UNDERSTANDING MOLECULAR ASPECTS OF *Antheraea assamensis*

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of

Doctor of Philosophy

By

Hasnahana Chetia

Department of Biosciences and Bioengineering

Indian Institute of Technology Guwahati

Guwahati, Assam-781039, India

August 2019

INDIAN INSTITUTE OF TECHNOLOGY
GUWAHATI
Dept. of Biosciences and Bioengineering

DECLARATION

This is to declare that the content embodied in this thesis entitled **"Understanding molecular aspects of *Antheraea assamensis*"** is the result of investigations carried out by me under the supervision of **Prof. Utpal Bora**, and is submitted to the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam-781039, India for the award of degree of **Doctor of Philosophy in Biosciences and Bioengineering**. This work has not been submitted elsewhere for any degree or diploma of any institute or university to the best of my knowledge and belief.

In keeping with the general practice of reporting scientific investigations, due acknowledgements have been made wherever the work of other investigators are referred.

Guwahati

August, 2019

Hasnahana Chetia
Roll No- 136106032
Department of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati
Guwahati, Assam- 781039, India

INDIAN INSTITUTE OF TECHNOLOGY

GUWAHATI

Dept. of Biosciences and Bioengineering

CERTIFICATE

This is to certify that the work embodied in the thesis entitled **"Understanding molecular aspects of *Antheraea assamensis*"** is the result of the investigations carried out by **Hasnahana Chetia (Roll No- 136106032)** under my supervision in the Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati and is submitted for the award of degree of **Doctor of Philosophy in Biosciences and Bioengineering**. This work has not been submitted elsewhere for a degree.

01/08/19

Guwahati

August, 2019

Prof. Utpal Bora

Thesis Supervisor,

Department of Biosciences and Bioengineering,

Indian Institute of Technology Guwahati

Guwahati, Assam- 781039, India

ACKNOWLEDGEMENTS

As I am about to end another chapter of my career, I would like to express my gratitude towards the people who have helped me construct it into a successful one.

I am extremely grateful and indebted to my thesis supervisor ***Prof. Utpal Bora*** for introducing me to the exciting world of omics and data science. My pre-IITG era form was not a huge fan of bioinformatics and joining Prof. Bora's laboratory changed it. I thank him for providing me the opportunity to work on one of the most significant biological resources of Assam and contribute towards its knowledgebase. I also thank him for allowing me to conduct my research independently and for coaching me to develop my scientific communication and interpersonal skills. I hope that I have been able to imbibe his enthusiasm and boldness in me before diving into this ocean of scientific adventures.

I would like thank my doctoral committee members ***Prof. Ranjan Tamuli, Prof. Bosanta Ranjan Boruah, Dr. Soumen Kumar Maiti*** and ***Prof. Venkata V. Dasu*** for their valuable suggestions, motivation and scientific guidance which always helped me to make my work better.

I would also like to convey my gratitude to the ***Department of Biosciences and Bioengineering***, ***Institutional Biotech Hub at the Centre for the Environment*** and ***Param-Ishan, IIT Guwahati's high-performance computing cluster*** for providing me all the necessary facilities to pursue my research.

I sincerely acknowledge the financial support from ***Ministry of Human Resource Development (MHRD)***, Government of India for providing me fellowship as well as

Department of Biotechnology (DBT), Government of India and ***Central Silk Board*** for funding our laboratory.

I would like to express my gratitude to ***Dr. Kartik Neog*** and ***Mr. Palash Dutta*** from ***Central Muga Eri Research and Training Institute (CMER&TI) Lahdoigarh*** for their cooperation in sample collection for my Ph.D research.

My laboratory mates have been a wonderful lot with diverse personalities and interests and interacting with them has shaped me. A note of gratitude to my seniors ***Deepika, Arghya, Suradip, Sunita*** and ***Swagata*** for their guidance and companionship; to my peers ***Vimal*** and ***Kabiraj*** for their friendship, guidance and all-round support; to my juniors ***Jon, Biju, Dharitri, Manash, Adhiraj, Tinka, Pulak*** and our friendly neighbor, ***Dibakar*** for sharing responsibilities, insights and laughter on all occasions; and to ***Pragya Ma'am*** for her affection, humor and culinary prowess.

I express my deepest sense of gratitude and love to ***Maa, Deta*** and ***Bhaiti*** for their overwhelming love, support and patience. Finally, I am in lifetime indebtment of gratitude for ***Prerana*** who has been exploring science with me since teenage hood. Ph.D. journey would've been a lifeless, uneventful journey without her. Thanks for spoiling me with love, care and friendly competitions(!) and gently nudging me back on the track when the need arose.

-Hasnahana

TABLE OF CONTENTS

SYNOPSIS

SYNOPSIS

Antheraea assamensis (Helfer), also known as the “Muga silkworm” is a multivoltine, polyphagous Saturniid Lepidopteran endemic to the Brahmaputra valley of Assam and adjoining hilly areas of North-East India. It produces muga silk, a unique golden silk which is commercially significant. Apart from textiles, the core constituents of *A. assamensis* silk also have prospective applications in the field of biomedical sciences, skincare, tissue engineering and so on.

Muga silkworm is a strict herbivore (folivore) and depends upon host plants for its diet. Based on geographical distribution and extent of commercial exploitation, the muga host plants are divided into primary, secondary and tertiary. Muga silkworm is predominantly reared on primary host plants from Lauraceae for commercial needs. Som (*M. bombycina)* is a primary host plant of muga silkworm which is evergreen and available almost all around the year. It is predominantly used for commercial rearing of muga silkworm in North-East India. Muga silkworm reared upon *M. bombycina* produces golden-yellow colored cocoon. On the other hand, Mejankari (*L. citrata*), is a secondary host plant of the same insect with extremely rare usage for commercial rearing. It is a deciduous shrub or tree, primarily cultivated for essential oil. Muga silkworms reared on *L. citrata* leaves are known to produce creamy white cocoons.

Muga silkworm is semi-domesticated and reared in outdoor conditions. This practice makes it vulnerable to many biotic and abiotic challenges. Of the biotic challenges, the microbial pathogen-induced diseases are the most disabling in nature. Pebrine is one of the most common diseases of muga silkworm and is

reportedly caused by a microsporidian (*Nosema* sp.). It is prevalent among the late instars to the adult moth and the eggs it lays; the reason behind this is transovarial transmission, which makes this disease difficult to control and demands manual monitoring. The causal organism, microsporidia, is an obligate intracellular parasite that survives inside a host till depletion of its cellular resources and can have severe impact on the fitness and silk biosynthesis of the infected larvae.

One major hurdle impeding research on muga silkworm is the unavailability of its genomic data and associated host plants as well as pests and pathogens. Following objectives were formulated to gain a deeper understanding of *A. assamensis,* its host plants (*L. citrata* and *M. bombycina*) and its microsporidian pathogen Nosema-

i. *De novo* transcriptome of *Antheraea assamensis* (Muga silkworm)

ii. *De novo* transcriptome of two host plants of muga silkworm, namely, *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari)

iii. Transcriptomic profile of *Antheraea assamensis* with respect to host-plant and development

iv. Development of MugaSeqDB, a database on *A. assamensis* and its associated host plants

v. Comparative transportome study of Nosema, a genus of microsporidia causing pebrine in silkworms

Chapter 1 of the thesis discusses general information on *A. assamensis,* its host plants (*L. citrata* and *M. bombycina*) and its microsporidian pathogen Nosema and reviews the existing studies of similar nature.

Chapter 2 reports the *de novo* transcriptome of *A. assamensis* (muga silkworm) using high-throughput sequencing of three of its tissues (alimentary canal, silk gland and residual body) from its 5th instar larvae. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were annotated using a combination of databases- UniProt, NCBI-NR (Non-redundant), Pfam, GO, COG and KEGG. Analysis of the resultant transcriptome lead to identification of differentially expressed candidate genes involved in silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. A set of candidate antimicrobial peptides of *A. assamensis* with antifungal, antibacterial, antiviral and antiparasitic potential were also identified. Finally, the transcriptome was validated by quantitative real-time PCR (qPCR) amplification of eight random candidate transcripts.

