to that term which are included in the database and selected in the current view. BLAST (Altschul et al., 1990) search offered by WildSilkbase allows users to compare any query sequence against *A. assama*, *S. c. ricini* and *A. mylitta* ESTs and putative unigene sequence datasets. BLAST search results are returned directly to the user's web browser in HTML format (Figure 4). The sequence IDs on the BLAST result page are further linked to respective sequence information such as organism name, tissue of origin, sequence length, unigene ID and sequence. A link to ClustalW (Thompson et al., 1994) alignment file of the sequences matched in the databases is also provided on the result page.

The 'cSNP' web page provides direct access to cSNPs of *A. mylitta*. The results include information such as, contig ID, contig sequence length, the ESTs included in the contig, SNP location, alleles and consensus sequence.

The database also provides information on wild silkmoth biology, cDNA library construction and EST processing pipeline. The 'Picture Gallery' section has been incorporated in the database to give access to pictures of wild silkmoths. Links to several other databases and resources related to ESTs and insects are provided on 'Useful Links' webpage. A 'General Help' page is included for easy and efficient use of the database. The technical terms occurring in the database are hyperlinked to the 'Glossary' page for quick reference. In general, WildSilkbase allows the users to access all applications. The EST sequences of wild silkmoths are also deposited in NCBI and can be accessed at the NCBI EST sequence database, dbEST (accession numbers: *A. assama*; FE952359-FE963860 and FG203277-FG226965, *A. mylitta*; EB742119- EB743530, *S. c. ricini*; DC858270-DC878540).

Data Analysis

Gene Ontology annotation

GO annotation generates a dynamic controlled vocabulary that can be applied to all organisms, even while knowledge on genes and proteins in cells is still accumulating. The closest annotated homologue in the GO database was used for assigning these categories. Based on the GO annotation of the closely related homologues, the ESTs were assigned a molecular function, biological process and cellular component from the GO database (Ashburner et al., 2000). To accomplish this, the Seqdblite FASTA sequence flat file was downloaded from the GO database. By running BLAST against Seqdblite, the closest homologue was identified. From the BLAST output, molecular functions, biological process and cellular localization were parsed by building an in-house GO database in MySQL from the GO-term database flat file, downloaded from GO Database Downloads (Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Craig, 1997)(Grimaldi and Engel, 2005). The output is graphically represented (Figures 5 and 6).

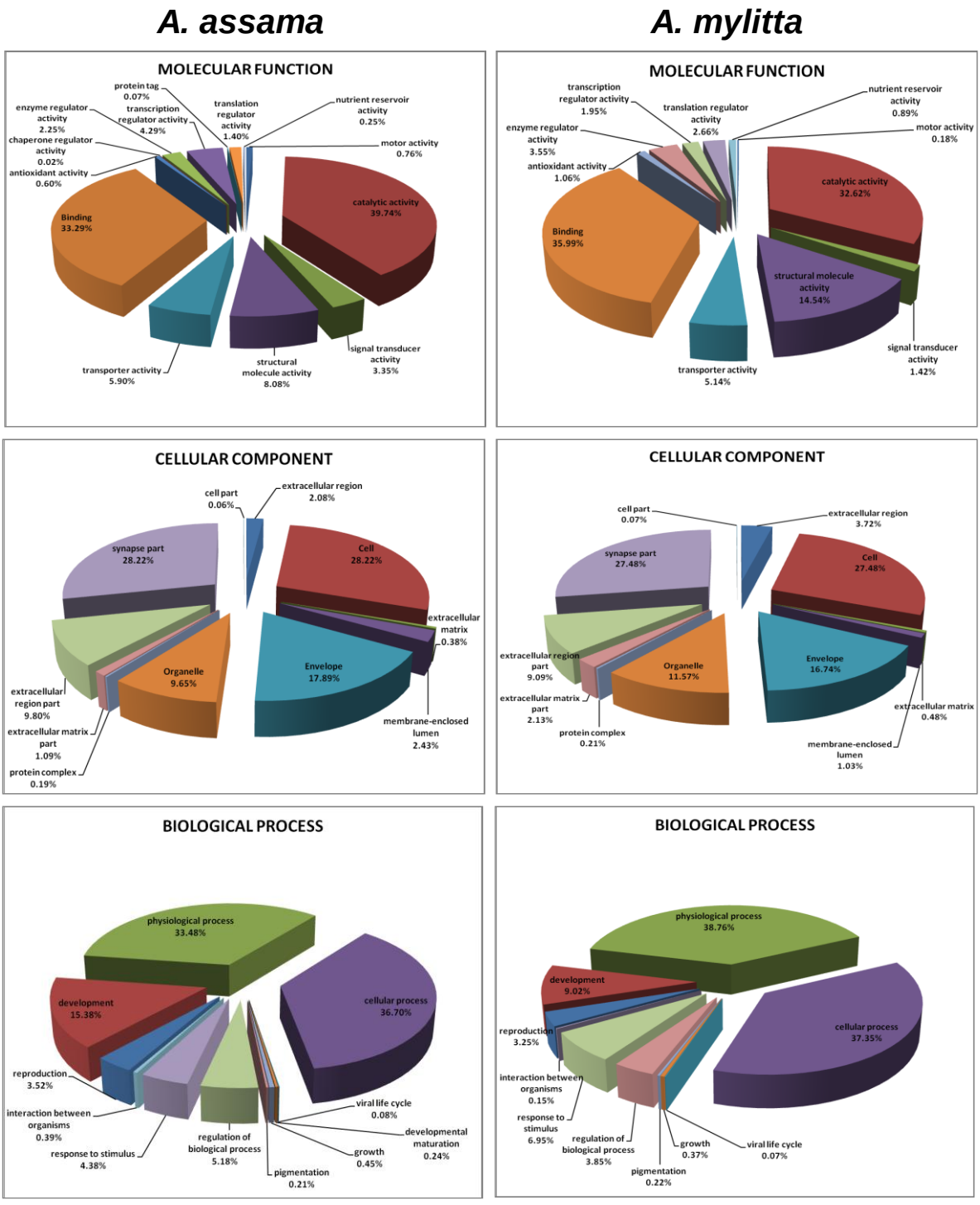


Figure 5: Gene Ontology representation of *A. assama* and *A. mylitta* clusters is shown for each organizing principle of GO: biological process, cellular component and molecular function. The chart is based on percentage representation of GO mappings.

S. c. ricini

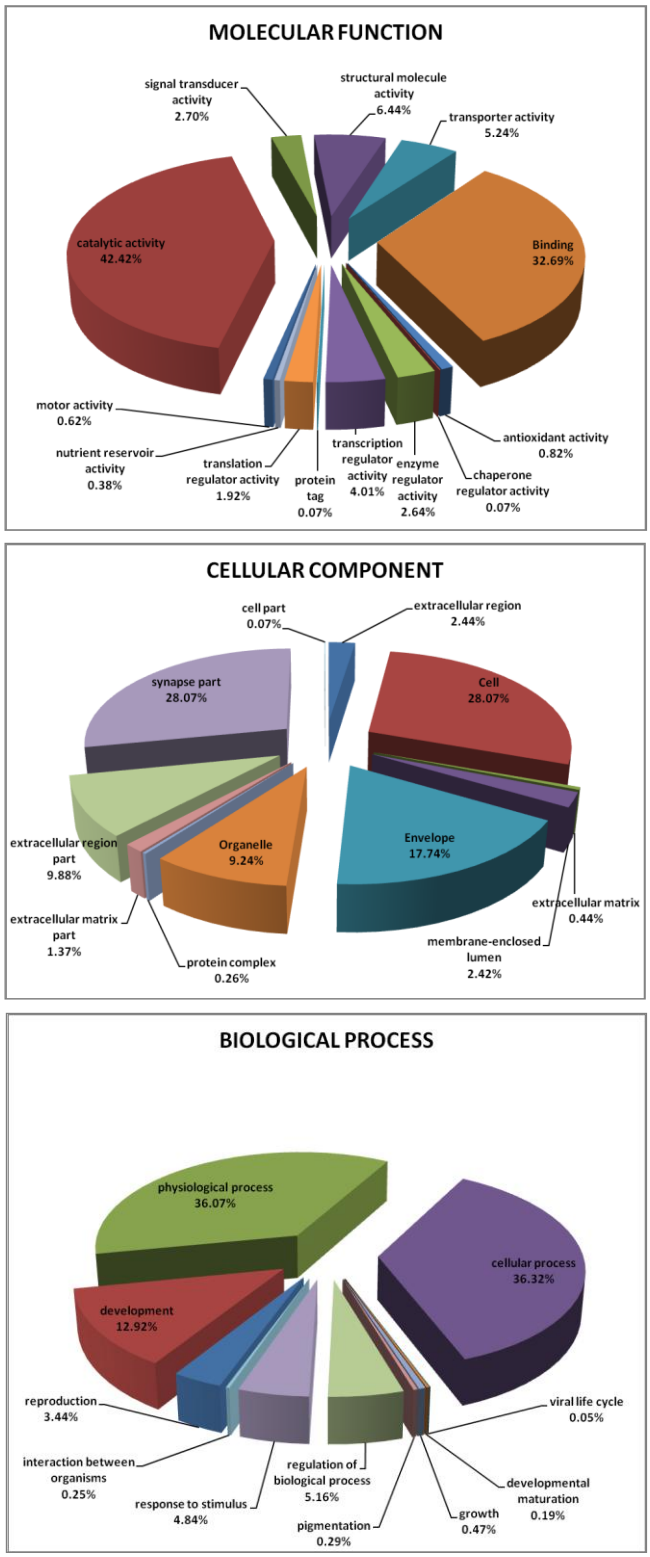


Figure 6: Percentage representation of gene ontology (GO) mappings for *S. c. ricini* clusters by biological process, cellular component, and molecular function.

GO mapping to molecular function revealed that a majority of genes from the tissue transcriptomes of wild silkmoths have almost equal distribution for 'binding' function (property of binding macro-molecules) and catalytic activity. The next abundant molecular function observed in the transcriptomes was structural molecule activity. In case of biological process, majority of the ESTs belonged to the category of physiological processes and cellular processes. Based on cellular localization results, most of the gene products were found to be localized in cell and synapse part (Figures 5 and 6). Putative unigenes for each category of GO, of any of the three wild silkmoth species can be browsed, viewed and downloaded from the 'GO Viewer' option of WildSilkbase.

Comparative genomics

Information on sequence similarity among genomes is a major resource for finding functional regions and for predicting their functions. Comparison of the genomes of closely related species is useful for finding the key sequence differences that may account for the differences in the organisms. Comparative genomics is thus a powerful and burgeoning discipline and has become more and more informative as genomic sequence data accumulate (Hardison, 2003). The lepidopteran insects have taxonomically specific biological phenomena including sex-determination, pheromone-dependent sexual communication, silk production, silk protein organisation, circadian rhythms, insect–plant interactions and insect-microbe interactions. Comparing genes of the wild silkmoths with other lepidopterans and other model insect species would shed light on conservation and divergence of different gene families. In the present study we compared the putative unigenes of the three wild silkmoth species between each other and with the unigenes of four insect species, *B. mori*, *D. melanogaster*, *A. mellifera* and *T. castaneum*.

The putative unigenes of each of the three wild silkmoth species were compared among each other using TBLASTX at the cutoff E-value of 1e-5. The resultant homologs between each pair and homologs common to all the three species and unique genes present only in one of the species are represented graphically (Figure 7). Gene comparisons, for example, revealed that *A. mylitta* has more number of homologous genes with *A. assama* than with *S. c. ricini*. This is in accordance with the phylogenetic relationship among silkmoth species (Mahendran et al., 2006), as both of them belong to the genera *Antheraea*.

A total of 9939, 17178, 9747 and 9028 unigene sequences of *B. mori*, *D. melanogaster*, *A. mellifera* and *T. castaneum* respectively, was downloaded from Unigene database of NCBI (Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec,

2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Sehnal and Zurovec, 2004)(Craig, 1997). Putative unigenes from each of the three wild silkmoth species were compared with these four insect species through TBLASTX search with E-value limit of 1e-5. Total number of homologs obtained was plotted on the graph. In all these comparisons *B. mori* was found to share the highest number of homologous genes with all the three wild silkmoth species and *D. melanogaster* standing next only to it, followed by *T. castaneum* and *A. mellifera* (Figure 8), which is expected as all the wild silkmoth species and *B. mori* share common ancestor (Arunkumar et al., 2006) and belong to the order, Lepidoptera.

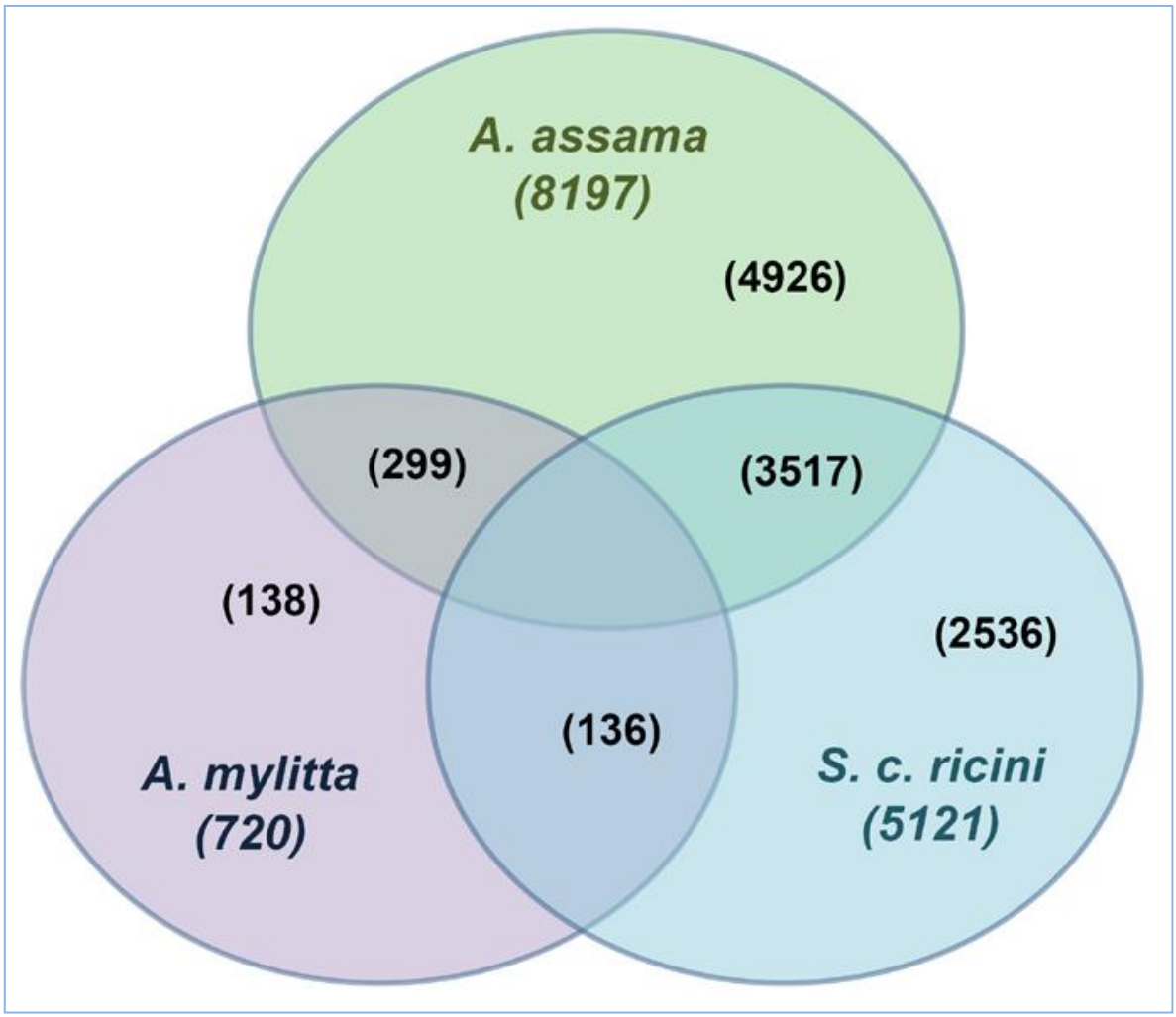


Figure 7: Venn diagram illustrating the number of unigenes shared among the 3 wild silkmoths, *A. assama*, *S. c. ricini* and *A. mylitta*.