Chapter 3 reports the *de novo* transcriptomes of the two *A. assamensis* host plants, *Machilus bombycina* (som) and *Litsea citrata* (mejankari). The study identified 55,400 and 1,38,690 transcripts, respectively. ~50% transcripts in both the transcriptomes were annotated using a combination of databases (UniProt Viridiplantae, NCBI NR, Pfam, MetaCyc and GO). We also identified the putative transcripts related to plant immune system, namely, glucosinolate-myrosinase pathway (which is an herbivore defense system of plants) and antimicrobial peptides. We were able to identify homologs of almost all the enzymes which mediate glucosinolate biosynthesis and activation using long-chain aliphatic and

aromatic amino acid precursors in both the host plant transcriptomes. We were also able to identify a myriad of potential peptides with specific- and broad-spectrum antimicrobial activity, chiefly, against bacteria, fungi and virus. Our findings generated a novel resource of sequence data on these two host plants from Lauraceae family and also provided a foundation for future studies on plant defense for benefit of the sericulture industry.

Chapter 4 discusses the overall variation of the transcriptomic profile of *A. assamensis* with respect to host plant and development. We sequenced the *de novo* transcriptomes of 5th instar larvae of *A. assamensis* reared on two of its host plants, *L. citrata* and *M. bombycina*. We also sequenced the *de novo* transcriptomes of 4th instar larvae of *A. assamensis* reared on *M. bombycina.* Using the data generated in this study, we reconstructed the transcriptome for *A. assamensis,* identified the top most expressed transcripts in each tissue and observed how biological processes associated with each tissue varies with respect to host plant and larval development (4th instar and 5th instar). We found that translation was the most unanimous process expressed in each tissue of silk gland of *A. assamensis* regardless of the developmental stage or host plant. Other than this process, other processes like oxidative stress management, redox homeostasis, transcriptional regulation, etc. had variable representation across different stages. Analysis of these patterns showed how the transcriptional profile of *A. assamensis* can vary in different anatomical sections and variation in host plants.

Chapter 5 discusses the construction of MugaSeqDB, a database on *A. assamensis* and its associated host plants. MugaSeqDB is a comprehensive,

freely accessible database hosting the transcriptome data of muga silkworm (*A. assamensis)* and its two host plants, Som *(M. bombycina)* and Mejankari (*L. citrata)*. This database also hosts transcripts, predicted proteins, their respective functional and ontological annotations for these three species. Additionally, it provides secondary information on pest, pathogen and patents of the muga silkworm and its host plants. A combination of MySQL and phpMyAdmin was utilized to develop its back-end while the front end was created using a combination of HTML, php and java scripts. The complete architecture was hosted at a Linux-based commercial server. Features like search, browse, download, secondary database cross-linking, patent information, informative help pages and scope for user data submission has been incorporated in MugaSeqDB. The ultimate goal of this database will be to perform as a one-stop database for information on muga silkworm (*A. assamensis*) and other species associated with it. This database is now available online as an open-access resource at http://mugaseqdb.in.

Chapter 6 reports a comparative transportome study of four species of Nosema, a genus which causes pebrine in silkworms and honeybees. We predicted the putative transportomes of four Nosema species, viz. *Nosema apis*, *Nosema bombycis*, *Nosema ceranae* and *Nosema antheraea.* Our results indicated that the transportomes of *N. apis, N. bombycis, N. ceranae* and *N. antheraea* have a dominant share of secondary carriers and primary active transporters. The comparatively rich and diverse transportome of *N. bombycis* indicates the role of transporters in its remarkable capability of host adaptation. We identified a set of twelve transporter families core to the *Nosema* genus with possible role in osmo-regulation, intra- and extra-cellular pH regulation, energy compensation and self-

defense mechanism. We also identified a set of ten species-specific transporter families within Nosema which may be involved in species-specific host adaptations. Both the core and species-specific transporter proteins of Nosema constituted a valuable resource that will come handy in development of inhibitor-based *Nosema* management strategies in future and thereby, help the sericulture and apiculture scenario of the industrial world.

Overall, our study was able to address the existing lacunae in muga silkworm by reporting the *de novo* transcriptomes of muga silkworm, *A. assamensis* and its two host plants, *M. bombycina* and *L. citrata* by application of RNA-Seq. These processes include silk biosynthesis, biosynthesis of allelochemicals, antimicrobial peptides, variation of overall transcriptome profile of silk glands with respect to host plant and developmental stages. The study also utilized proteome information on Nosema, a genus of pebrine microsporidia, to identify crucial transporter proteins in the species. Finally, an open-access database was crated on the information generated in this study which will be useful for the broader community of seri-researchers across the world. In summary, the outcome of the current study will provide a foundation for future studies on muga silkworm and the biotic components associated with it. This in turn will benefit the greater goal of enhancing productivity of the sericulture industry and conservation of this crucial endemic species of North-East India.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

20E: 20- hydroxyecdysone

AaFHC: *A. assamensis* Fibroin heavy chain

AC: Alimentary Canal

ASG: Anterior Silk Gland

BUSCO: Benchmarking Universal Single-Copy Orthologs

Cbp: Carotenoid binding-protein

CDD: Conserved Domain Database

CEG: Core Eukaryotic Genes

COG: Cluster of Orthologous sequences

FPKM values: Fragments Per Kilobase of transcript per Million mapped reads

GO_BP: GO Biological Process

GUP: Glycerol Uptake

Hly: Hemolysin protein

HMMTOP: Hidden Markov Model for Topology of Transmembrane Proteins

HOX protein: Homeodomain Box protein

HP: Host plant

HSP: Heat shock proteins

JH: Juvenile Hormone

KAAS: KEGG Automatic Annotation Server

MF: Methyl Farnesoate

MSG: Middle Silk Gland

NAn: *Nosema antheraea*

NAp: *Nosema apis*

NB: *Nosema bombycis*

NC: *Nosema cerenae*

PNTAa: Putative Novel Transcripts of *Antheraea assamensis*

PSG: Posterior Silk Gland

RB: Residual Body

RDBMS: Relational Database Management System

RSEM: RNA-Seq by Expectation-Maximization

SBH: Single-Directional Best Hit

SEO: Search Engine Optimization

SG: Silk Gland

SOLiD: Sequencing by Oligonucleotide Ligation and Detection

SQL: Structured Query Language

SRA: Short Read Archive

TMHMM: Hidden Markov Model for Topology Prediction

TPM: Transcript per million bases

WEGO: Web Gene Ontology Annotation

CHAPTER 1

Introduction and Review of Literature

CHAPTER 1

Introduction and Review of Literature

1.1. SILKWORMS: THE SILK-PRODUCING LEPIDOPTERANS

Lepidoptera is one of the most speciose order of insects (>150,000 species) consisting of moths and butterflies [1]. Lepidopterans are primarily phytophagous, holometabolous insects found predominantly in terrestrial habitats. Some distinct anatomical features of Lepidoptera are presence of scales (modified, flattened hairs) over their body, wings and proboscis. The holometabolan lifestyle of Lepidoptera ensures complete metamorphosis in the following order- eggs→ caterpillar or larvae→ pupa → adult moth or butterfly → eggs (Fig. 1.1). Many species of this order have had a significant impact on human societies. The positive impact has been due to their role as pollinators, natural food sources as well as producers of biomaterials of anthropogenic interest while negative impact has been on agriculture as pests [2–4]. Silkworms are one of the most popular group of Lepidopterans due to the economic benefits provided by it. As their name suggests, they produce silk in their anatomically distinct, pair of labial glands and spin silk cocoons for metamorphosis into pupa (the third stage of life cycle). Larval stage is divided into five instars where silkworm larvae molt or shed their skins at each transition and change appearance. Completion of the fifth instar is marked by termination of feedings and initiation of cocoon spinning in

preparation for pupation. Other lepidopterans, i.e., butterflies usually follow the same life cycle but via chrysalis (hardened cuticle) rather than a silk cocoon.

For centuries, sericulture has been one of the major agronomic practices in countries of Asia and Europe. India has been one of the largest producers of silk in the globe with exports to ~50 countries, earning foreign exchanges of >2000 crores per year and contributing roughly 15.5% of the global silk produce (Source- Central Silk Board). Indian sericulture industry also provides employment to ~8 million rural as well as semi-urban people (Source- Central Silk Board). There are five varieties of commercial silk produced in India- mulberry silk (*Bombyx mori),* muga silk (*Antheraea assamensis),* eri silk (*Samia cynthia ricini),* tasar or tusser (*A. mylitta)* and oak tasar (*A. pernyi).* Commercial silk is produced by silkworms of two lepidoptera families- Bombycidae and Saturniidae. Of the five moths described above, *B. mori* belongs to Bombycidae family while the remaining belong to Saturniidae family and are collectively called non-mulberry silk.

1.1.A Silk and its biosynthesis in silkworms

Silk is a secretory material produced in specialized labial glands of silkworms called silk glands. These secretions are stored within ectodermal cells of the silk gland in the form of hydrated jelly and are polymerized into water-insoluble filaments as they are spun into cocoons into the external environment [5]. Silk or silk-like materials are also produced by other Arthropod taxa like Arachnida, Myriapoda, Hexapoda, etc. and is speculated to have evolved via more than one independent occasions[3]. Labial glands typically produce saliva and their initial development starts with embryogenesis itself; as the silkworm grows, they also

grow larger via polyploidization, often reaching considerably large sizes (20-40% of the body weight).

Origin of silk produced by silkworm larvae are speculated to have evolved from the last common ancestor of Lepidoptera and Trichoptera (aquatic insect order that synthesizes silk as well) about 250 million years ago[3]. Other than transition, other Lepidoptera sometimes use silk as a domicile or a girdle-like support to suspend from during molting. The silk fiber is a highly organized structure derived from multiple proteins synthesized in different anatomical sections of the silk gland which is basically a pair of labial glands [3,6]. The fiber has two filaments (one from each gland) which is composed of polymers of heavy chain and light chain fibroin as well as p25 glycoprotein or fibrohexamerin. Bombycidae silkworms have all these three components, but Saturniidae lacks p25 [7]. Fibroin chains and p25 are produced in the posterior silk gland (PSG). These proteins are further engulfed by sericin protein produced in the middle silk gland (MSG) and polymerized into a fiber during its movement through anterior silk gland (ASG) and spinneret for cocoon construction. Other than the two genes for Fibroin, *B. mori* reportedly harbors up to five sericin genes [3,8].

In Bombycidae and Saturniidae larvae, production of silk is relatively low in the early instars followed by drastically greater production during the commencement of cocoon spinning. The use of cocoon as a thermally regulated, durable and strong metamorphosis chamber is ensured by the presence of crystalline motifs like poly(Ala) or GlyAla repeat motifs in the fibroin sequence [9].

1.2. *Antheraea assamensis* (MUGA SILKWORM) AND ITS SILK

A. assamensis (Helfer), also known as the "Muga silkworm" is a multivoltine, polyphagous Saturniid Lepidopteran endemic to the Brahmaputra valley of Assam and adjoining hilly areas of Northeast India (Fig. 1.1) [10]. It is the sole producer of globally acclaimed "Muga silk", a unique lustrous golden yellow fabric and thus, contributes hugely towards of the Indian sericulture industry. It has an average life span of ~50 days and is reared five times a year during late winter (Jarua), early spring (Chatua), spring (Jethua), early summer (Aherua) and late summer or early winter (Kotia) (Fig. 1.1) [11].

The historical roots of muga silk can be traced back to the mention of the "Pitambara vastra" adorned by Lord Krishna which is strongly believed to have been woven from muga silk[12]. The kings of Ahom dynasty which ruled Assam during the 12th century were also appreciative and encouraging of muga silk rearing practices to the extent that they provided an honorary title of *Mugachungia* to prominent rearers [11]. On the initiative of Assam Science Technology & Environment Council (ASTEC), muga silk was awarded a geographical indication tag and logo in 2007 (Government of India Geographical Indications Journal No.82, 2016). Overall, muga silk contributes towards the greater Indian sericulture industry which involves ~8 million people and earns ~2000-2500 crore every year (http://texmin.nic.in/sites/default/files/note-on-sericulture2017-18-ThirdQtr.pdf).

The core constituents of *A. assamensis* silk also have prospective applications in the field of biomedical sciences as its silk has novel characteristics[13]. Muga fibroin has been demonstrated as a biocompatible, biomimetic product previously

[14,15]. There has been relatively less studies on muga sericin as its full genetic information is not known. Despite this, muga sericin, which glues fibroin (heavy and light chains) can prove to be a promising biomaterial due to the properties it shares with its other counterparts like *B. mori* sericin which has found multiple applications in the field of skincare, tissue engineering and so on [16]. Another aspect that adds to the commercial demand of Muga silk is its lustrous golden colour, whose genetic basis is not known at the moment.

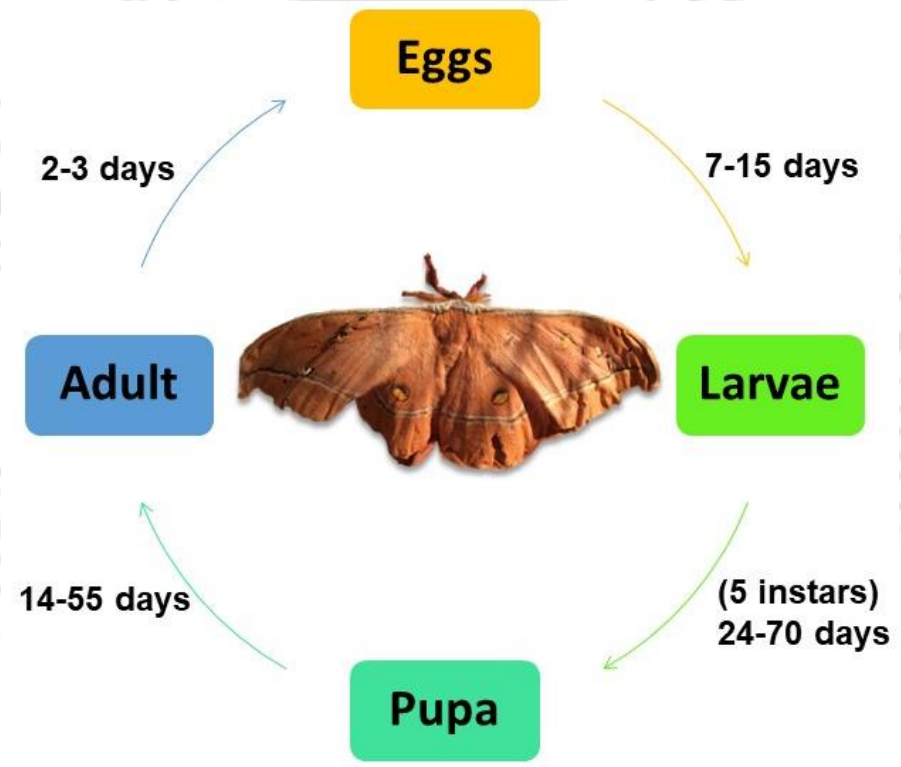


Fig. 1.1 Life cycle of muga silk moth (*A. assamensis*). The various life stages are depicted with approximate number of days required for completion of that stage. The variation in number of days per stage is due to seasonal variations with winter requiring longer duration for cycle completion.

1.3 IMPORTANT BIOTIC COMPONENTS OF *A. assamensis'* ECOSYSTEM

1.3.A Host plants

Muga silkworm is a strict herbivore (folivore) and depends upon host plants for its diet. Based on geographical distribution and extent of commercial exploitation, the muga host plants are divided into primary, secondary and tertiary. Muga silkworm is predominantly reared on primary host plants from Lauraceae for commercial needs but ~20 additional plants from Lauraceae, Magnoliaceae, Rutaceae, etc. have been reported as its tertiary host plants [17]. However, availability of experimental data on rearing on this silkworm on the non-Lauraceae host plants is scarce or non-existent.

Som (*M. bombycina)* is a primary host plant of muga silkworm which is evergreen and available almost all around the year. It is predominantly used for commercial rearing of muga silkworm in North-East India. Muga silkworm reared upon *M. bombycina* produces golden-yellow colored cocoon [17]. The other primary host plant of muga silkworm is Soalu (*Litsea polyantha* or *L. monopetala*). Muga larvae reared on Soalu are believed to be healthier than those reared on Som [18]. Mejankari (*L. citrata*), on the other hand, is a secondary host plant of the same insect, however, its usage for commercial rearing is extremely rare. It is a deciduous shrub or tree, primarily cultivated for essential oils [10,19]. Muga silkworms reared on *L. citrata* leaves are known to produce creamy white cocoons [17]. Common folklore and historical accounts suggest that Mejankari silk were rare due to the difficulties associated with breeding of *L. citrata* plants and low yield of cocoons. Thus, due to limited availability, Mejankari silk was

restricted for royal usage only, while muga silk were worn by elites and bureaucrats of ancient Assam [20].