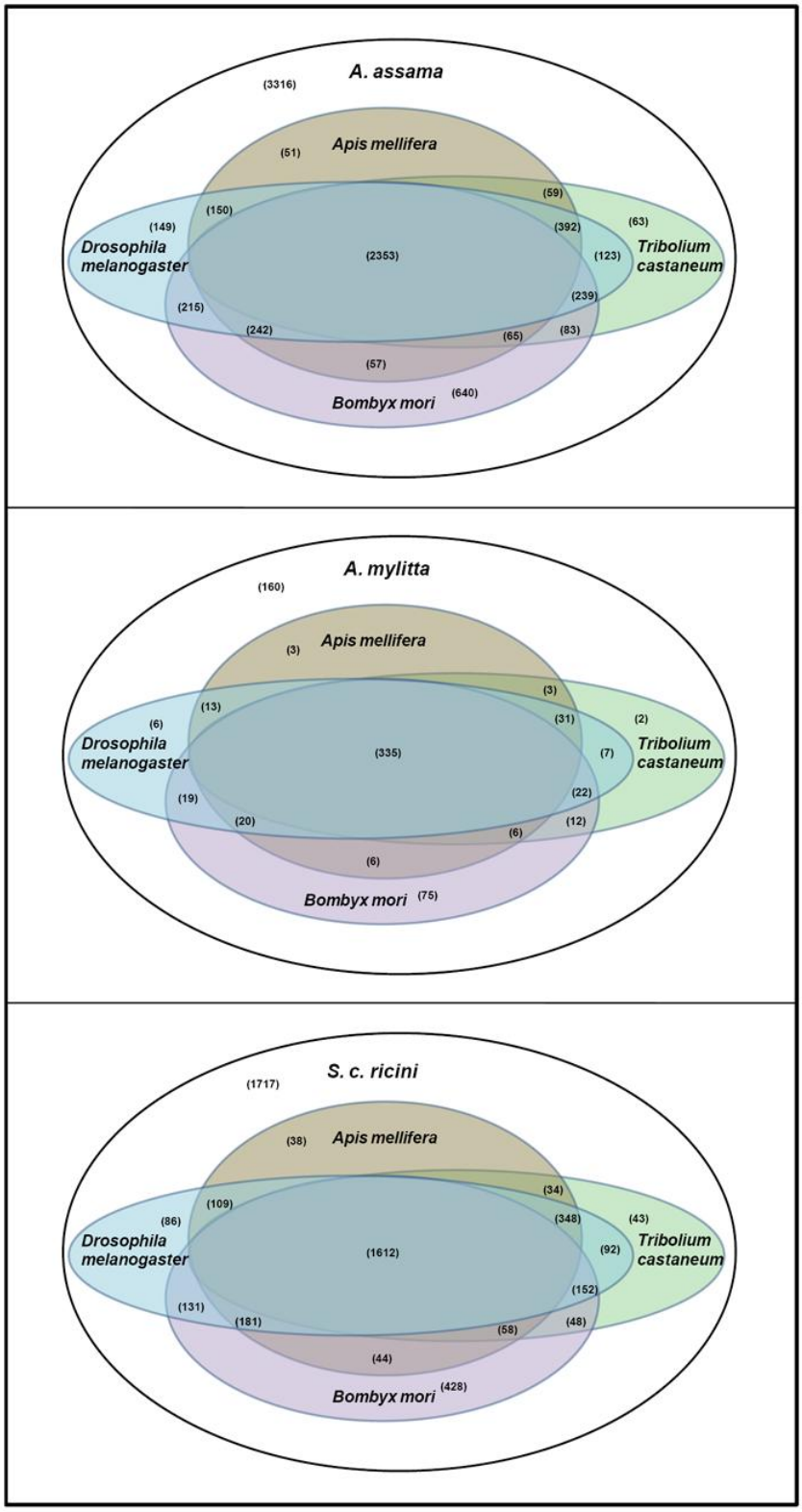


Figure 8: Venn diagrams showing the number of shared and species-specific genes among each of the 3 wild silkmoth species and 4 model insect species.

Conclusions

WildSilkbase aims to provide user-friendly access to EST data on wild silkmoths. Database will be continuously updated as and when the new information is available on wild silkmoth ESTs. Researchers working on silkmoths are encouraged to submit their wild silkmoth EST data to WildSilkbase, so that it can be made a single window portal for all information on wild silkmoths. WildSilkbase could be extremely useful for the researchers working in the areas of ecology, evolution, functional and comparative genomics, genetics and biochemistry of insects.

Chapter III

Bombyx mori testis transcriptome analysis and physical mapping of testis specific genes

Male germ cells in animals differentiate into sperm cells with extensive morphological and physiological changes. This complex and highly regulated process is orchestrated by an organized network of thousands of genes encoding proteins that play essential role during specific phases of germ cell development. Presumably each stage requires that a complement of genes is activated and others repressed to satisfy developmental demands.

A series of testicular events in the silkworm, *Bombyx mori* seems to be under the control of a large number of tissue- and stage- specific genes like those in many other organisms. In previous studies, a few testis-specific genes involved in silkworm spermatogenesis, such as the genes coding for BmAHA1 (Miyagawa et al., 2005), BmDmc1 (Kusakabe et al., 2001) and testis-specific tektin (Ota et al., 2002) have been reported. However, no comprehensive data on testis transcriptome at global level is available for *B. mori*. Among insects, only in *Drosophila melanogaster*, testis transcriptome has been studied in detail (Andrews et al., 2000, Mikhaylova et al., 2008, Parisi et al., 2003, Ranz et al., 2003).

Recent high throughput genomics projects have focused on the construction, annotation and analysis of cell and tissue specific transcriptomes that have provided fundamental insights into biological processes. Data gathered from Expressed Sequence Tags (ESTs) and microarray based gene expression profiles provide important repertoire for discovery of novel genes that are expressed in tissue-specific manner. In this context, we analyzed the *B. mori* testis transcriptome to identify testis specific genes and their functions. Further, we annotated the testis-specific ESTs by using protein homology data and Gene Ontology (GO), and deduced gene structures of a few testis specific genes using Whole Genome Shotgun (WGS) sequence data and ESTs. These studies led to the identification of a testis specific splice form of *intersex* (*ix*) mRNA which is reported to be transcribed from a single exon gene in *D. melanogaster* (Garrett-Engele et al., 2002). RT-PCR and RNase protection assays confirmed the two splice forms; one, accumulated exclusively in testis and the other being present ubiquitously in all the tissues including testis.

Another scientifically rewarding exercise of analyzing testis transcriptome is to examine whether there is any correlation between the sex specificity of the genes and their chromosomal location. In every eukaryote, unequal complement of sex determining chromosomes exists between the two sexes. Unlike mammals and dipterans, which have XX/XY system of sex chromosome composition, lepidopterans have ZZ/ZW or ZZ/ZO system of sex chromosome composition. Two hypotheses that are in vogue propose contradictory fate for the genes that reside on the sex chromosomes and are preferentially/exclusively expressed in one sex. Rice's hypothesis (Rice, 1984) proposes that genes with sex-biased / sex-specific expression are enriched

on X chromosome, which has been tested in Human (Saifi and Chandra, 1999, Vallender and Lahn, 2004) and Mouse (Wang et al., 2001). An alternate hypothesis suggests feminized X because the X chromosome is present in females 50% more frequently than in males, providing evolution with more opportunity to act on genes benefiting females (Reinke et al., 2000). Thus, genes with female biased expression should reside on the X chromosome. So which hypothesis is correct? Or are the two hypotheses mutually compatible? Corroborating the Rice hypothesis, the studies by Wang et al. (2001) on mouse male germ cells and by Saifi and Chandra (1999) and Lercher et al. (2003) on human X chromosome concluded that genes expressed specifically in males were X-linked far in excess of the autosomes. In contrast, the male-biased genes in *Drosophila* (Parisi et al., 2003, Ranz et al., 2003) and genes expressed in spermatogenic and oogenic cells in *Caenorhabditis elegans* (Reinke et al., 2004) are under-represented on the X chromosome. Khil et al. (2004), studying on sex-biased expression in mouse show that genes involed in spermatogenesis are relatively underrepresented on the X chromosome and female biased genes are enriched on it. Further, they show that the meiotic sex chromosome inactivation (MSCI), in which sex chromosome becomes heterochromatic and transcriptionally inactive, accounts for the depletion of testis-expressed genes on the X chromosome. This was so inferred for the reason that the genes expressed before MSCI are over-represented on the X chromosome. To test this phenomenon in a lepidopteran model system, *B. mori*, we analyzed the distribution of testis specific genes on chromosomes. A total of 1104 testis specific genes, identified by microarray analysis (Xia et al., 2007), was mapped onto their specific location on *B. mori* chromosomes. Our results show that Z chromosome (Linkage Group 1) harbors significantly higher number of testis specific genes as compared to autosomes. We speculate that lack of dosage compensation and sexual antagonism have possibly led to the accumulation of male specific genes on Z chromosome. In the course of evolution, proteins which are required in higher amounts in males, than in females would have been favored on Z chromosome. Finally, initial analysis of testis specific paralogs gave evidence for possible translocation of male advantageous genes onto Z chromosome.

Sequence source

More than 100,000 ESTs, are available for *B. mori* in NCBI dbEST (Mita et al., 2003, Xia et al., 2004). We downloaded 9614 testis ESTs and a total of 95,051 ESTs derived from tissues other than testis. ESTs generated from other tissues were also downloaded to identify testis specific genes.

Sequence analysis

Since many ESTs may be derived from same gene, the sequences were assembled into clusters with the TGICL program (Pertea et al., 2003). A cluster is defined as a unique set of sequences, which share common sequence similarity. A cluster containing only one sequence is termed a singleton.

Based on the GO annotation of the closely related homologs, ESTs were assigned with their molecular functions, biological process and cellular component from the GO database (Ashburner et al., 2000).

To identify putative homologies to known proteins, the clusters (3,622) were subjected to BLASTx searches against the *nr* of NCBI where a cutoff E-value of 1e-05 was used for parsing output files (Altschul et al., 1990). In our annotation, an E-value of 1e-20 was used as an upper limit to assign significant homology. In addition, in cases where no significant homology was found, an E-value limit of 1e-05 was used to assign weak homology. We found this additional category of weak homology useful for data mining as many clusters do not represent full-length sequences. Hence it is possible that only a highly divergent region of a gene sequence is available in our collection. The category of weak homology allowed us to find the potential homologs in such situations.

Digital differential display

We carried out *in silico* differential display to identify genes which are expressing specifically in testis. This was accomplished by carrying out MegaBLAST with cut-off: score $\geq$50, percent similarity $\geq$94 and E-value $\leq$1, against non-redundant EST set obtained by clustering and assembly of *B. mori* ESTs from tissues other than testis and using non-redundant testis ESTs as query. Testis ESTs that did not show any similarity with other tissue ESTs were considered as putative testis specific genes.

Experimental validation of a few evolutionarily conserved testis specific genes

To confirm the testis specificity of genes as predicted by digital differential display, semi-quantitative RT-PCR was carried out for a select set of 15 transcripts using total RNA from six different tissues, midgut, fatbody, head, silk gland, epidermis and gonads of 5th instar silkworm larvae as template. The sex-limited strains of silkworm carrying translocation of chromosome 2 harboring gene for larval markings to W-chromosome were utilized for this purpose. The males and females of this silkworm stock can be distinguished from 3rd instar onwards by virtue of markings on thoracic segment only in females. Total RNA was isolated separately from all the tissues using Trizol reagent (Invitrogen, Carlsbad, CA, U.S.A.), followed by DNase (Invitrogen) treatment to remove genomic DNA.

PCR was carried out using thermal cycler (Eppendorf, Germany) under the following conditions- 94°C, 2 minutes- initial denaturation, 30 cycles (94 °C - 30 seconds, 58 °C- 30 seconds, 72°C- 2 minutes) and a final elongation at 72°C for 10 minutes. Actin cDNA was amplified as an endogenous control. PCR reaction components included: 1X buffer, 100µM dNTPs, 1.5 mM $MgCl_2$, 0.5 unit Taq polymerase (MBI), 0.5 µM primers.

Expression analysis of B. mori intersex (Bmix) gene

The *in silico* analysis of the testis transcriptome led to the identification of two splice forms of *Bmix* gene. Complementary DNA was synthesized from total RNA isolated from midgut, fatbody, head, silk gland, epidermis and gonads of 5th instar silkworm larvae, by oligo(dT) priming using MMLV reverse transcriptase (Invitrogen). To determine the tissue distribution of *Bmix* expression, PCR experiments were performed using cDNAs from multiple tissues of males and females. *Bmix* gene-specific primers were designed using primer 3 software (Rozen and Skaletsky, 2000). Five primers designed for all but second exon are shown in Figure 3. A pair of primers that bind to first and second exons was also designed to specifically amplify the second splice form. PCR was performed for 35 cycles of 94^{o}C for 30 seconds, 60^{o}C for 30 seconds, and 72^{o}C for 1 minute. The primers for β-actin, actin-F CACTGAGGCTCCCCTGAAC and actin-R GGAGTGCGTATCCCTCGTAG, were used as endogenous control. PCR products amplified using *Bmix* gene specific primers were cloned into pCRII- TOPO vector (Invitrogen) and then sequenced to confirm the testis specific splice form.