1.3.B Pathogenic challenges

Muga silkworm is semi-domesticated and is reared in outdoor conditions. This cultural practice makes it vulnerable to many biotic and abiotic challenges. Of the biotic challenges, the microbial pathogen-induced diseases are the most devastating in nature. Some of the major diseases in muga silkworm are pebrine, flacherie, grasserie and muscardine [21]. Pebrine is one of the most common diseases of muga silkworm and is reportedly caused by a microsporidian (*Nosema* sp.) [22]. It is prevalent among the late instars to the adult moth and the eggs it lays; the reason behind this is transovarial transmission, which makes this disease difficult to control and demands manual monitoring. The causal organism, microsporidia, is an obligate intracellular parasite that survives inside a host till depletion of the host's cellular resources and can severely impact the fitness and reduce silk biosynthesis in the infected larvae. Pebrine disease in also prevalent in other silkworms and usually, prophylactic measures are used to control the disease [23]. Muscardine is another lethal disease of fungal origin. Its causal agent is *Beauveria bassiana* and the disease is more prevalent in rainy seasons as the environmental humidity and temperature is favorable for this pathogen. Flacherie are caused by viral and bacterial pathogens all around the year in muga silkworm, but especially prevalent in rainy season. In this disease, the silkworm larvae turn lethargic, have black hemolymph and lethal diarrhea. Viral flacherie is caused by viruses from Iflaviridae, Parvoviridae and Reoviridae families while common bacterial agents are *Serratia* sp., *Streptococcus sp.* and

Staphylococcus sp. Grasserie is a nuclear polyhedrosis virus induced lethal disease of muga silkworm that prevails all around the year. The disease causes swelling of the silkworm induced by disintegration of hemolymph and other bodily tissues marking the fatality of the disease. All these diseases are usually managed by- (i) manual monitoring of the eggs used for rearing, (ii) host plant disinfection as well as (iii) other prophylactic measures [21].

1.4 CURRENT STATUS OF GENOMICS OF *A. assamensis* AND ITS CONTEMPORARY SILKWORMS

B. mori is not just a model organism for Lepidoptera but also considered as a popular model organism from Insecta after Drosophila. Two primary reasons behind this are its domesticated status and ease of rearing due to dependence on a single host plant. Existence of a fully sequenced genome and transcriptome resource also Another reason that hugely contributes towards this status [8,24]. Presence of whole genome has facilitated a myriad range of genomic studies on the species ranging from production of fluorescent silk production to CRISPR-Cas based genome editing for disease targeting in silkworms [25,26]

Both mulberry and muga silk has a common basis of human interest, their silk. However, the growth of their scientific knowledgebase has not been at par. Despite the existing commercial necessity or potentiality of its biomaterials, there has been a dearth of studies addressing genetics and genomics of *A. assamensis*. The existing studies on muga silkworm are mostly focused on rearing practices and their improvement [27,28]. The vast majority of the studies are also focused on usage of muga silk as a biomaterial for various tissue engineering applications [15,29]. DNA barcode sequences of a few selected

mitochondrial genes constituted almost the entirety of genetic resources available on the species till 2017.

An efficient way to address the lack of genomic information is the application of next-generation sequencing (NGS). NGS provides comprehensive information of entire genomes and transcriptomes. It can be used for gene discovery, gene expression profiling, investigation of insect-environment interactions, evolutionary studies and much more [30]. NGS was applied to sequence, assemble and report the whole mitochondrial genome of *A. assamensis* in 2017 by our laboratory [31]. The study uncovered the complete gene sequences of 37 genes present in the mitochondrial genome of which 13 protein coding genes were crucial. The study was able to use this mitogenomic information to establish the phylogenetic position of *A. assamensis* among the Bombycoidae superfamily. Another noteworthy study on the species was the report of whole sequence of Fibroin-H gene for the first time [7]. Whole genome sequencing (WGS) can provide even greater support to studies of such nature, however, the sequencing cost and computational resource necessity for WGS is enormous. Hence, scientists often address the molecular data scarcity on non-model organisms by *de novo* transcriptome sequencing [32,33]. A transcriptome ideally represents the complete set of RNAs transcribed in any organism. Whole transcriptome constitutes an essential genomic resource to facilitate future studies on lesser-studied species.

1.5 TRANSCRIPTOMICS AS A TOOL FOR ORGANISMAL STUDIES

Transcriptomics is a sequencing-based approach to facilitate functional genomics in an organism. mRNA usually serves as the desirable molecule for this technology and it represent the complete information content that is being

expressed at any particular condition or time. Ideally, a transcriptome can capture a snapshot of the total transcripts present in a cell and reveal details about the organism with an unprecedented level of sensitivity and accuracy [34,35].

Earlier, transcriptomics was performed using microarray while nowadays RNA sequencing (RNA-Seq) is a more common approach as it can achieve a base pair level resolution and enables scientists of *de novo* annotation [34]. Reads generated by sequences like Illumina, SOLiD etc. are short in nature (100-500 bps), so efficient computational algorithms and existing knowledgebase in the form of databases are combinatorially applied to provide meaningful output from RNAseq. Currently, transcriptomes can be assembled either using a pre-existing genome sequence (reference-based assembly) or *de novo*. Here, we discuss *de novo* assembly which was adopted for this study.

De novo assembly uses the redundancy of short reads to find overlaps between the reads, commonly using De-Bruijn graph-based approaches and assemble these reads into transcripts [34,36]. Assembling transcriptomes of prokaryotes are relatively easier than eukaryotes due to larger dataset, presence of the splicing machinery and abundance of isoforms for many genes. One common method of such assembly called Trinity have addressed these issues by parallelization of the whole assembly process [36]. It first uses a “greedy algorithm” to assemble unique sequences from the reads followed by pooling the ones that overlap and finally creates an independent De Bruijn graph for each set of sequences and assembles isoforms within the group.

De novo transcriptomics have several advantages such as providing an initial set of transcripts for expression analysis, recovery of sequences missing in

genomes, detection of transcripts from exogenous sources, prediction of novel splice sites and chromosomal rearrangement [34]. Among disadvantages of *de novo* assembly are the needs for a greater sequencing depth, overwhelming need of computational resources and prediction of chimeric reads. The first two issues are usually addressed by using sequencing approaches that provide more depth and large computational clusters which can run parallelized assembly programs. The chimeric reads can be removed using stringent quality control and annotation approaches for transcript identification.

In the current study, we have applied NGS for transcriptomics of *A. assamensis* and its host plants to uncover certain molecular processes occurring in these species. These processes include biosynthesis of allelochemicals and their corresponding detoxification systems in muga silkworm, variation of muga silk genes with respect to host plant and developmental stages as well as identification of antimicrobial peptides in the three species. We have also utilized genome sequencing-derived proteome information on a silkworm pathogen, microsporidia, to identify crucial transporter proteins in the species.

Based on our literature review, the following objectives were formulated to gain a deeper understanding of *A. assamensis,* its host plants (*L. citrata* and *M. bombycina*) and its microsporidian pathogen Nosema-

i. *De novo* transcriptome of *Antheraea assamensis* (Muga silkworm)

ii. *De novo* transcriptome of two host plants of muga silkworm, namely, *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari)

iii. Transcriptomic profile of *Antheraea assamensis* with respect to host-plant and development

iv. Development of MugaSeqDB, a database on *A. assamensis* and its associated host plants

v. Comparative transportome study of Nosema, a genus of microsporidia causing pebrine in silkworms

REFERENCE

1. Mitter C, Davis DR, Cummings MP. Phylogeny and Evolution of Lepidoptera. *Annu Rev Entomol*. 2017;62(1):265-283. doi:10.1146/annurev-ento-031616-035125

2. van Eldijk TJB, Wappler T, Strother PK, et al. A Triassic-Jurassic window into the evolution of Lepidoptera. *Sci Adv*. 2018;4(1):e1701568. doi:10.1126/sciadv.1701568

3. Sehnal F, Sutherland T. Silks produced by insect labial glands. *Prion*. 2008;2(4):145-153. http://www.ncbi.nlm.nih.gov/pubmed/19221523.

4. Zhong H, Li F, Chen J, Zhang J, Li F. Comparative transcriptome analysis reveals host-associated differentiation in Chilo suppressalis (Lepidoptera: Crambidae). *Sci Rep*. 2017;7(1):13778. doi:10.1038/s41598-017-14137-x

5. Craig CL. Evolution of arthropod silks. *Annu Rev Entomol*. 1997;42(1):231-267. doi:10.1146/annurev.ento.42.1.231

6. Merritt DJ, Hayashi CY, Weisman S, Sutherland TD, Young JH. Insect Silk: One Name, Many Materials. *Annu Rev Entomol*. 2009;55(1):171-188. doi:10.1146/annurev-ento-112408-085401

7. Gupta K A, Mita K, Arunkumar KP, et al. Molecular architecture of silk fibroin of Indian golden silkmoth, *Antheraea assama*. *Sci Rep*. 2015;5(1):12706. doi:10.1038/srep12706

8. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori. Insect Biochem Mol*

Biol. 2008;38(12):1036-1045. doi:10.1016/J.IBMB.2008.11.004

9. Chen F, Porter D, Vollrath F. Structure and physical properties of silkworm cocoons. *J R Soc Interface*. 2012;9(74):2299-2308. doi:10.1098/rsif.2011.0887

10. Tikader A, Vijayan K, Saratchandra B. Muga silkworm, *Antheraea assamensis* (Lepidoptera: Saturniidae) - an overview of distribution, biology and breeding. *Eur J Entomol*. 2013;110(2):293-300. doi:10.14411/eje.2013.096

11. S K Borthakur. Ethnobiological wisdom behind the traditional muga silk. *Indian J Tradional Knowl* . 2003;2(1):230-235. https://www.niscair.res.in/sciencecommunication/researchjournals/rejour/ijtk/Fulltextsearch/2003/July 2003/IJTK-Vol 2(3)-July 2003-pp 230-235.htm.

12. Das S, Bora U, Borthakur BB. Silk Biomaterials for Tissue Engineering and Regenerative Medicine. Elsevier; 2014. doi:10.1533/9780857097064.1.41

13. Kundu S. Silk Biomaterials for Tissue Engineering and Regenerative Medicine. Woodhead Publishing; 2014.

14. Kasoju N, Bhonde RR, Bora U. Preparation and characterization of *Antheraea assama* silk fibroin based novel non-woven scaffold for tissue engineering applications. *J Tissue Eng Regen Med*. 2009;3(7):539-552. doi:10.1002/term.196

15. Kasoju N, Bora U. *Antheraea assama* silk fibroin-based functional scaffold with enhanced blood compatibility for tissue engineering applications. *Adv*

Eng Mater. 2010;12(5):B139-B147. doi:10.1002/adem.200980055

16. Padamwar MN, Pawar AP. Silk Sericin and Its Applications: A Review. Vol 63.; 2004. https://pdfs.semanticscholar.org/14dc/8c69431c2af83f3682b519a04321390bbb85.pdf.

17. Bindroo BB, Singh NT, Sahu AK, Chakravorty R. Muga silkworm host plants. *Indian Silk*. 2006;44:13-17. http://mugadbase.com/pdf/Bindroo_et_al_2006.pdf.

18. Mazumdar-Leighton S*;Rabha's Weave;* http://www.mugadbase.com/pdf/Rabha%27s_weave.pdf. Accessed July 30, 2019.

19. Han X-J, Wang Y-D, Chen Y-C, Lin L-Y, Wu Q-K. Transcriptome Sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba.* Schönbach C, ed. *PLoS One*. 2013;8(10):e76890. doi:10.1371/journal.pone.0076890

20. William Robinson. A Descriptive Account of Asam: With a Sketch of the Local Geography - William Robinson (of Gowhatti Government Seminary) - *Google Books*. Ostell and Lepage; 1841. https://books.google.co.in/books?hl=en&lr=&id=LoNFAQAAMAAJ&oi=fnd&pg=PA28&dq=%22muga%22+AND+%22silk%22&ots=6RMrxhgNmy&sig=ir3YtBbYnM7XsdR9HBCWcjWQAPQ#v=onepage&q=muga silk&f=false.

21. Problems of muga culture. https://shodhganga.inflibnet.ac.in/bitstream/10603/69589/14/14_chapter 5.pdf. Accessed July 30, 2019.

22. Talukdar JN. Prevalence of transovarian infection of a microsporidian

parasite infecting muga silkworm, *Antheraea assamensis*. *J Invertebr Pathol*. 1980;36(2):273-275. https://www.cabdirect.org /cabdirect/abstract/19800579474. Accessed July 30, 2019.

23. Ishihara R, Fujiwara T. The spread of pebrine within a colony of the silkworm, *Bombyx mori* (Linnaeus). *J Invertebr Pathol*. 1965;7(2):126-131. doi:10.1016/0022-2011(65)90023-6

24. Li Y, Wang G, Tian J, et al. Transcriptome analysis of the silkworm (*bombyx mori*) by high-throughput rna sequencing. Gibas C, ed. *PLoS One*. 2012;7(8):e43713. doi:10.1371/journal.pone.0043713

25. Dong Z, Dong F, Yu X, et al. Excision of nucleopolyhedrovirus form transgenic silkworm using the crispr/cas9 system. *Front Microbiol*. 2018;9:209. doi:10.3389/fmicb.2018.00209

26. Tatemastu K, Sezutsu H, Tamura T. Utilization of transgenic silkworms for recombinant protein production. *J Biotechnol Biomater*. 2012;s9(01):1-9. doi:10.4172/2155-952X.S9-004

27. Dutta BM. Economics of muga rearing. *Glob J Res Manag*. 2013;3(1):32-45. http://www.i-scholar.in/index.php/gjrm/article/view/40499. Accessed December 10, 2015.

28. Barman H, Rana B. Early stage indoor tray rearing of muga silkworm (*Antheraea assamensis* helfer) – a comparative study in respect of larval characters. _*Mun Ent Zool*. 2011;6(1).

29. Kar S, Talukdar S, Pal S, Nayak S, Paranjape P, Kundu SC. Silk gland

fibroin from indian muga silkworm *Antheraea assama* as potential biomaterial. *Tissue Eng Regen Med*. 2013;10(4):200-210. doi:10.1007/s13770-012-0008-6

30. Dheilly NM, Adema C, Raftos DA, Gourbal B, Grunau C, Du Pasquier L. No more non-model species: The promise of next generation sequencing for comparative immunology. *Dev Comp Immunol*. 2014;45(1):56-66. doi:10.1016/j.dci.2014.01.022

31. Singh D, Kabiraj D, Sharma P, et al. The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. *PLoS One*. 2017;12(11). doi:10.1371/journal.pone.0188077

32. Birol I, Behsaz B, Hammond SA, Kucuk E, Veldhoen N, Helbing CC. De novo transcriptome assemblies of *Rana (lithobates) catesbeiana* and *Xenopus laevis* tadpole livers for comparative genomics without reference genomes. *PLoS One*. 2015;10(6):e0130720. doi:10.1371/journal.pone.0130720

33. Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*. 2010;11(1):400. doi:10.1186/1471-2164-11-400

34. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet*. 2011;12(10):671-682. doi:10.1038/nrg3068

35. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics

technologies. *PLOS Comput Biol*. 2017;13(5):e1005457. doi:10.1371/journal.pcbi.1005457

36. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644-652. doi:10.1038/nbt.1883

CHAPTER 2

De novo transcriptome of *Antheraea assamensis* (Muga silkworm)

CHAPTER 2

De novo transcriptome of *Antheraea assamensis* (muga silkworm)

This chapter has been published as:

Chetia H, Kabiraj D, Singh D, et al. *De novo* transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). Gene. 2017; 611. doi:10.1016/j.gene.2017.02.021

ABSTRACT

Antheraea assamensis (Lepidoptera: Saturniidae), is a semi-domesticated silkworm known to be endemic to Assam and the adjoining hilly areas of Northeast India. It is the only producer of a unique, commercially important variety of golden silk called “muga silk”. Herein, we report the *de novo* transcriptome of *A. assamensis* reared on *Machilus bombycina* (Som) leaves for the first time. Reads generated by high throughput sequencing of cDNA libraries from multiple tissues, viz. alimentary canal, silk gland and residual body of the 5th instar of muga silkworm were assembled into transcripts via a *de novo* transcriptome assembly pipeline followed by functional annotation and classification. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were either allocated a functional annotation or categorized under Pfam/COG/KEGG categories. Identification of differentially expressed transcripts and their comparative sequence analysis revealed candidate genes related to silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. Several candidate anti-microbial peptides which may have potential anti-bacterial,

anti-fungal or anti-parasitic activity in A. assamensis were also identified. Transcriptome validation was carried out by quantitative real-time PCR (qPCR) amplification of eight transcripts. The resources generated by this study will expand the periphery of existing genomic data on *A. assamensis* facilitating future in-depth studies on its unknown aspects.