RNase protection assays were carried out to study the expression pattern of *Bmix* splice forms. A radiolabelled anti-sense RNA probe of 284 bp was prepared for region comprising first

and second exons using MegaScript *in vitro* transcription kit (Ambion, Texas, USA) after cloning the fragment into pCRII- TOPO vector (Invitrogen). This probe contains 89bp sequence complementary to second exon and the remaining 195 bp complementary to part of first exon. RNA samples (10 μg each) and aliquots containing 2 x 10^5 cpm of the RNA probe were hybridized overnight at 45°C in 20 μl hybridization solution containing 75% formamide/0.5 M NaCl/10 mM Tris HCl, pH 7.5. After addition of 300 μl of 300 μM NaCl/5 mM EDTA containing RNase A at 60 ug/ml, the mixture was incubated for one hour at 37°C and subsequently treated with proteinase K (300 μg/ml) for 30 minutes at 37°C, in the presence of 0.1% SDS, extracted twice with 1:1 phenol/chloroform and once with chloroform, precipitated with ethanol, and subjected to electrophoresis in a 8% polyacrylamide/8 M urea gel in 90 mM Tris borate buffer (pH 8.3). RNase protected fragments were quantified with a PhosphorImager.

Sequencing and phylogenetic analysis of intersex homologs in other Lepidoptera

Putative homologs of *ix* in lepidopteran insects were identified either by searching the EST databases or by sequencing with primers designed to conserved regions. DNA sequences found to encode proteins with significant similarity to BMIX amino acid sequence were identified by translated BLAST (tblastn) search against WildSilkbase, an EST database of wild silkmoths (Arunkumar et al., 2008). The search resulted in 6 ESTs from *Antheraea assama* (Ac. Nos. Aaov1642, Aaov0427, Aaov0387, Aaov1877 and Aaov2957 from ovary ESTs, and Aabr4525 from brain ESTs) and 1 EST from *Samia cynthia ricini* (Ac. No. Sc_96Hrs4747). These ESTs were individually inspected and assembled to get partial sequences of *ix* homologs in these species.

To identify putative *ix* homolog in *Antheraea mylitta*, we performed reverse transcription of total RNA from *A. mylitta*, followed by PCR amplification of cDNA using primers designed to regions conserved between *A. assama* and *S. c. ricini ix* gene sequences. PCR amplicons were then sequenced after cloning into TA vector. Based on the partial sequence, RACE (Rapid Amplification of cDNA Ends) primers were designed and 5' end sequence was obtained. All the three putative *ix* cDNA sequences obtained had complete 5' end sequences. These were conceptually translated and aligned with other homologs from several taxa.

The amino-acid sequences of IX homologs available in other organisms (*Rattus norvegicus*, XP_214868; *Mus musculus*, NP_080318; *Homo sapiens*, AAH19015; *H. sapiens*, AAU43732; *Canis familiaris*, XP_855358; *Bos taurus*, XP_871568; *Xenopus laevis*, AAH78525; *Tetraodon nigroviridis*, CAF9084; *Danio rerio*, AAM34654; *Strongylocentrotus purpuratus*, XM_001178265; *Apis mellifera*,

XP_395989; *Tribolium castaneum*, XP_970547; *Anopheles gambiae*, XP_321918; *D. melanogaster*; *Drosophila pseudoobscura*, EAL26072; *Drosophila virilis*, AAV65894; *Megaselia scalaris*, AAV65895) were retrieved from NCBI protein database. The protein sequences of newly identified homologs were aligned with BMIX and other IX homologs downloaded from NCBI. Phylogenetic analysis was carried out using ClustalX and an unrooted tree was constructed.

To examine whether there are any alternative splice forms present in *A. assama*, we performed RT-PCR of total RNA isolated from whole larvae. The agarose gel profile revealed the presence of 2 bands hinting at presence of a second splice form in *A. assama* also. Therefore the amplicons were cloned into TA vector and sequenced to identify the exon(s) leading to second splice form.

Physical mapping of testis specific genes and analysis of testis specific gene paralogs

Recently, through microarray analysis, 1104 genes were identified to be testis specific in *B. mori*. Generally, the tissue-specific gene expression features have been viewed traditionally as predictors of tissue-specific function. If the intensity of expression of a gene in one surveyed tissue exceeded twice that in other tissues then it is regarded as tissue-specifically expressed gene (Xia et al., 2007).

These microarray validated testis specific genes were assigned chromosomal position on *B. mori* genome, to examine their distribution on different chromosomes. Mapping was carried out using the *B. mori* physical chromosome map as implemented in KAIKOBLAST and UTGB (University of Tokyo Genome Browser). In brief, all the gene sequences were queried against physical map database using BLAST program. The results were then manually parsed to determine the exact location of genes. Total number of genes on each chromosome and number of genes per Mb of each chromosome were calculated using the map data. We also mapped 465 genes from the other tissues (only from somatic tissues as identified in Xia et al 2007, through microarray analysis) onto chromosomes to compare their distribution with that of testis specific genes.

From the testis specific genes identified through microarray analysis, paralogs were selected by combining BLAST parameters and homology to published annotated genes. Translated BLAST (tBLASTx) was performed for each of the testis specific genes against complete set of testis specific genes. The BLAST result was parsed so as to get information on hits having score >100. The parsed data was manually checked to group the paralogs and were assigned chromosomal locations. Each group was then analyzed by carrying out BLAST against NCBI protein nr database.

Only those groups which had same functional annotation for all the genes included in the group, were regarded as genuine paralogs.

Further, through *in silico* subtraction a set of 2559 testis specific genes was identified, using full length cDNAs (fl-cDNAs) (Mita et al., unpublished data) and ESTs derived from testis. From these, genes that were common to microarray validated testis specific genes were removed and the remaining 1857 genes were mapped onto *B. mori* chromosomes to assign their chromosomal distribution.

Spermatozoa are produced through spermiogenesis, a complex process involving cell proliferation, cell differentiation, meiosis, and powerful selective pressures on male reproductive success. During spermatogenesis, transcriptional regulation within germ cells is cautiously orchestrated (Sassone-Corsi, 2002). In the present study 9614 ESTs derived from fifth instar larval testis were analyzed to identify such genes in *B. mori*. *In silico* analysis revealed that several testis specific genes are evolutionarily conserved from insects to mammals. There are many reports on genes expressed in a specific and restricted pattern in the testis. The computational analyses carried out in the present study supports the idea that the testis expresses a complex set of transcripts as observed in *Drosophila* (Andrews et al., 2000). The present study also identified several families of testis specific genes like *tektins*, *dyneins*, *kinases* and *tubulins*, involved in spermatogenesis of silkworm.

High gene diversity in testis transcriptome

The accessibility of 9X coverage of the *B. mori* genome sequence and EST resources for this insect facilitated the identification of several testis-specific genes. Clustering and assembly of 9,614 testis ESTs resulted in a total of 3,622 unique clusters, containing 1,112 contigs and 2,510 singletons. We obtained 24,857 unique clusters (8,819 contigs and 16,038 singletons) by clustering and assembly of the other tissue ESTs.

To identify putative homologies to known proteins, we subjected the clusters to BLASTx searches against the non-redundant (*nr*) database of NCBI. A total of 2,385 (66%) unique sequences shared homology with known proteins from *nr* and could be assigned a putative identity. Of the 3,622 clusters, 2.6% matched proteins with an E-value of >1e-99 and were considered to be genuine orthologs. Thirty percent of the clusters found a hit with an E-value between 1e-20 and 1e-99 and were assigned significant homology. Finally, 16% of clusters had a first hit with an E-value between 1e-19 and 1e-05 and were assigned weak homology to a protein from the *nr* database. A majority of the sequences (52%) showed either no significant similarity (34%) or were having homolog (18%) in the *nr* with an E-value $\geq$1e-04.

A large proportion of the transcripts showed no significant match to known genes. This points to the presence of a multitude of novel genes expressed exclusively in testis. Since many testis ESTs do not have a homolog in NCBI protein database and many genes were found to be expressed only in testis, there seems to be high gene complexity in testis transcriptome. This is consistent with the observations in EST studies of mouse, wherein, it was estimated that more

than 2300 genes are specifically expressed in meiotic and post-meiotic male germ cells (Schultz et al., 2003). Many of these genes encode proteins that are functionally unique to the developing germ cells, or protein isoforms, which can take over functions of proteins encoded by genes that are silenced during spermatogenesis (Baarends and Grootegoed, 1999, Eddy, 2002, Wang, 2004). Analyses of ESTs and microarrays indicated that an unusually diverse set of mRNAs is expressed in mouse, human, and *Drosophila* testes with large number of these mRNAs being expressed in spermatogenic cells (Andrews et al., 2000, Kerr et al., 1994, Pawlak et al., 1995).

In silico differential display of B. mori testis ESTs

Several interesting features about the uniqueness of testis transcriptome were revealed by *in silico* subtraction of testis derived non-redundant transcripts from those of other tissues. Out of 3,622 unique clusters of testis origin >900 were found to express only in testis. These were regarded as putative testis specific genes. Similar results were also reported by Xia and co-workers (Xia et al., 2007) wherein 1104 genes were found to be testis specific using microarray analysis.

Testis specificity of several genes is conserved

In *B. mori*, using the EST database, we identified four *β-tubulin* and three □- *tubulin* genes which could be classified into at least three distinct subfamilies: ubiquitously expressed, developmentally regulated and testis specific (Kawasaki et al., 2003). In the present study we identified another hitherto unreported *α-tubulin* gene named *bmtua4* (Ac. No., Bmo.6186), which has single exon, similar to the three previously reported genes. The *bmtua4* showed ~76% similarity to *bmtua1* and *bmtua2* and ~71% similarity to *bmtua3* genes. Earlier reports have identified testis specific *β-tubulin* (*bmtub4*) (Mita et al., 1995) and *α-tubulin* (*bmtua3*) (Kawasaki et al., 2003) genes; whereas *bmtua4* showed male specificity with enhanced expression in testis.

Spermatogenesis appears to be distinctive because many gene families include paralogs that are expressed in somatic cells as well, and paralogs that are expressed solely in spermatogenic cells (Eddy and O'Brien, 1998) as in case of *tubulin*, *dynein light chain* genes reported in the present study; whereas gene families such as tektins have conserved testis specificity.

Gene structure deduction of 10 testis specific genes revealed a few interesting characteristics of organisation of testis specific genes. Two genes were found to show more than

one alternative splice forms in testes, which was confirmed through RT-PCR analysis. The unusual feature of spermatogenesis is that many genes that are expressed in both spermatogenic and somatic cells produce transcripts that differ in structure due to alternative promoters, alternative splicing, and upstream polyadenylation sites (Kleene, 2001, Venables, 2002, Walker et al., 1999). The presence of multiple splice forms of many genes within same tissue would increase the diversity of the proteome (Graveley, 2001).

Determination of testis specificity

Through BLAST analysis we discovered 34 testis specific genes, which are known to have specific and specialized function in testis, and are conserved from insects to mammals. All these genes have testis specific homolog in diverse species like mammals (*H. sapiens*, *M. musculus*, *R. norvegicus*, *C. familiaris*, *B. taurus*), urchin (*S. purpuratus*), ascidian (*Ciona intestinalis*), avian (*Gallus gallus*), fishes (*T. nigroviridis*, *D. rerio*) and amphibian (*X. laevis*). Among them, 22 were present exclusively in testis ESTs (testis specific expression), 8 showed elevated expression in testis (testis enhanced expression) and 4 did not show any tissue specificity of expression (equal expression). We found three genes coding for dynein light chain protein, of which two were present only in testis transcriptome. Many genes involved in different processes of spermatogenesis were identified that are evolutionary conserved across phyla (Table 1).

To confirm the testis specificity of evolutionarily conserved genes, RT-PCR expression analysis of selected 15 genes was carried out. Of these, 9 (including 717bp testis specific splice product of sperm mitochondria associated protein) were expressing exclusively in testis and the remaining six genes showed male specificity with testis enhanced expression. Only one gene, a homolog of *TEGT* (*Testis Enhanced Gene Transcript*), showed equal expression in all the tissues examined (Figure 1). Expression analysis is in accordance with the tissue distribution of ESTs as revealed by *in silico* analysis. Conservation of testis specific genes from insects to mammals suggests that the proteins encoded by these genes are probably indispensable for spermatogenesis. Expression analysis also revealed that most of the testis-enhanced genes were male specific.

Expression analysis confirmed the presence of predicted alternative splice forms of *Sperm mitochondria associated protein* and *Serine protease* genes. The *Sperm mitochondria associated protein* gene showed three transcripts with amplicons of 800, 717 and 417 bp, whereas *Serine protease* gene showed two products of 575 and 331 bp (Figure 1).

Table 1: Testis specific gene homologues identified by BLAST analyses of *B. mori* testis ESTs that are conserved in different animal phyla.

Unigene / dbEST ID	Gene description	Tissue specificity
Bmo.521	*Bm-Meichroacidin*	Testis specific
CK537614	*Dynein light chain 3* (*BmDLC3a*)	Testis specific
Bmo.2773	*Dynein light chain*	Testis specific
Bmo.1163	*Tubulin beta chain* (*bmtub4*)	Testis specific
Bmo.525	*Tubulin alpha chain* (*bmtuα3*)	Testis specific
Bmo.764	*Bmtektin-1* (BA3388)	Testis specific
CK537544	*Bmtektin-2*	Testis specific
CK536102	*Serine protease-6*	Testis specific
CK537191	*Serine kinase 1*	Testis specific
Bmo.6365	Sperm mitochondria-associated cysteine-rich protein	Testis specific
CK536677	*Testis protein*	Testis specific
Bmo.110	*BLu protein*	Testis specific
Bmo.709	*ATPase inhibitor (BAB39164)*	Testis specific
Bmo.4264	*Male germ cell-associated kinase*	Testis specific
Bmo.514	*Mage-d1*	Testis specific
CK534071	*Bm-tesmin*	Testis specific
CK533275	*Bm TPX1*	Testis specific
Bmo.4264	*Serine/threonine protein kinase MAK*	Testis specific
CK533681	*Structural sperm protein*	Testis specific
CK533950	*Male sterility protein 2-like protein*	Testis specific
CK537063	*Sperm ion channel*	Testis specific
CK536763	*Sperm associated antigen 6*	Testis specific
CK533294	*Sperm associated antigen 9*	Testis enhanced
Bmo.6386	*Outer dense fiber of sperm tails 2*	Testis enhanced
Bmo.2817	*Outer dense fiber of sperm tails 3*	Testis enhanced
Bmo.2777	*Sperm nuclear basic protein*	Testis enhanced
CK537618	*Channel, sperm associated 4*	Testis enhanced
Bmo.956	*Dynein light chain 3* (*BmDLC3*)	Testis enhanced
Bmo.6404	*Tubulin alpha chain* (*bmtuα4*)	Testis enhanced
Bmo.6412	*Testis intracellular mediator protein*	Testis enhanced
Bmo.2293	*Bm-tegt*	Equal expression
BP127933	*Testis specific 10*	Equal expression
Bmo.5282	*Outer dense fiber of sperm tails 1*	Equal expression
Bmo.6298	*Motile sperm domain containing 3*	Equal expression

Note: Many clusters did not have a corresponding unigene ID. Such clusters are represented here with the dbEST ID of first EST in the cluster.

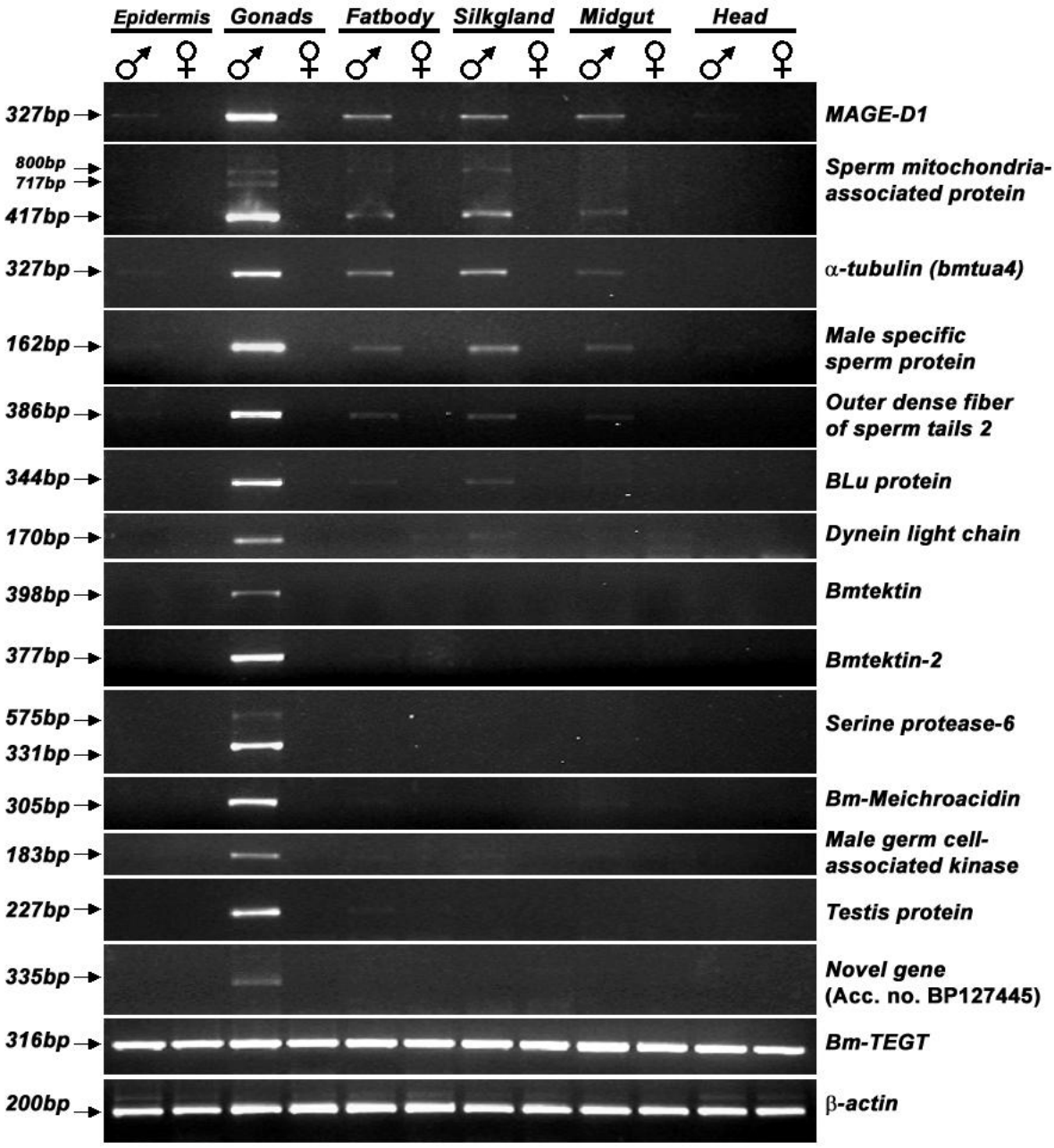


Figure 1: RT-PCR expression analysis of 15 predicted testis specific genes. β-actin was used as internal control. Details of accession numbers and tissue specificity as predicted *in silico*, are given in Table 1.

We found one testis specific transcript (Acc. no. BP127445), having no apparent similarity to any protein from nr of NCBI (Table 1) and was confirmed to be testis specific through RT-PCR (Figure 1). A few genes like TEGT have lost their testis specificity in insects whereas a few like the above mentioned one (BP127445), have gained testis specificity in silkworm.

Tektin gene is duplicated in B. mori

In the present study, we identified two *tektin* genes, one corresponding to the already reported *B. mori tektin* (*Bmtektin*) (Acc. no. BAB33388) and the second copy (hereafter called *Bmtektin-2*) being probably a product of a duplication event. In the BMTEKTIN-2, 220 amino acids at C-terminal end were deleted. The two tektins shared about 50% aa similarity. A BLAST search against *B. mori* WGS revealed that these two genes are single exon transcripts, located adjacent to each other in the genome about 4 kb apart (Figure 2). These genes were mapped on to a WGS contig with accession number AADK01000430 (36.8 kb). Expression analysis *in silico* with the available *B. mori* ESTs in dbEST, showed that both are expressed exclusively in testes (Table 1). Three ESTs similar to *Bmtektin and* two ESTs similar to *Bmtektin-2,* were present in testis ESTs.

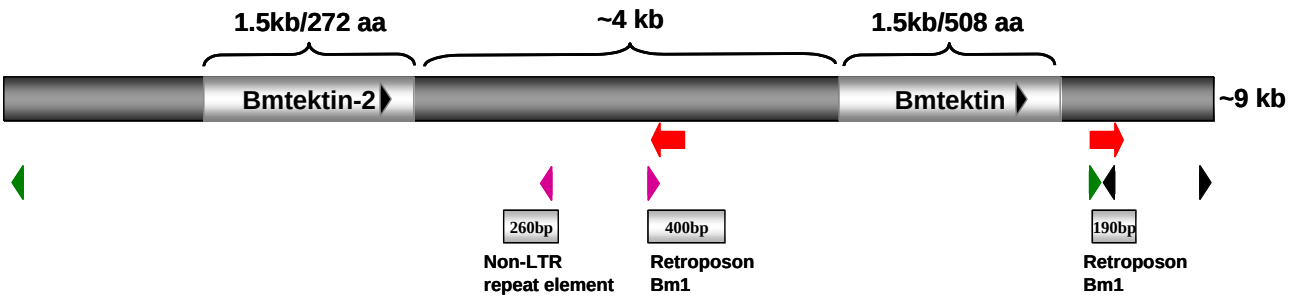


Figure 2: Tektin genes of *B. mori*, *Bmtektin* and *Bmtektin-2*. *Bmtektin-2* has originated as a result of duplication of *Bmtektin*. Arrows (red, black, violet and green) indicate the location and direction of inverted repeats, which are assumed to have brought about duplication, mediated by transposons. Arrows within genes represent the direction of transcription of two genes. Only ~550 bp of 5' end of the gene is duplicated leading to a truncated protein of 272 aa.

Further examination of the genomic contig harboring *Bmtektin* genes revealed the presence of four inverted repeats bordering the two genes (Figure 2). Around 200bp (red arrows) inverted repeats were present adjacent to *Bmtektin*, surrounded by *Bm1* retroposon (~400bp). A part of the repeat region of gypsy-Ty3-like retrotransposon was also found at 1kb upstream of *Bm1*, near to *Bmtektin-2*. A 130bp inverted repeat unit was present after the *Bmtektin* (black arrow in figure 2), and ~50bp inverted repeats (violet and green arrows in figure 2) were scattered around both the genes. Previous studies have indicated that transposition can occur in germ cells or in early embryogenesis, before the germ line becomes a distinct lineage (Ostertag et al., 2002, Prak et al., 2003). This appears to be the case in *B. mori* testis also, as many transposable elements are found expressed in the testis. Previous analyses have shown that transpositions occur in the expressing genes (Gilbert et al., 2002, Moran et al., 1999, Symer et al., 2002). Empirical evidence from a recent study shows that transposons active in a particular tissue, can insert into the genes expressing in that tissue (Muotri et al., 2005). Since *B. mori tektins* are expressed exclusively in testis and transposable elements are highly expressed in the testis, a transposition event might have led to the duplication of this gene.

The gene duplication provides material for functional diversification of genes. A duplicated gene may subsequently acquire new function or may retain some of its progenitor gene's functional repertoire, or a combination of both (He and Zhang, 2005, Lynch and Katju, 2004, Rastogi and Liberles, 2005). Earlier studies on Tektin family proteins have revealed their involvement in the formation and reinforcement of ciliary and flagellar microtubules during spermatogenesis (Larsson et al., 2000). Expression analysis in a previous study (Ota et al., 2002) has shown the testis specific expression of *tektin* during sperm maturation in *B. mori*. The protein was immunologically detected exclusively in the fraction expected to contain the 9+2 flagellar axonemes of sperms. It was thus inferred that the *tektin* is possibly involved in the spermatogenesis of *B. mori*. Significant divergence between the two paralogs shows that the duplication has taken place long back. RT-PCR based expression analysis showed that testis specificity of *Bmtektin* gene has been retained even after the duplication event.

Intersex is alternatively spliced in B. mori

Intersex, a gene implicated in female sexual development in *Drosophila*, is expressed in both sexes and functions together with *doublesex* (*dsx*) to regulate terminal sex differentiation. It is known to express more in females of *Drosophila* and produce only one splice form in all the tissues of *Drosophila* males and females (Garrett-Engele et al., 2002). *B. mori* homolog of *ix* was identified

recently (Siegal and Baker, 2005). In the same study, attempts were made to rescue the wild type phenotype in *ix* mutants of *Drosophila* using *B. mori ix* cDNA. Our study suggests that this gene produces a second splice form expressing only in testis. Since the *ix* phenotype is female specific and some genes in the somatic sex-determination hierarchy are regulated at the level of splicing, it is speculated that the *Bmix* pre-mRNA would be sex-specifically spliced.

Through BLAST analysis of non-redundant set of *B. mori* testis transcripts, we identified a homolog of *ix*. Further, mining for this transcript in other tissue ESTs revealed a transcript in ovary. Pair-wise alignments showed distinct alternative splice forms in ovary and testis. A contig of 16,856 bp with accession number AADK01005119, harboring all the six exons was obtained by searching *B. mori* WGS database. The complete gene structure was deduced by aligning ovary and testis *Bmix* ESTs to this contig. The total gene size was found to be 7.8 kb which includes a total of ~7.2 kb intron region (Figure 3a).

Expression profile of *Bmix* gene alternative splice forms was carried out by designing a common forward primer for exon 1 (primer name: ixA) and separate reverse primers for exons 3 (ixB), 4 (ixC), 5 (ixD) and 6 (ixE). RT-PCR analysis using different combination of above mentioned primer sets (ixA-ixB, ixA-ixC, ixA-ixD, and ixA-ixE) revealed the presence of different exons in the tissues tested (data not shown). The common splice form of *Bmix* containing all but exon 2 was present in all the tissues tested. However in testis it showed an additional splice form containing all the exons (Figures 3b & 3c). The RT-PCR amplicons were cloned into TA vector and subsequently sequenced to verify the sequence of the two splice forms. Further confirmation of testis specific splice form was conducted through RNase protection assay, which also showed the presence of testis specific splice form (Figure 3c). Conceptual translation of both the transcripts showed that common splice form has a protein product of 192 amino acids and since the second exon of *Bmix* has a stop codon at the beginning, the testis specific splice form protein product had 72 amino acids.