2.1 INTRODUCTION

Silk is an important cultural and commercial fibre largely obtained from silkworms. Some silkworms have been domesticated by humans over a period of time to exploit their potential in textiles. Still, majority of them remain semi-domesticated or wild. The silk proteins- fibroin and sericin of domesticated mulberry silkworm, *Bombyx mori* (Family: Bombycidae) has been extensively used for tissue engineering applications. Research on these has been further accelerated by discovery of its complete genome and transcriptome [1,2]. Similar as well as novel research applications can also be expected from biomaterials of other less-studied semi- or undomesticated silkworms. The dearth of information and hindrances in domestication of these silkworms currently restricts their usage in such applications.

One such semi-domesticated silkworm is *Antheraea assamensis* (Helfer), also known as the “muga silkworm”. *A. assamensis* (n=15) is a multivoltine, polyphagous silkworm (Lepidoptera:Saturniidae). It is mostly endemic to the Brahmaputra valley of Assam and adjoining hilly areas of Northeast India [3]. It is the sole producer of globally acclaimed “muga silk”, a unique lustrous golden yellow fabric and thus, contributes hugely towards the Indian sericulture industry.

Muga silk has a geographical indication tag and logo since 2007 (Government of India Geographical Indications Journal No.82, 2016).

Apart from textiles, the core constituents of *A. assamensis* silk also have prospective applications in the field of biomedical sciences as its silk has novel characteristics in comparison to *B. mori* silk [4]. The potential of one of the two core structural protein components of silkworm, namely, fibroin has been demonstrated as a biocompatible, biomimetic product previously [5,6]. The other component, sericin, which glues fibroin (heavy and light chains) also shows promises as a biomaterial. This is apparent from *B. mori* sericin which has found multiple applications in the field of skincare, tissue engineering and so on [7]. Both of these proteins are exclusively produced and processed in anatomically distinct glands named the silk glands. The silk gland transcriptome of *A. assamensis* can help us in understanding the molecular components of silk synthesis and regulation in *A. assamensis* better.

Another aspect that adds to the commercial demand of muga silk is its lustrous golden colour. In *B. mori*, co-relation studies between diet and cocoon colour in *B. mori* has shown that exogenous pigments like carotenoids and flavonoids are absorbed from dietary mulberry leaves, transported via carotenoid-binding protein in alimentary canal to the silk gland and are responsible for its cocoon colour [8,9]. While the role of any plant pigment in muga silk coloration has not been hinted at yet, identification of candidate pigment-binding proteins can be a resource for future studies on the same. Also, learning about the digestive profile of a silkworm gut during active feeding stage is interesting. Sequencing of the alimentary canal will facilitate answering these aspects as the above-mentioned processes predominantly occur here.

Despite the existing commercial necessity or potentiality of its biomaterials, genomic studies on *A. assamensis* are scarce and has an additional hindrance, lack of its whole genome sequence. Given the sequencing cost and enormous bioinformatic efforts involved in sequencing genomes, scientists often address the molecular data scarcity on non-model organisms by *de novo* transcriptome sequencing [10,11]. A transcriptome ideally represents the complete set of RNAs transcribed in any organism. Elucidation of silkworm transcriptome have evidently uncovered molecular information on hitherto unreported aspects of silkworm like the fibroin-H gene of muga silkworm [12]. So, given the unprecedented potential of transcriptome studies, we also aimed to elucidate a *de novo* multi-tissue transcriptome for *A. assamensis.* To this end, we carried out the sequencing and *de novo* assembly of *A. assamensis* transcriptome via high throughput RNA-Seq using three of its tissues, namely, silk gland, alimentary canal and residual body. Analysis of the assembled transcripts uncovered an array of candidate genes related to silk synthesis and innate immunity in muga silkworm and constituted an essential genomic resource to facilitate future studies on this endemic species.

Table 2.1 General information on *A. assamensis*

Common name	Muga moth
Scientific name	*Antheraea assamensis*
Host plants	• Primary: *Machilus bombycina* (Som) and *Litsea polyantha* (Soalu).[13] • Secondary: *Litsea citrata* (Mejankari), *Cinnamomum camphora* (Korpur), etc. [13] • Tertiary: *Litsaea salicifolia* (Dighloti), *Actinodaphne obovata* (Petarichawa), etc. [13]
Homotypic synonym	*Antheraea assama*

NCBI Taxonomy ID	91021
Taxonomical Classification (Order and Family)	Lepidoptera: Saturniideae
Chromosome number (n)	15
Whole Genome	Not available
Transcriptome	Available from this study [14]
Mitochondrial genome	Available (NCBI RefSeq ID- NC_030270.1) [15]

2.2 MATERIALS AND METHODS

2.2.1 Sample collection, RNA isolation, cDNA library preparation and sequencing

Fifth-instar larvae of semi-domesticated *A. assamensis* reared on *Machilus bombycina* leaves (Som) in outdoor conditions (~25 °C, 70– 80% relative humidity) were collected from Tura district of Meghalaya, India (courtesy - Central Silk Board). Three tissue samples were dissected and pooled from three silkworm larvae under sterile conditions; the first and second sample consisted of the complete silk gland or SG and the complete alimentary canal or AC, respectively. The remaining tissues including fat bodies, Malpighian tubules, etc. were pooled together to constitute the third sample termed as residual body or RB. The samples were stored in RNAlater stabilization solution (Ambion™) at −80 °C for further processing.

Total RNA from AC, SG and RB was isolated using the standard Trizol method. The concentration, purity and integrity of the isolated RNA were estimated using Nanodrop spectrophotometer and High Sensitivity Bioanalyzer Chip (Agilent

Technologies). High quality of total RNA with A260/A280 ratio ≥ 1.8 and RIN number ≥ 8 was used for polyA-mRNA enrichment followed by cDNA library preraparation using the TruSeq RNA Library Preparation Kit (Part # 15008136) to prepare a cDNA library for each tissue- AC, SG and RB. These libraries were then sequenced on Illumina HiSeq™ 2000 sequencer platform using the paired-end sequencing protocol of Illumina at Genotypic Technology, Bangalore.

2.2.2 Quality control of raw data and *de novo* assembly of transcriptome:

The resultant 101 bp paired-end raw reads were subjected to quality control measures using SeqQC (https://www.genotypic.co.in/Products/7/Seq-QC.aspx). Raw reads were trimmed of adapters, polyA sequences, duplicated reads, low-quality bases towards 3′ end and with phred quality score of ≤ 30 using Trimmomatic 0.35 [16]. The resulting raw reads were mapped to ribosomal rRNA reads from the SILVA rRNA database project using bowtie and unmapped read pairs were retained [17,18]. The resultant reads were re-examined using SeqQC to check for attainment of desirable quality features.

These filtered paired-end reads were selected for assembly pipeline. The reads from AC, SG and RB were assembled into separate transcriptomes each, using Velvet (v1.2) - Oases 0.2 pipeline for *de novo* assembly lengths [19,20]. Multiple k-mer lengths were tested for assembly and assessed based on metrics like total number of transcripts generated, total transcript length and fewer number of N's (N = number of gaps between scaffolded contigs) following which 51 (AC), 45 (SG) and 51 (RB) were selected as optimal. The quality of transcriptome assemblies was checked using parameters like N50 value, mean transcript length, percentage of reads mapped to the transcriptome, etc. using the scripts

provided under Trinity package [21]. The percentage of transcripts mapped to 248 Core eukaryotic genes (CEG) and EST set of *A. assamensis* (WildSilkBase) were also assessed using blastn program (e-value cutoff- 1-e03) [22,23].

2.2.3 Annotation and functional classification of the transcriptome:

AC, SG and RB transcriptomes were clustered at 95% identity using CD-Hit to create a set of comprehensive transcriptome [24]. These transcripts were annotated against NCBI Protein database (Insecta) and UniProtKB (Insecta) using blastx (e-value cutoff- 1e03). Pfam A database was utilized to identify functional domains within transcripts using blastx (e-value cutoff- 1e03) [25]. The transcripts were further classified into functional categories of COG (Cluster of Orthologous sequences) and KEGG (Kyoto Encyclopedia of Genes and Genomes) using blastx (e-value cutoff- 1e03) and single-directional best hit (SBH) method of the KAAS server, respectively [26,27]. The set of transcripts that remained unannotated after performing the steps described above was checked for similarity to known silkworm sequences from the following datasets- ESTs of *A. assamensis*, *S. cynthia ricini* and *A. mylitta* from WildSilkBase and "Bombyx mori Comprehensive gene set" available from Kaikobase using blastn (e-value cutoff- 1e03) [23,28]. The dissimilar transcripts were tested for homology with proteins of other taxa using blastx (e-value 0.001) against UniProt Ensemble database for plant, virus, bacteria and fungi (UniProt Consortium, 2012). The remaining unannotated transcripts with neither annotation nor classification were grouped together and named as "Putative Novel Transcripts of *Antheraea assamensis*" or PNTAa.

2.2.4 Identification of differentially expressed transcripts:

The differential expression of transcripts in AC and SG relative to RB was analyzed to identify important candidate genes which are known to be produced in tissue-specific manner. The QC-filtered raw reads were aligned to the comprehensive transcriptome to generate read counts using bowtie 2 and RSEM [18,29]. Then, R language-based tool, DeSeq (v1.14.0) tool was used to identify the differentially expressed genes (p-value cutoff 0.05) [30]. Furthermore, the Gene Ontology (GO) terminologies associated with the functional annotations of upregulated transcripts in AC and SG were retrieved using the "Retrieve ID/Mapping" tool of UniProt database. GO enrichment analysis was performed using the online tool WEGO (Web Gene Ontology Annotation Plot) which demarcates the statistically significant GO terms on the basis of its in-built Pearson chi-square test [31].

2.2.5 Quantitative reverse transcription PCR (RT-qPCR) for transcriptome validation:

Eight transcripts- six annotated and two unannotated, were randomly selected for experimental validation of the transcriptome assembly using RT-qPCR. RB was used as a control sample for the study while AC and SG were regarded as treated samples. Transcript-specific primers were designed by Primer-BLAST and aligned to the transcripts to rule out non-specific amplification prior to experiment via blastn [32]. Total RNA was extracted using standard Trizol method and converted into cDNA using Brilliant II SYBR Green cDNA synthesis Master mix (Cat# 600559, Agilent Technologies, USA) as per manufacturer's protocol. 30 ng/reaction of cDNA was used as input template concentration which was

amplified for the eight transcripts using Brilliant II SYBR Green qPCR Master mix (Cat#. 600,828, Agilent Technologies, USA) in 25 µl reaction. The cycling conditions were as follows: 95°C for 10 min and 40 cycles of 95°C for 15 secs, 60°C for 30 secs and 72°C for 30 secs. The reaction melting-curve analysis was applied to all reactions to ensure consistency and specificity of the amplified product. The qPCR products were resolved in 1.5% agarose gel and their sizes were confirmed using 100 bp DNA ladder (Cat# SM0241, GeneRuler 100 bp DNA ladder, Thermo Scientific, USA). The relative expression of the six transcripts as well as PNTAa_1 and PNTAa_2 were normalized against the internal reference gene, alpha-tubulin. The fold change values were calculated using the $2^{-\Delta\Delta Ct}$ method.

2.3 RESULTS AND DISCUSSION-

2.3.1 Transcriptome assembly of *A. assamensis*-

High-throughput sequencing of AC, SG and RB generated 70.37, 84.27 and 76.86 million paired-end raw reads were generated out of which 68.42, 80.92 and 74.83 million reads were used for transcriptome assembly post-quality control. The raw reads were deposited in the NCBI Short Read Archive (SRA) database with the following accession numbers: AC- SRX1293136, SG- SRX1293137, RB-SRX1293138. The statistics of transcriptome assembly and annotation are depicted in Table 2.2. Mean transcript length and N50 values of the transcriptomes were found to be consistent with those observed in other *de novo* transcriptome assemblies [33,34]. We also measured the overall coverage of CEGs (a set of 248 genes which are highly conserved in eukaryotes) to evaluate the quality of the assembled transcripts [22]. ~99% of the CEGs were present in AC,

SG and RB, respectively. Similarly, 69% of AC, 62% of SG and 89% of RB transcripts were found to be homologous with the *A. assamensis*-specific EST data from WildSilkBase [23]. These results further ascertained the robustness of our assemblies and our quality metrics were at par with previous sequencing studies [11,35].

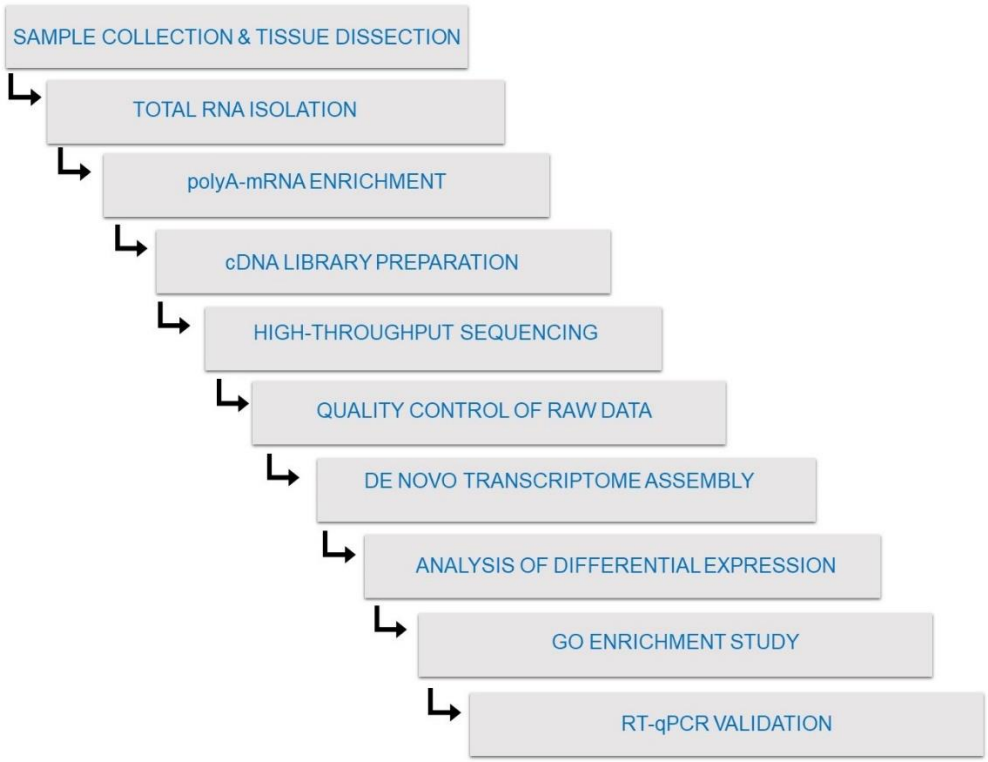


Fig. 2.1 Diagrammatic representation of the workflow followed for assembly, annotation and differential expression study of the *de novo* transcriptome of *Antheraea assamensis*. [AC - Alimentary Canal, SG - Silk Gland and RB - Residual Body]

Table 2.2 Transcriptome assembly statistics for Alimentary Canal (AC), Silk Gland (SG) and Residual Body (RB) of *Antheraea assamensis*. [Table re-used in compliance with publisher guidelines [14]]

	AC	SG	RB
k-mer length	51	45	51
Total number of transcripts	39,784	40,518	41,131
Total transcript length (in megabases or mb)	74.23	74.08	75.56
Average transcript length (in bases)	1865.9	1828.5	1837.2
Maximum transcript length (in bases)	22,903	25,783	53,354
Minimum transcript length (in bases)	200	200	200
N50 value of transcripts (in bases)	3562	3484	3451

2.3.2 Annotation and classification of the collective transcriptome of *A. assamensis*

The collective transcriptome of *A. assamensis* was used as an input for the annotation process. A combination of insect-specific (NCBI-Protein, KEGG, COG and Pfam) and silkworm-specific (WildSilkBase and KaikoBase) databases was strategically adopted so that annotation of maximum number of transcripts could be possible (Fig. 2.1). Out of 1,21,433 transcripts, we were able to annotate 87,281 (Table 2.3). The remaining transcripts were matched with Ensemble: Plants, Fungi, Bacteria and Virus proteins resulting in identification of 1302 transcripts homologous to organisms from other taxa. In totality, we were able to annotate 88,583 (74%) of the transcripts while 31,850 (26%) of the transcripts which remained unannotated were classified under the PNTAa dataset.