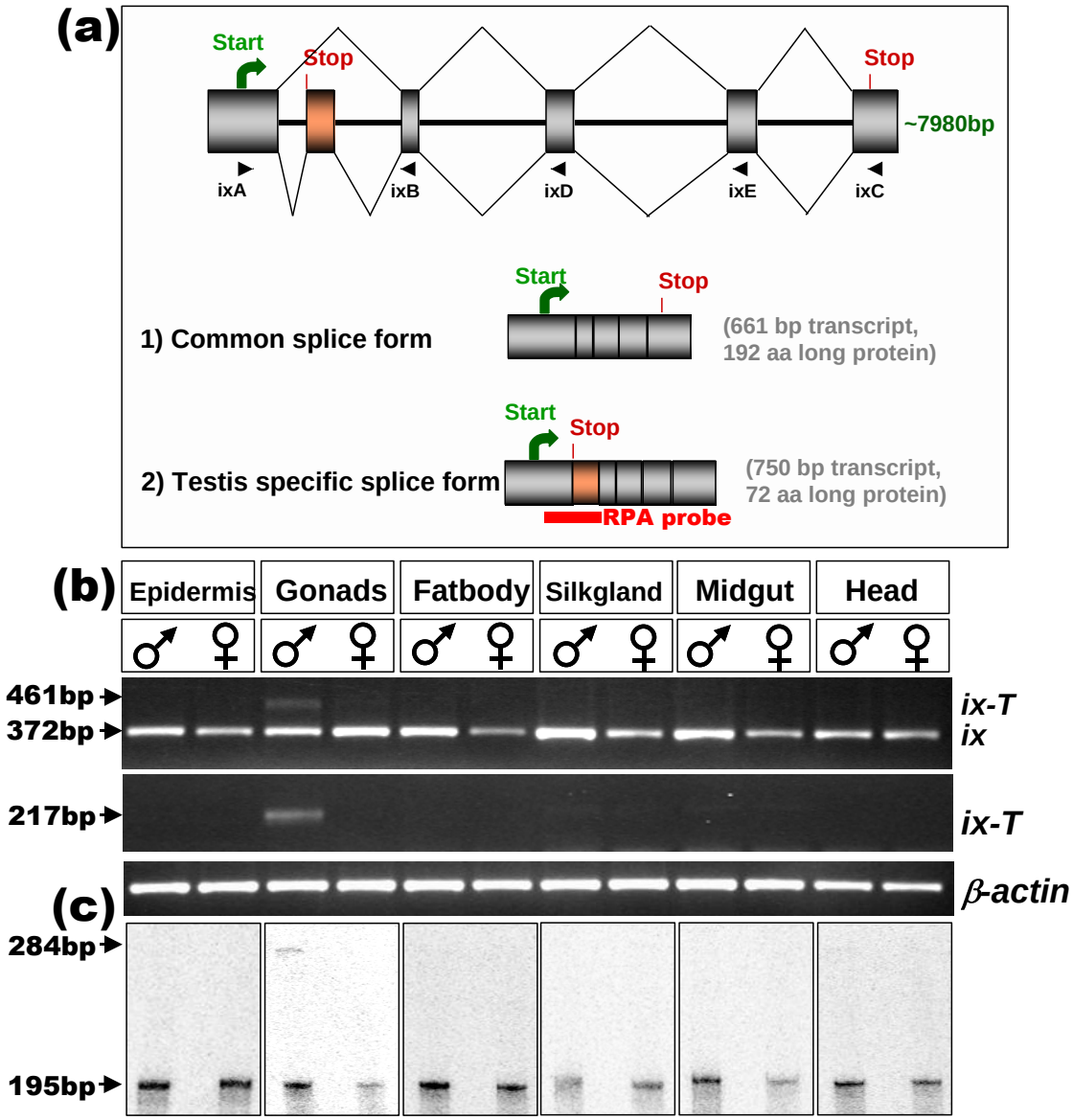


Figure 3: Gene structure and alternative splice forms of *Bmix* (a) and expression of *Bmix* mRNA during different developmental stages of *B. mori* as revealed by RT-PCR. The primers ixA, ixB, ixC, ixD and ixE, are designed for different exons and black arrows indicate their location and direction of primer extension. (b) Upper panel shows RT-PCR profile of different tissues from 5th instar larvae amplified using ix-A and ix-E primers, which will amplify both common (*Bmix*) and testis specific (*Bmix-T*) splice forms. Middle panel is the RT-PCR profile, obtained by using primers binding to first and second exons which will specifically amplify *Bmix-T*. Lower panel shows β-actin levels as control. (c) Shows the RNase protection assay for six tissues of both sexes separately, of fifth instar larvae, wherein only testis is showing the second splice form. 'RPA probe' in (a) indicates the cDNA region used as probe for RNase protection assay.

In contrast to *Drosophila*, where *ix* is single exon gene and is not sex specifically spliced, the *Bmix* possesses six exons among which only five exons are represented in transcripts expressed in all the tissues. Whereas all exons are present in one of the two alternative splice forms present in testes. Earlier studies in *Drosophila* (Baker and Ridge, 1980, Chase and Baker, 1995) had suggested that the IX is required to function with the female-specific product of the *dsx* gene to implement female sexual differentiation in diplo-X animals. However, other studies have favored that IX might function not only with DSX to control many aspects of somatic sex, but also independently of DSX to regulate other aspects of somatic sex (McRobert and Tompkins, 1985). Chase & Baker (1995) observed that, while DSX proteins are capable of binding to the sex-specific enhancer site of the *yolk protein* (YP) genes, the product of *ix* may not be required to achieve high levels of YP transcription (Chase and Baker, 1995). Another study (Acharyya and Chatterjee, 2002), revealed the possible functions of *ix* in *Drosophila* males by characterizing an allele of *ix* (*ix5*). The presence of additional splice form of *ix* in *B. mori* strongly supports the above findings in *Drosophila*, that *ix* is also essential for male differentiation. Attempts made to rescue the wild type phenotype in *ix* mutants of *Drosophila* using *B. mori ix* cDNA resulted in near complete rescue of the *ix*-mutant females. This suggests partial functional divergence between the *D. melanogaster* and *B. mori* IX proteins. We suspect that *ix* may have many other roles in sex differentiation, than the previously reported ones. Also, testis specific *Bmix* splice form possibly may have a role in spermatogenesis. However these have to be confirmed experimentally through RNAi knock down.

Previous study (Siegal and Baker, 2005) hypothesizes that potential diversity of biological processes in which IX homologs participate is at least as great as that in which DSX homologs participate, because it is possible that IX homologs have interaction partners beyond the DM protein family. In that study though the authors report the almost complete rescue of wild type female phenotype and they also opine that mere ability of a homolog to replace *D. melanogaster ix* does not guarantee that the homolog is functioning in sex determination in the donor species. Bucking trend of *ix* in *B. mori* tempts us to speculate that IX in lepidopterans may have a different function. However, experimental evidence is needed to unravel the precise function of ix in *B.mori*.

Sequencing and phylogenetic analysis of intersex homologs in other lepidopteras

Putative homologs of IX in *A. assama*, *S. c. ricini* and *A. mylitta* were identified either by searching WildSilkbase (Arunkumar et al., 2008) or by *de novo* sequencing. Other homologs were

downloaded from NCBI. Protein sequences were aligned and an unrooted tree was constructed. The resulted tree grouped the insect species together in one group and mammals in another group. The phylogenetic analysis showed close nesting of lepidopteras. All the lepidopteras studied belonged to one super family Bombycoidae. However the tree showed that IX is highly conserved within these species when compared to *Drosophila* species (*D. melanogaster*, *D. pseudoobscura* and *D. virilis*). Though all the three *Drosophila* species belonged to same family Drosophilidae, they showed loose clustering compared to members of Bombycoidae (Figure 4).

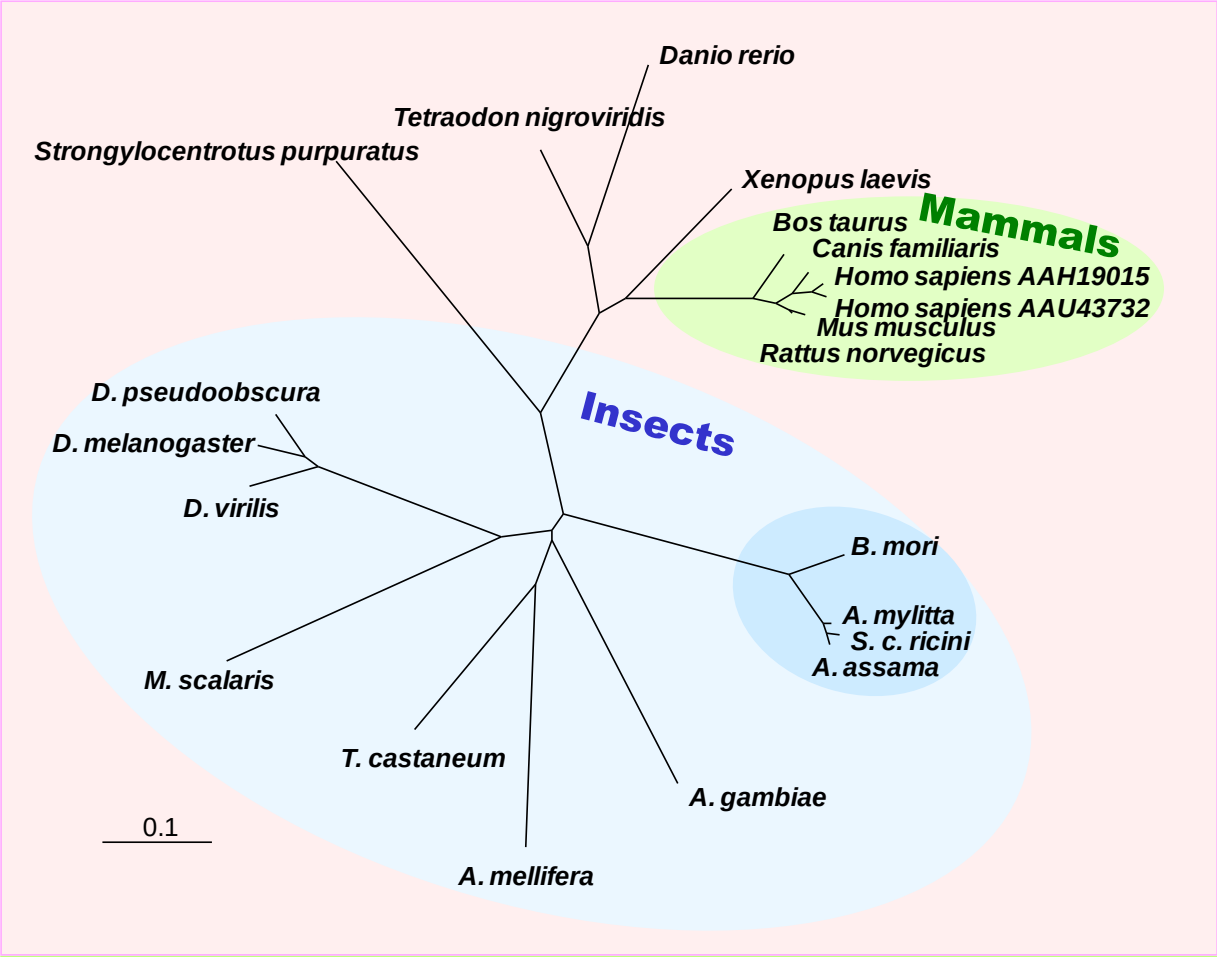


Figure 4: Dendrogram showing the relationship of Intersex protein between different species.

To study whether the *ix* homolg in *A. assama* also shows presence of alternative splice forms, RT-PCR products were obtained from whole larval RNA and sequenced. We obtained 2 kinds of sequences with a stretch of overlapping sequences between them. On alignment the

presence of an additional 51 bp sequence in one of the isoforms was apparent. Conceptual translation yielded a protein with extra 17 aa as the extra sequence did not possess stop codons. The presence of this extra exon is on same region as the occurrence of testis specific exon in *Bmix*. However there is no sequence similarity between second exon of *B. mori* (present only in testis specific splice form) and the extra exon of *A. assama*.

Physical mapping reveals abundance of testis specific genes on Z chromosome

Evolutionary significance of sex chromosomes probably involves not only the mechanism of sex determination but also the evolution of sexually dimorphic traits. Sexual dimorphism results from natural selection that favors different phenotypic characteristics in the two sexes (Rice, 1984). The adult males of many lepidopteras fly more actively than females, while female moths usually wait for males and attract them by releasing sex pheromones. Although neither males nor females of *B. mori* can fly, males flap their wings more vigorously than females. Possibly during domestication process, *B. mori* would have lost its flight character. Recent findings show that there is no dosage compensation in *B. mori* and other lepidopteras (Koike et al., 2003, Suzuki et al., 1998, 1999). In butterflies, it has been suggested that the absence of dosage compensation affects adaptation and sexual selection (Grula and Taylor Jr, 1980, Sperling, 1994, Stehr, 1959). Several genes for mate selection in females and the courtship signals in males are located on the Z chromosome in butterflies (Grula and Taylor Jr, 1980). Analyzing the distribution of sex specific genes between sex chromosomes and autosomes would help delineate the underlying molecular mechanisms of sexual dimorphism in *B. mori*.

Studies have been carried out to analyze the distribution of sex specific and/or sex enhanced genes on different chromosomes of model organisms. Human genome analysis and gene expression studies have revealed that the X chromosome has more than its fair share of genes involved in sex and reproduction. In contrast, the male-biased genes in *Drosophila* (Parisi et al., 2003, Ranz et al., 2003) and genes expressed in spermatogenic and oogenic cells in *C. elegans* (Reinke et al., 2004) seem to be underrepresented on the X chromosome.

The biased representation of sex specific genes on sex chromosomes is attributed to two main reasons, sexual antagonism and dosage compensation. According to the hypothesis of sexual antagonism (Hurst, 2001, Rice, 1984), an unusual homogametic sex chromosome gene content reflects a non-random accumulation of sexually antagonistic mutations on this chromosome. This is caused by the different time that the homogametic sex chromosome has spent in the two sexes

and by its hemizygous exposure in the heterogametic sex. The other hypothesis concerns the epigenetic modifications of the sex chromosomes associated with meiotic sex chromosome inactivation and dosage compensation (Khil et al., 2005, Parisi et al., 2003, Reinke et al., 2004, Rogers et al., 2003). However, the exact role of these mechanisms still remains elusive, as all the studies have been carried out in male heterogametic sex chromosome systems (XX/XY system).

Alternatively, the mechanisms responsible for the non-random representation of sex-biased and sex-specific genes on the homogametic sex chromosome may be clarified by analyzing the gene content of the Z chromosome, a homogametic sex chromosome in heterogametic female organisms. Although the X chromosome occurs more frequently in female individuals, the Z chromosome spends more time in male individuals. If sexually antagonistic selection were the primary mechanism affecting the sex chromosome gene content, opposite trend would be expected in the representation of sex-biased and sex-specific genes on the X and Z chromosomes (Storchova and Divina, 2006). Towards this effect, recently a study was reported in chicken, which is a female heterogametic system like *B. mori*. Compared with the genes expressed in other somatic tissues, the male-biased genes expressed in chicken brain were significantly enriched on the Z chromosome, whereas the female biased genes were deficient on the Z chromosome. However, the study did not find any biasness in distribution of testis specific genes on Z chromosome (Storchova and Divina, 2006).

Unlike in *Drosophila*, the sex chromosome constitution of the *B. mori*, is ZW in the female and ZZ in the male and there is lack of dosage compensation. This has been demonstrated by analyzing the mRNA levels of genes on Z chromosome. It is found that mRNA levels of several Z chromosome genes in males are twice as in females (Koike et al., 2003, Suzuki et al., 1998, 1999). Speculation is that the Z chromosome has evolved through a process of genome shuffling to carry only genes whose products are required at higher levels in males. Perhaps genes that must be expressed in equal amounts in the two sexes have already translocated onto autosomes or acquired gene-specific regulatory mechanisms. These issues can be addressed by studying the sex specific or sex enhanced gene composition of Z chromosome in *B. mori*.

Recently, through microarray analysis a total of 1104 genes was identified to be testis specific in *B. mori* (Xia et al., 2007). In the present study, these microarray validated testis specific genes were assigned chromosomal position on *B. mori* genome, to examine whether there is any biasness in their representation on Z chromosome. We were able to successfully map 1028 genes onto their respective chromosomal location. Physical mapping revealed many interesting features about the frequency and distribution of testis specific genes on *B. mori* chromosomes.

Surprisingly, Z chromosome harbored highest number (82) of testis specific genes. Average number of testis specific genes on autosome was calculated to be 35 (standard deviation $\pm$10). The student t-test revealed that Z chromosome harbored significantly higher ($P<0.001$) number of genes compared to that on autosomes. Among autosomes, chromosome B harbored lowest number (9) of testis specific genes (Figure 5).

To study the frequency of occurrence of testis specific genes on different chromosomes, number of genes per Mb of each chromosome was calculated. Z showed highest frequency of testis specific genes, with 4.03 genes per Mb of chromosome; whereas Chromosome B had least frequency (0.86) of these genes. The student t-test showed that Z harbored significantly high frequency of testis specific genes per Mb ($P<0.001$) compared to autosomes. Physical mapping of these genes revealed that there is apparent biasness in distribution of these testis specific genes on the chromosomes. In order to identify if there are any clusters of testes specific genes, we constructed a physical map using Karyoview (http://www.ensembl.org/index.html). The map showed the spatial distribution of genes on *B. mori* chromosomes (Figure 6). It was evident from the map that genes are quite evenly and densely distributed on Z chromosomes. The analysis confirmed the presence of several clusters of testis specific genes on different chromosomes. The information on exact location of each gene on chromosomes can be found in the Appendix.

We also assessed the allocation of the tissue-specific genes (the genes expressed exclusively in one tissue which excludes tissues testis and ovary) between autosomes and the Z chromosome. For this purpose, we mapped 465 other tissues specific genes onto *B. mori* chromosomes, which showed no chromosome bias in the distribution of these genes (Figure 5). This shows that enrichment of testis-specific genes on Z is not due to its sheer size or due to higher frequency of genes on Z chromosome.

To further verify this phenomenon we mapped testis specific genes as identified through fl-cDNA and EST analysis. We mapped 1857 testis specific genes onto *B. mori* chromosomes. Here also we could find the significant enrichment of testis specific genes on Z chromosome (Figure 7). These results are consistent with our observation that the Z chromosome is enriched for testis-specific genes.

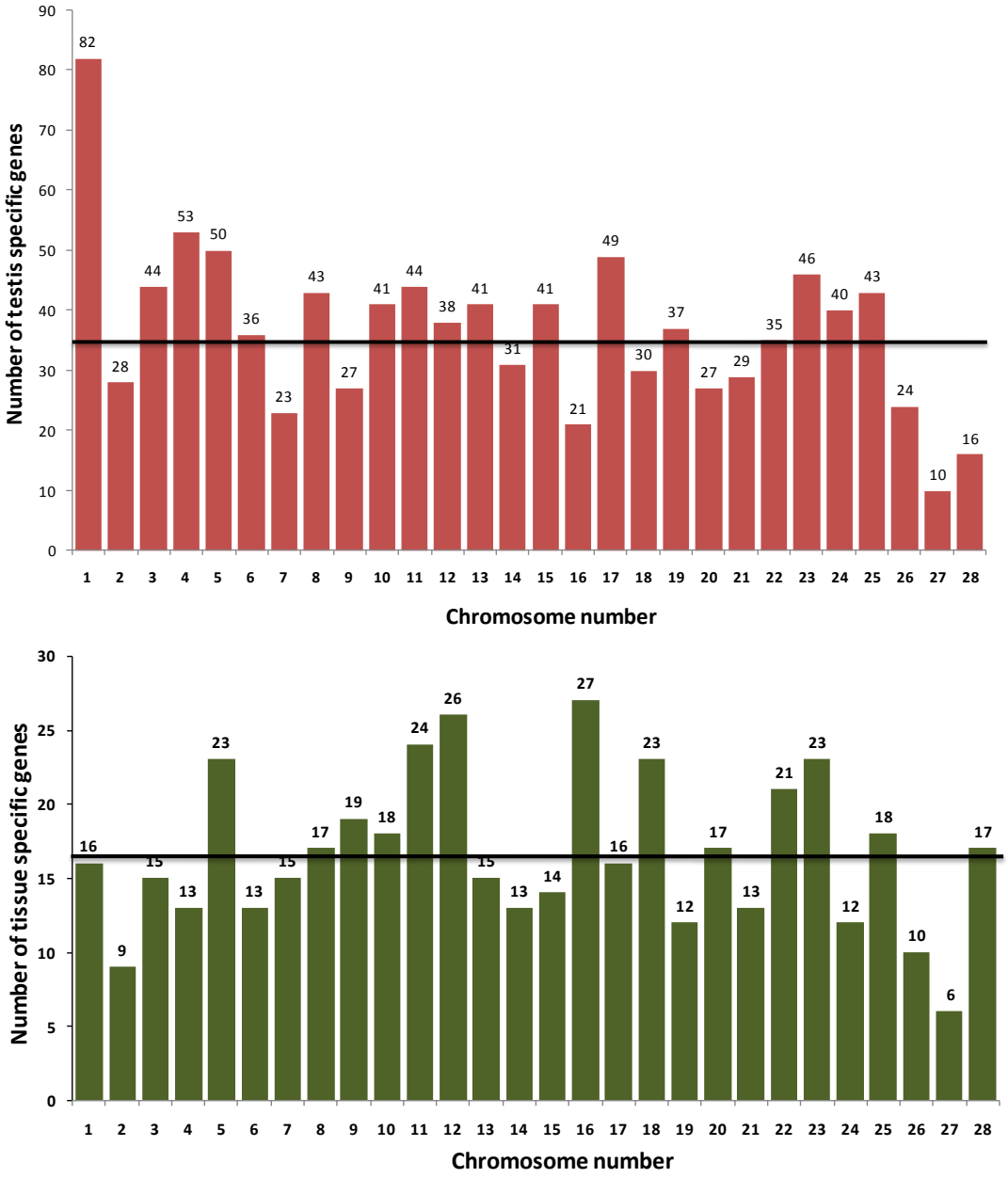


Figure 5: Distribution of testis specific genes (Upper panel) and other tissue specific genes (Lower panel), as identified through microarray analysis, on different chromosomes of *B. mori*. There is a significant difference (P<0.001) between number of testis specific genes present on Z chromosome and autosomes. Out of 1102 microarray validated testis specific genes, 1029 were successfully mapped on to *B. mori* chromosomes (Upper panel). Average number of testis specific genes on autosomes was calculated to be 35, which is indicated by a black horizontal line on the histogram. Out of 501 other tissue specific genes only 465 were successfully mapped (Lower panel). Average number of other tissue specific genes on autosomes was calculated to be 16, which is indicated by a black horizontal line on the histogram in the lower panel.

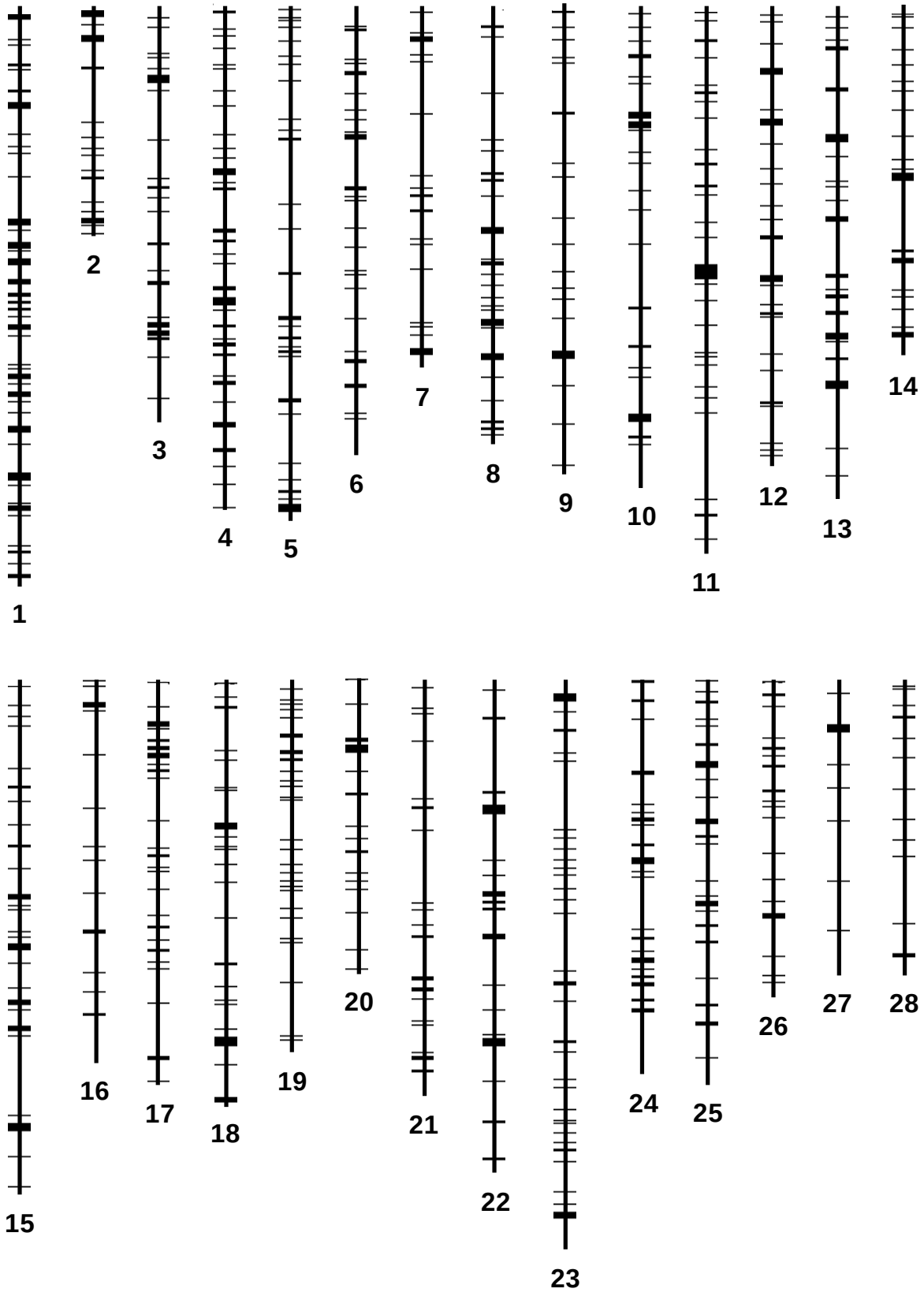


Figure 6: Physical map of *B. mori* chromosomes showing the distribution of testis specific genes on different chromosomes.

Previous studies have shown that in butterflies a disproportionate number of genes related to sexuality, reproduction and speciation are located on the Z chromosome, which forms approximately one-sixtieth of the genome in females. Female mate-selection behavior, male courtship signals, female limitation of color polymorphism and mimicry are thought to result largely from interactions between autosomal genes and uncompensated Z-linked regulatory genes (Cook, 1964, Grula and Taylor Jr, 1980, Sheppard, 1961, Stehr, 1959). The present study in *B. mori* corroborates these findings.

The Rice thesis (Rice, 1984) suggests that when mutations occur favoring accumulation of sexually antagonistic genes, they will be selected naturally if they are located on sex chromosomes rather than autosomes. X linkage facilitates the initial build-up of recessive and partially recessive sexually antagonistic gene that favors the heterogametic sex. The evolutionary dynamics of male-benefit mutations were considered when they first appear as rare alleles on X chromosomes as opposed to autosomes. When they are rare, autosomal recessive alleles would be of no advantage to (heterozygous) males and thus would be unlikely to spread widely in the population. By contrast, X-linked recessive alleles would immediately benefit hemizygous males, greatly increasing the alleles' likelihood of permeating the population. Eventually, as an allele's frequency increased in the population, female fitness would be diminished by the detrimental effects of homozygosity. This would generate adaptive pressure to limit the gene's expression to males, through additional mutations. Therefore, it was postulated that X chromosomes should evolve to carry a disproportionate share of male-specific genes functioning in male differentiation (Rice, 1984, Wang et al., 2001).

By contrast, Z linkage can facilitate the increase of traits which favor homogametic sex also. The dominant mutations favoring the homogametic sex have a greater chance to be fixed on the Z chromosome where they are exposed to selection, because $2/3^{rd}$ of the Z chromosomes reside in homogametic individuals compared to only $1/3^{rd}$ in heterogametic individuals and due to absence of dosage compensation. In Z chromosome, dominant sexually antagonistic genes that favor the homogametic sex are expressed at a higher rate in the sex where they are positively selected. Our findings are in accordance with this prediction. However, studies on the X/Z chromosome gene content in different organisms offer conflicting results. In case of Z chromosome systems, results obtained in *B. mori* physical mapping of testis specific genes, are in contrary to what was observed in chicken.

Possible translocation of male advantageous genes on to Z chromosomes from autosomes

Duplicated genes may acquire novel functions and altered expression patterns (Ohno et al., 1968) and thus contribute to diversification of tissues during development (Mikhaylova et al., 2008). Analysis of such paralogs provides opportunity to study genome evolution as it sheds light on history of gene duplication and gene trafficking between chromosomes through translocation events. In *Drosophila*, a number of testis-specific genes have been reported to be generated by gene duplications, where the duplicated gene is specifically expressed in the male reproductive system while the parental gene is ubiquitous (Betran and Long, 2003, Nurminsky et al., 1998). In the present study we have used the data on paralogous genes that express in *B. mori* testis to investigate the evolutionary causes that led to the enrichment of testis specific genes on Z chromosome. Through BLAST analysis we obtained 30 groups of paralogs (comprising 74 out of 1104 testis specific genes), among them only 12 were found in clusters on different chromosomes. Whereas the remaining 18 did not show any clustering, though some paralogs were mapped on to same chromosome but located far apart from each other. Duplications were more frequent among genes coding for dynein proteins (3 out of 30 paralogous groups), probably because of its requirement in large amounts in sperms for motility (Moss et al., 1992). None of the clustered co-expressed testis specific genes showed similarity to transposable elements, which was not the case in non-clustered paralogous groups where 6 paralogous groups were found to code for transposable elements (e.g., reverse transcriptase, transposon polyprotein and transposase).

One or more genes in 6 non-clustered paralogous groups were located on Z chromosome. In other words, around 30% of the non-clustered paralogs have a copy on Z chromosome, which is the highest for any chromosome. Our speculation is that during the course of evolution male advantageous genes were translocated onto Z chromosome and were selected positively. This may be the reason for higher occurrence of one of the copies of testis specific non-clustered paralogs on Z chromosomes. Also, three (25%) of the 12 co-expressed clusters were present on Z chromosome, which is the highest for any chromosome. Based on these results we surmise that testis specific paralogs are concentrated more on Z chromosomes either by clustering or by the location of one of the copies of non-clustered paralogs (Figure 7). These translocations followed by the gene fixation on Z chromosome possibly may be responsible for enrichment of testis specific genes on Z chromosome. Analysis of expression patterns of the paralogous genes in testes would provide valuable indications to elicit the functional relationships between these genes (Mikhaylova et al., 2008).

Functional annotation of testis specific genes originating from Z chromosomes revealed presence of genes coding for a variety of proteins. Of the 82 testis specific genes on Z chromosome, 5 showed no match to any proteins in nr database of NCBI. Six genes showed similarity to transposable elements and three genes code for dynein proteins.

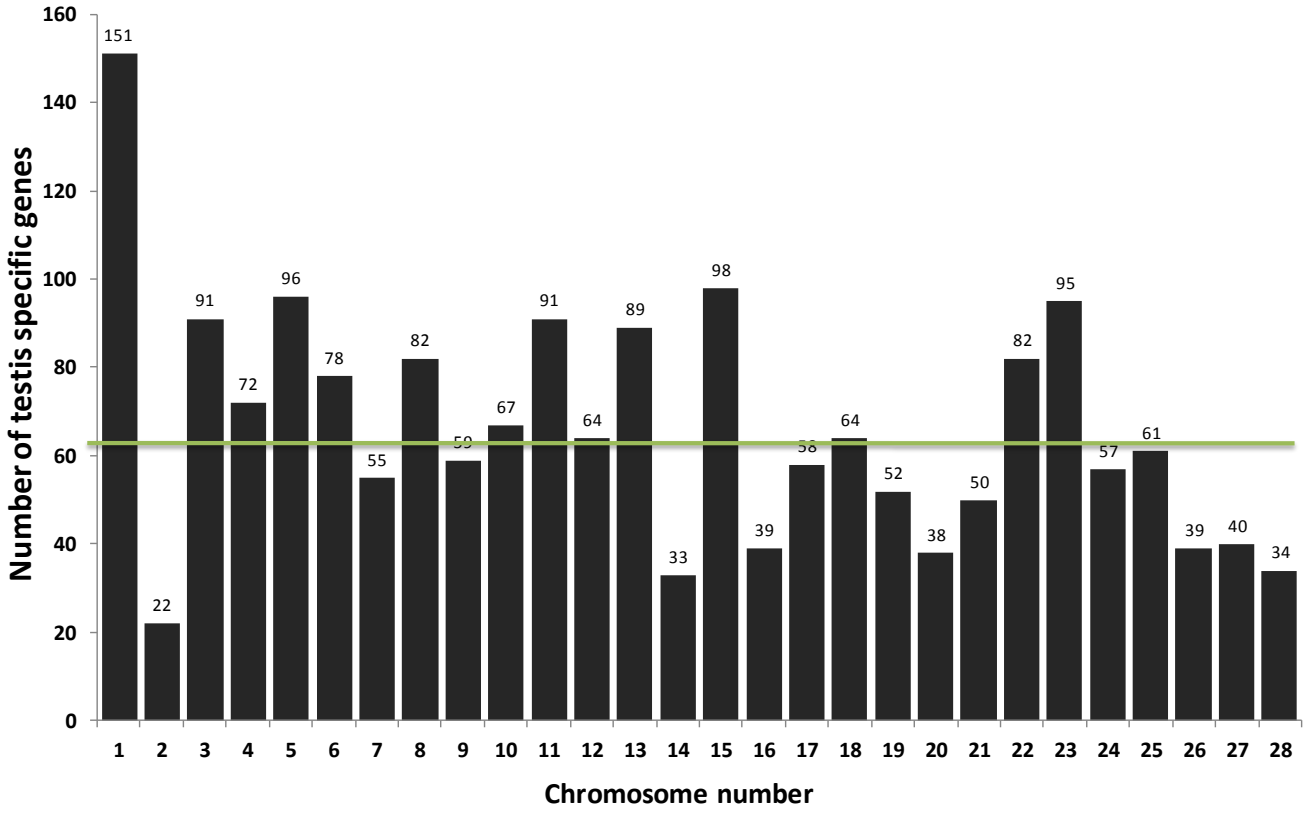


Figure 7: Distribution of testis specific genes as identified through EST and fl-DNA sequence analysis on different chromosomes of *B. mori*. Out of 1984 testis specific genes identified through analysis of testis derived fl-cDNAs and ESTs, 1857 genes could be mapped onto *B. mori* chromosomes. Average number of genes on autosomes was calculated to be 63 (green horizontal line).

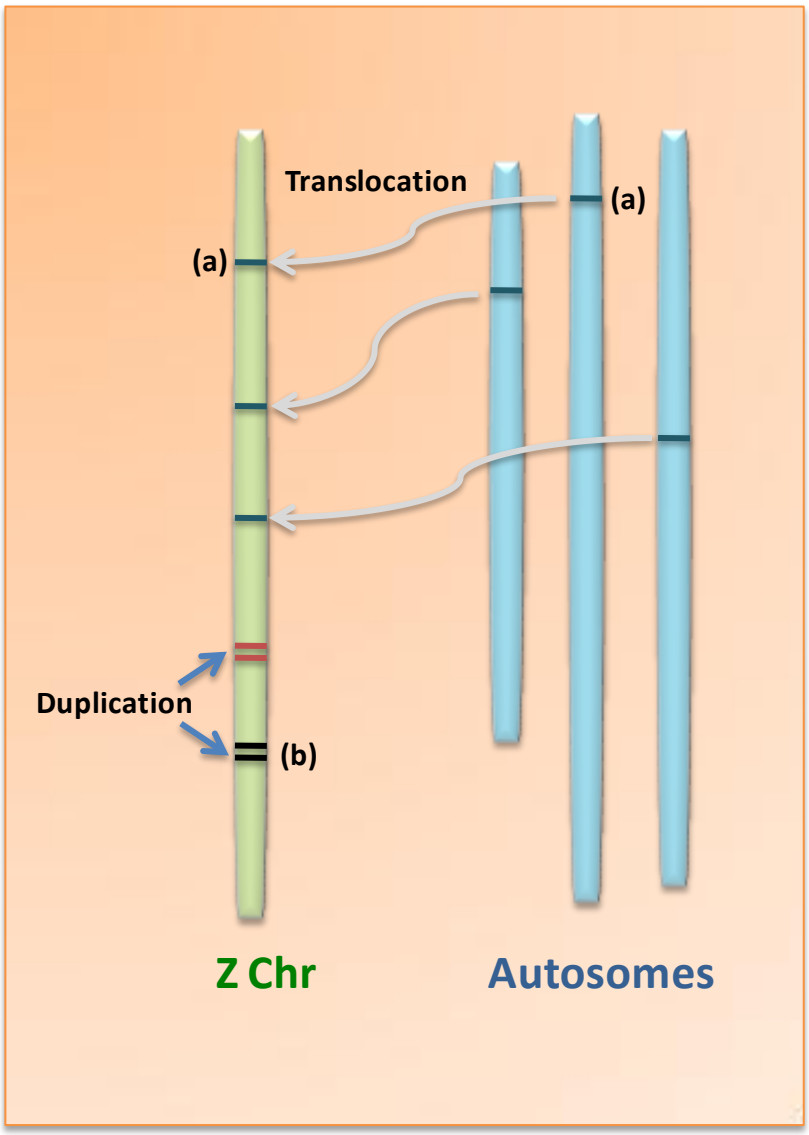


Figure 8: Possible mechanism of accumulation of male advantageous genes on Z chromosome through translocation from autosomes and local duplication events, and their fixation on Z. (a) non-clustered paralogs, (b) clustered paralogs.

Conclusions

The complex regulatory mechanisms underlying male germ cell development in lepidoptera offer many challenges. Apparently spermatogenic stage and cell specific gene expression are crucial for the developmental changes. *In silico* analyses of *B. mori* transcriptome provide a foundation for further in-depth analysis of activities in male germ cells. High gene diversity in the testis transcriptome is suggestive of the complex mechanisms of spermatogenesis. Information produced by the sequencing of *B. mori* cDNA libraries and identification of testis-specific genes in the present study provide new insights into the structure, diversity and molecular evolution of genes involved in spermatogenesis in lepidopteras in particular and insects in general. Further functional analysis of several testis specific genes identified, would reveal mechanisms of male germ cell development and differentiation. The present study gives evidence for the presence of alternative splice form of *ix* gene reported for the first time in any organism. Future experiments should be carried out to find the function of the *Bmix* gene in both the sexes, by using RNAi and transgenesis approaches. The study on distribution of testis specific genes on *B. mori* chromosomes has shown that male advantageous genes are accumulated on Z chromosomes either by translocation from other chromosomes or by tandem duplication of such genes on Z, which supports the hypothesis of sexual antagonism.

1. Acharyya M, Chatterjee RN (2002) Genetic analysis of an intersex allele (ix5) that regulates sexual phenotype of both female and male *Drosophila melanogaster*. Genet Res 80: 7-14.

2. Ahmad R, Kamra A, Hasnain SE (2004) Fibroin silk proteins from the nonmulberry silkworm *Philosamia ricini* are biochemically and immunochemically distinct from those of the mulberry silkworm *Bombyx mori*. DNA Cell Biol 23: 149-154.

3. Akai H (2000) Cocoon filament characters and post-cocoon technology. Int J Wild Silkmoths Silk 5: 255-259.

4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410.

5. Andersen SO (1998) Amino acid sequence studies on endocuticular proteins from the desert locust, *Schistocerca gregaria*. Insect Biochem Mol Biol 28: 421-434.

6. Andersen SO (2000) Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. Insect Biochem Mol Biol 30: 569-577.

7. Andersen SO, Hojrup P, Roepstorff P (1995) Insect cuticular proteins. Insect Biochem Mol Biol 25: 153-176.

8. Andrews J, Bouffard GG, Cheadle C, Lu J, Becker KG, et al. (2000) Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res 10: 2030-2043.

9. Arunkumar KP, Metta M, Nagaraju J (2006) Molecular phylogeny of silkmoths reveals the origin of domesticated silkmoth, *Bombyx mori* from Chinese *Bombyx mandarina* and paternal inheritance of *Antheraea proylei* mitochondrial DNA. Mol Phylogenet Evol 40: 419-427.

10. Arunkumar KP, Tomar A, Daimon T, Shimada T, Nagaraju J (2008) WildSilkbase: An EST database of wild silkmoths. BMC Genomics 9: 338.

11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25-29.

12. Baarends WA, Grootegoed JA (1999) Molecular biology of male gametogenesis. In: Fauser B, editor. Molecular Biology in Reproductive Medicine. New York: The Parthenon Publishing Group. pp. 271–295.

13. Baker BS, Ridge KA (1980) Sex and the single cell. I. On the action of major loci affecting sex determination in *Drosophila melanogaster*. Genetics 94: 383-423.

14. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids

Res 27: 573-580.

15. Betran E, Long M (2003) Dntf-2r, a young *Drosophila* retroposed gene with specific male expression under positive Darwinian selection. Genetics 164: 977-988.

16. Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, et al. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33: D212-215.

17. Castelo AT, Martins W, Gao GR (2002) TROLL--tandem repeat occurrence locator. Bioinformatics 18: 634-636.

18. Chase BA, Baker BS (1995) A genetic analysis of intersex, a gene regulating sexual differentiation in *Drosophila melanogaster* females. Genetics 139: 1649-1661.

19. Cohen E (1987) Chitin biochemistry: synthesis and inhibition. Ann Rev Entomol 32: 71-93.

20. Consortium GO (2006) The Gene Ontology (GO) project in 2006. Nucleic Acids Res 34: D322-326.

21. Cook AG (1964) Dosage compensation and sex-chromatin in non-mammals. Genet Res 5: 354-365.

22. Couble P, Michaille JJ, Garel A, Couble ML, Prudhomme JC (1987) Developmental switches of sericin mRNA splicing in individual cells of *Bombyx mori* silkgland. Dev Biol 124: 431-440.

23. Craig CL (1997) Evolution of arthropod silks. Annu Rev Entomol 42: 231-267.

24. Dash R, Mukherjee S, Kundu SC (2006) Isolation, purification and characterization of silk protein sericin from cocoon peduncles of tropical tasar silkworm, *Antheraea mylitta*. Int J Biol Macromol 38: 255-258.

25. Deodikar GB, Chowdhury SN, Bhuyan BN, Kshirsagar KK (1962) Cytogenetic studies in Indian silkworms. Curr Sci 31: 247-248.

26. Eddy EM (2002) Male germ cell gene expression. Recent Prog Horm Res 57: 103-128.

27. Eddy EM, O'Brien DA (1998) Gene expression during mammalian meiosis. Curr Top Dev Biol 37: 141-200.

28. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8: 175-185.

29. Fedic R, Zurovec M, Sehnal F (2002) The silk of Lepidoptera. J Insect Biotech Sericol 71: 1-15.

30. Gamo T, Inokuchi T, Laufer H (1977) Polypeptides of fibroin and sericin secreted from the different sections of the silk gland in *Bombyx mori*. Insect Biochem 7: 285-295.

31. Gandhe AS, Arunkumar KP, John SH, Nagaraju J (2006) Analysis of bacteria-challenged wild silkmoth, *Antheraea mylitta* (lepidoptera) transcriptome reveals potential immune genes. BMC Genomics 7: 184.

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

33. Garrett-Engele CM, Siegal ML, Manoli DS, Williams BC, Li H, et al. (2002) intersex, a gene required for female sexual development in *Drosophila*, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation. Development 129: 4661-4675.

34. Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. Cell 110: 315-325.

35. Glenn TC, Schable NA (2005) Isolating microsatellite DNA loci. Methods Enzymol 395: 202-222.

36. Graveley BR (2001) Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17: 100-107.

37. Grimaldi DA, Engel MS (2005) Evolution of the Insects. New York: Cambridge University Press. 755 p.

38. Grula JW, Taylor Jr OR (1980) The Effect of X-Chromosome Inheritance on Mate-Selection Behavior in the Sulfur Butterflies, *Colias eurytheme* and *C. philodice*. Evolution 34: 688-695.

39. Gupta ML, Narang RC (1981) Karyotype and meiotic mechanism in Muga silkmoths, *Antheraea compta* Roth. and *A. assamensis* (Helf.) (Lepidoptera: Saturniidae) Genetica 57: 21-27.

40. Hardison RC (2003) Comparative genomics. PLoS Biol 1: E58.

41. Hasimoto H (1933) The role of the W-chromosome in the sex determination of *Bombyx mori*. Jpn J Genet 8: 245-247.

42. He X, Zhang J (2005) Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics 169: 1157-1164.

43. Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S (2000) Blocks-based methods for detecting protein homology. Electrophoresis 21: 1700-1706.

44. Huang X, Madan A (1999) CAP3: A DNA sequence assembly program. Genome Res 9: 868-877.

45. Huntley D, Baldo A, Johri S, Sergot M (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics 22: 495-496.

46. Hurst LD (2001) Evolutionary genomics. Sex and the X. Nature 411: 149-150.

47. Iizuka E (2000) Physical properties of silk thread from cocoons of various wild silkmoths including domestic silkmoth. Int J Wild Silkmoths Silk 5: 266-269.

48. Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314: 67-73.

49. Jolly MS, Sen SK, Sonwalker TN, Prasad GK (1981) FAO Agriculture Services Bulletin: Non-Mulberry silks. Rome: FAO

50. Kawasaki H, Sugaya K, Quan GX, Nohata J, Mita K (2003) Analysis of alpha- and beta-tubulin genes of *Bombyx mori* using an EST database. Insect Biochem Mol Biol 33: 131-137.

51. Kerr SM, Vambrie S, McKay SJ, Cooke HJ (1994) Analysis of cDNA sequences from mouse testis. Mamm Genome 5: 557-565.

52. Khil PP, Oliver B, Camerini-Otero RD (2005) X for intersection: retrotransposition both on and off the X chromosome is more frequent. Trends Genet 21: 3-7.

53. Kleene KC (2001) A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev 106: 3-23.

54. Koike Y, Mita K, Suzuki MG, Maeda S, Abe H, et al. (2003) Genomic sequence of a 320-kb segment of the Z chromosome of *Bombyx mori* containing a kettin ortholog. Mol Genet Genomics 269: 137-149.

55. Kusakabe T, Kawaguchi Y, Maeda T, Koga K (2001) Role of interaction between two silkworm RecA homologs in homologous DNA pairing. Arch Biochem Biophys 388: 39-44.

56. Larsson M, Norrander J, Graslund S, Brundell E, Linck R, et al. (2000) The spatial and temporal expression of Tekt1, a mouse tektin C homologue, during spermatogenesis suggest that it is involved in the development of the sperm tail basal body and axoneme. Eur J Cell Biol 79: 718-725.

57. Lucas F, Rudall KM (1968) Extracellular fibrous proteins: The silks; Florkin M, Stota EH, editors. Amsterdam: Elsevier. 594 p.

58. Lynch M, Katju V (2004) The altered evolutionary trajectories of gene duplicates. Trends Genet 20: 544-549.

59. Magkrioti CK, Spyropoulos IC, Iconomidou VA, Willis JH, Hamodrakas SJ (2004) cuticleDB: a relational database of Arthropod cuticular proteins. BMC Bioinformatics 5: 138.

60. Mahendran B, Ghosh SK, Kundu SC (2006) Molecular phylogeny of silk-producing insects based on 16S ribosomal RNA and cytochrome oxidase subunit I genes. J Genet 85: 31-38.

61. McRobert SP, Tompkins L (1985) The effect of transformer, doublesex and intersex mutations on the sexual behavior of *Drosophila melanogaster*. Genetics 111: 89-96.

62. Mikhaylova LM, Nguyen K, Nurminsky DI (2008) Analysis of the *Drosophila melanogaster* testes transcriptome reveals coordinate regulation of paralogous genes. Genetics 179: 305-315.

63. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, et al. (2004) The genome sequence of silkworm, *Bombyx mori*. DNA Res 11: 27-35.

64. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, et al. (2003) The construction of an EST database for *Bombyx mori* and its application. Proc Natl Acad Sci U S A 100: 14121-14126.

65. Mita K, Nenoi M, Morimyo M, Tsuji H, Ichimura S, et al. (1995) Expression of the *Bombyx mori* beta-tubulin-encoding gene in testis. Gene 162: 329-330.

66. Miyagawa Y, Lee JM, Maeda T, Koga K, Kawaguchi Y, et al. (2005) Differential expression of a *Bombyx mori* AHA1 homologue during spermatogenesis. Insect Mol Biol 14: 245-253.

67. Moran JV, DeBerardinis RJ, Kazazian HH, Jr. (1999) Exon shuffling by L1 retrotransposition. Science 283: 1530-1534.

68. Moss AG, Sale WS, Fox LA, Witman GB (1992) The alpha subunit of sea urchin sperm outer arm dynein mediates structural and rigor binding to microtubules. J Cell Biol 118: 1189-1200.

69. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature 435: 903-910.

70. Negre V, Hotelier T, Volkoff AN, Gimenez S, Cousserans F, et al. (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. BMC Bioinformatics 7: 322.

71. Neville AC (1975) Biology of the Arthropod Cuticle. Berlin: Springer.

72. Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL (1998) Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature 396: 572-575.

73. Nusslein-Volhard C, Wieschaus E (1980) Mutations affecting segment number and polarity in *Drosophila*. Nature 287: 795-801.

74. Ohno S, Wolf U, Atkin NB (1968) Evolution from fish to mammals by gene duplication.

Hereditas 59: 169-187.

75. Okamoto H, Ishikawa E, Suzuki Y (1982) Structural analysis of sericin genes. J Biol Chem 257: 15192-15199.

76. Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, et al. (2002) A mouse model of human L1 retrotransposition. Nat Genet 32: 655-660.

77. Ostrowski S, Dierick HA, Bejsovec A (2002) Genetic control of cuticle formation during embryonic development of *Drosophila melanogaster*. Genetics 161: 171-182.

78. Ota A, Kusakabe T, Sugimoto Y, Takahashi M, Nakajima Y, et al. (2002) Cloning and characterization of testis-specific tektin in *Bombyx mori*. Comp Biochem Physiol B Biochem Mol Biol 133: 371-382.

79. Papanicolaou A, Gebauer-Jung S, Blaxter ML, Owen McMillan W, Jiggins CD (2008) ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Res 36: D582-587.

80. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, et al. (2003) Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science 299: 697-700.

81. Paulsson G, Hoog C, Bernholm K, Wieslander L (1992) Balbiani ring 1 gene in *Chironomus tentans*. Sequence organization and dynamics of a coding minisatellite. J Mol Biol 225: 349-361.

82. Pawlak A, Toussaint C, Levy I, Bulle F, Poyard M, et al. (1995) Characterization of a large population of mRNAs from human testis. Genomics 26: 151-158.

83. Peakall R, Smouse PE (2006) genalex 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes 6: 288-295.

84. Pertea G, Huang X, Liang F, Antonescu V, Sultana R, et al. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19: 651-652.

85. Prak ET, Dodson AW, Farkash EA, Kazazian HH, Jr. (2003) Tracking an embryonic L1 retrotransposition event. Proc Natl Acad Sci U S A 100: 1832-1837.

86. Prasad MD, Muthulakshmi M, Arunkumar KP, Madhu M, Sreenu VB, et al. (2005) SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. Nucleic Acids Res 33: D403-406.

87. Prasad MD, Nurminsky DL, Nagaraju J (2002) Characterization and molecular phylogenetic analysis of mariner elements from wild and domesticated species of silkmoths. Mol Phylogenet Evol 25: 210-217.

88. Rajkhowa R (2000) Structure property correlation of non-mulberry and mulberry silk fibers. Int J Wild Silkmoths Silk 5: 287-297.

89. Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL (2003) Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science 300: 1742-1745.

90. Rastogi S, Liberles DA (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. BMC Evol Biol 5: 28.

91. Raymond CS, Shamu CE, Shen MM, Seifert KJ, Hirsch B, et al. (1998) Evidence for evolutionary conservation of sex-determining genes. Nature 391: 691-695.

92. Raymond M, Rousset F (1995) GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J Heredity 86: 248-249.

93. Rebers JE, Riddiford LM (1988) Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol 203: 411-423.

94. Rebers JE, Willis JH (2001) A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol 31: 1083-1093.

95. Reinke V, Gil IS, Ward S, Kazmer K (2004) Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. Development 131: 311-323.

96. Reinke V, Smith HE, Nance J, Wang J, Van Doren C, et al. (2000) A global profile of germline gene expression in *C. elegans*. Mol Cell 6: 605-616.

97. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.

98. Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735-742.

99. Rogers DW, Carr M, Pomiankowski A (2003) Male genes: X-pelled or X-cluded? Bioessays 25: 739-741.

100. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365-386.

101. Saifi GM, Chandra HS (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. Proc Biol Sci 266: 203-209.

102. Sassone-Corsi P (2002) Unique chromatin remodeling and transcriptional regulation in spermatogenesis. Science 296: 2176-2178.

103. Schultz N, Hamra FK, Garbers DL (2003) A multitude of genes expressed solely in meiotic or postmeiotic spermatogenic cells offers a myriad of contraceptive targets. Proc Natl Acad Sci U S A 100: 12201-12206.

104. Sehnal F, Zurovec M (2004) Construction of silk fiber core in lepidoptera. Biomacromolecules 5: 666-674.

105. Sezutsu H, Yukuhiro K (2000) Dynamic rearrangement within the *Antheraea pernyi* silk fibroin gene is associated with four types of repetitive units. J Mol Evol 51: 329-338.

106. Sheppard PM (1961) Some contributions to population genetics resulting from the study of the Lepidoptera. Adv Genet 10: 165-216.

107. Siegal ML, Baker BS (2005) Functional conservation and divergence of intersex, a gene required for female differentiation in *Drosophila melanogaster*. Dev Genes Evol 215: 1-12.

108. Sperling FAH (1994) Sex-linked genes and species differences in Lepidoptera. Can Ent 126: 807-818.

109. Stehr G (1959) Hemolymph Polymorphism in a Moth and the Nature of Sex-Controlled Inheritance. Evolution 13: 537-560.

110. Storchova R, Divina P (2006) Nonrandom representation of sex-biased genes on chicken Z chromosome. J Mol Evol 63: 676-681.

111. Suzuki MG, Shimada T, 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 (Pt 3): 275-283.

112. Suzuki MG, Shimada T, Kobayashi M (1999) Bm kettin, homologue of the *Drosophila* kettin gene, is located on the Z chromosome in *Bombyx mori* and is not dosage compensated. Heredity 82 (Pt 2): 170-179.

113. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, et al. (2002) Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110: 327-338.

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

115. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

116. Traut W, Marec F (1996) Sex chromatin in lepidoptera. Q Rev Biol 71: 239-256.

117. Vallender EJ, Lahn BT (2004) How mammalian sex chromosomes acquired their peculiar gene content. Bioessays 26: 159-169.

118. van Oosterhout C, Hutchison WF, Wills DPM, Shipley P (2004) MICRO-CHECKER: software for identifying and correcting genotyping errors in microsatellite data Molecular Ecology Notes 4: 535-538.

119. Varshney RK, Graner A, Sorrells ME (2005) Genic microsatellite markers in plants: features and applications. Trends in Biotechnology 23: 48-55.

120. Venables JP (2002) Alternative splicing in the testes. Curr Opin Genet Dev 12: 615-619.

121. Wahlund S (1928) Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas 11: 65-106.

122. Walker WH, Delfino FJ, Habener JF (1999) RNA processing and the control of spermatogenesis. Front Horm Res 25: 34-58.

123. Wang J, Xia Q, He X, Dai M, Ruan J, et al. (2005) SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res 33: D399-402.

124. Wang PJ (2004) X chromosomes, retrogenes and their role in male reproduction. Trends Endocrinol Metab 15: 79-83.

125. Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. Nat Genet 27: 422-426.

126. Willis JH (1996) Metamorphosis of the cuticle, its proteins, and their genes; Gilbert LI, Tata JR, Atkinson BG, editors. San Diego: Academic Press. 253-282 p.

127. Xia Q, Cheng D, Duan J, Wang G, Cheng T, et al. (2007) Microarray-based gene expression profiles in multiple tissues of the domesticated silkworm, *Bombyx mori*. Genome Biol 8: R162.

128. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, et al. (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306: 1937-1940.

129. Zarkower D (2001) Establishing sexual dimorphism: conservation amidst diversity? Nat Rev Genet 2: 175-185.

130. Zhou P, Li G, Shao Z, Pan X, Yu T (2001) Structure of *Bombyx mori* silk fibroin based on the DFT chemical shift calculation. J Phys Chem B 105: 12469-12476.

Publications

1. **Arunkumar KP,** Kifayathullah L and Nagaraju J (2008) Microsatellite markers for the Indian golden silkmoth, *Antheraea assama* (Saturniidae: Lepidoptera). ***Molecular Ecology Resources*** (In press).
2. **Arunkumar KP,** Tomar A, Daimon T, Shimada T and Nagaraju J (2008) WildSilkbase: An EST database of wild silkmoths. ***BMC Genomics*** 9: 338. **(Highly accessed)**
3. **Arunkumar KP** and Nagaraju J (2006) Unusually long palindromes are abundant in mitochondrial control regions of insects and nematodes. ***PLoS ONE*** 1(1): e110.
4. Gandhe AS*, **Arunkumar KP***, John SH, Nagaraju J (2006) Analysis of bacteria-challenged wild silkmoth, *Antheraea mylitta* (Lepidoptera) transcriptome reveals potential immune genes. ***BMC Genomics*** *7*:184. (*equal contribution)
5. **Arunkumar KP,** Metta M and Nagaraju J (2006) Molecular phylogeny of silkmoths reveals the origin of domesticated silkmoth, *Bombyx mori* from Chinese *B. mandarina* and paternal inheritance of *Antheraea proylei* mitochondrial DNA. ***Molecular Phylogenetics and Evolution*** 40: 417-427.
6. Prasad MD, Muthulakshmi M, **Arunkumar KP,** Madhu M, Sreenu VB, Pavithra V, Bose B, Swaminathan S, Nagarajaram HA, Mita K, Shimada T and Nagaraju J (2005) Silksatdb: a microsatellite database of silkmoth, *Bombyx mori*. ***Nucleic Acids Research*** 33: D403-D406.

Manuscripts communicated:

Arunkumar KP, Mita K and Nagaraju J (2008) Silkworm testis specific genes are enriched on Z chromosome and are evolutionarily conserved.

Arunkumar KP and Nagaraju J (2008) Large-scale gene discovery project in Indian golden silkmoth, *Antheraea assama*.

Arunkumar KP, Awasthi AK and Nagaraju J (2008) Genetic variability and population structure of Indian golden silkmoth (*Antheraea assama*) as revealed by SSR markers: need for conservation.