Table 2.3 Annotation summary of the collective transcriptome of *A. assamensis*. [Table re-used in compliance with publisher guidelines [14]]

Database	Number of annotated transcripts
NCBI "Insecta" proteins	74001
Pfam	46,415
KEGG	15,260
COG	33,015
WildSilkBase	49,436
Kaikobase	21,992
Ensemble plants#	75 (plants), 998 (bacteria), 22 (fungi), 207 (virus)
Total number of transcripts = 1,21,433	
Total number and percentage of annotated transcripts = 89.583, ~ 74%	
Total number and percentage of unannotated transcripts = 31.850, ~ 26%	

#- Ensemble databases were matched only with the unannotated transcripts.

Top hit species distribution of annotated transcripts (74,001 from NCBI Insecta) showed that *Danaus plexippus* (Lepidoptera: Nymphalidae) (15,479 transcripts), *B. mori* (Lepidoptera: Bombycidae) (12,910 transcripts), *Papilio xuthas* (Lepidoptera: Papilionidae) (1254 transcripts), *Manduca sexta* (Lepidoptera: Sphingidae) (914 transcripts) and *Tribolium castaneum* (Coleoptera: Tenebrionidae) (481 transcripts) were the top five organisms sharing homology with *A. assamensis* (Fig. 2.2A). The remaining transcripts (42,963) were, also,

found to be predominantly similar to proteins from related Lepidopteran insects like *Helicoverpa armigera, A. pernyi, Samia cynthia ricini, P. polytes* etc. To obtain an overall cue about the scenario of shared homology between *A. assamensis* and the top two similar organisms, *D. plexippus* and *B. mori*, the associated GO terms of their corresponding proteins were retrieved from UniProt. Count based analysis showed that ATP, RNA, zinc-ion, DNA binding and RNA-directed DNA polymerase activity were among the top-associated molecular functions in these three species while proteolysis, carbohydrate metabolic process and intracellular protein transport were the most commonly associated biological processes. These terms are often observed among the distribution of top GO classes in annotated Lepidopteran transcriptomes [36,37]. For example, the zinc-ion binding function is related to proteins with zinc-finger domain which form one of the most abundant group of proteins in eukaryotes. ATP or DNA/RNA binding function is vital for the smooth functioning of several fundamental processes of the cell such as transcription, translation, intracellular protein transport, etc. Overrepresentation of these set of GO terms suggests that despite undergoing genomic modifications under selective pressure and adapting to different environmental situations with variable characteristics such as food habits, few core fundamental pathways are highly conserved within the Lepidopteran order.

Presence of Pfam domains were also probed in the transcriptome. 48,418 transcripts were identified with 5896 Pfam domains and on ranking the Pfam domains by transcript count, “Zinc finger C2H2 type” (Znf) domain was observed to be predominantly present in *A. assamensis* transcripts (Fig. 2.2.C). Our previous observations that zinc-ion binding is one of the top molecular functions

conserved between *A. assamensis, D. plexxipus* and *B. mori* can be co-related with the Pfam annotation results where presence of a greater share of these domains is probably due to their broad spectrum of role as transcription factors, DNA-binding motifs, etc. in insects and other organisms [38]. Among other significantly abundant domains were “WD domain, G-beta repeat” involved in essential biological functions such as signal transduction, transcription regulation and apoptosis; “Zinc finger double domain” whose structure and functions are quite similar to Znfs; “Immunoglobulin I-set domain” found in hemolymph proteins and signalling molecules involved in immune response; and “Fibronectin type III domain” which are topologically similar to Ig-like domains and found in important insect proteins like chitinase [39,40].

KEGG and COG databases are often used to describe transcriptomes where KEGG-annotated transcripts are classified under pathway functions providing cues about the existing biological pathways while COG-annotated transcripts were classified into functional categories based on orthologous relationships. In *A. assamensis*, 15,034 of the transcripts were assigned to 316 KEGG pathways classified under 22 unique KEGG pathway functions. Based on the transcript count, “translation”, “folding, sorting and degradation”, “carbohydrate metabolism”, “signal transduction” and “transport and catabolism” were the top five KEGG molecular pathway networks (Fig 2.2B). The network associated with genetic information processing and metabolism was found to be the most abundant. Again, COG-classification classified 33,015 transcripts under a series of functional categories among which “general function prediction only”, “replication, recombination and repair”, “transcription”, “translation” and “post-transcriptional modification” were comparatively enriched in transcript count;

"general function only" category, which contains poorly categorized genes, formed the largest group of COG classified transcripts (Fig. 2.2D). Our results briefly indicated that the silkworm larvae are actively carrying out protein synthesis as well as downstream processing of the end-products. Pfam, KEGG or COG classification results were quite similar to that observed in the other transcriptomic studies of Lepidopterans like *B. mori* or *P. xylostella* [1,41]. A few homologues of bacterial, viral, fungal and plant-proteins were also observed in the transcriptome when matched with their respective UniProt Ensemble data. Similarity of *A. assamensis* transcripts with proteins of unrelated taxa suggests that these were sourced from the consumed host plant leaves and microflora associated with these leaves as well as muga silkworm gut. The complete annotation and classification results of the collective transcriptome have been combined and are currently being utilized for the construction of a database on *A. assamensis* transcriptome called MugaSeqDB (See Chapter 5).

2.3.3 Differentially expressed genes in *A. assamensis*-

The functional annotation and classification of the collective transcriptome provided us an understanding of the overall transcriptomic scenario of *A. assamensis*. However, discovery of the tissue-specific transcripts of interest, for e.g. silk synthesis related transcripts of silk gland, requires the study of differential expression of genes across different tissues. So, we analyzed differential expression of AC and SG relative to RB. 21.8% (1023 up and 1741 down) of transcripts in AC and 23.02% (348 up and 1949 down) of transcripts in SG were differentially expressed (p-value ≤ 0.05) relative to RB.

We analyzed the set of up-regulated annotated transcripts whereby we retrieved their associated GO terms from UniProt and performed an enrichment analysis via WEGO [31]. The results showed that eight biological processes and five molecular functions were significantly enriched in SG and AC (p-value ≤ 0.05) (Fig. 2.3). The biological processes commonly enriched in SG and AC were "catabolic process", "cellular metabolic process", "establishment of localization", "nitrogen compound metabolic process" and "transport". Catabolic processes refer towards chemical reactions or pathways involved in breakdown of carbon-based compounds for liberation of energy and are possibly involved in meeting the energy-requirements of up-regulated metabolic processes. The energetic stress incurred by silk glands for heightened production of silk proteins during fifth instar could also result in up-regulation of catabolic processes. Both tissues are actively producing transcripts involved in general metabolism of the cell as well as synthesis of organic or inorganic compounds containing nitrogen. Transport and localization are also active in both tissues suggesting the directional movement of secretory silk proteins or plant-derived nutrients in SG and AC, respectively.

Among the other biological processes, "pigmentation", "oxidation-reduction" and "regulation of biological processes" were significantly enriched in AC. Among the processes enriched only in AC was "biological process regulation" whose transcripts are involved in modulation of chemical reactions/pathway rates of metabolic processes. Additionally, oxidation-reduction processes which involve addition or removal of electrons from a substance were also enriched in AC. The conditions of the silkworm gut can be related with this observation; the

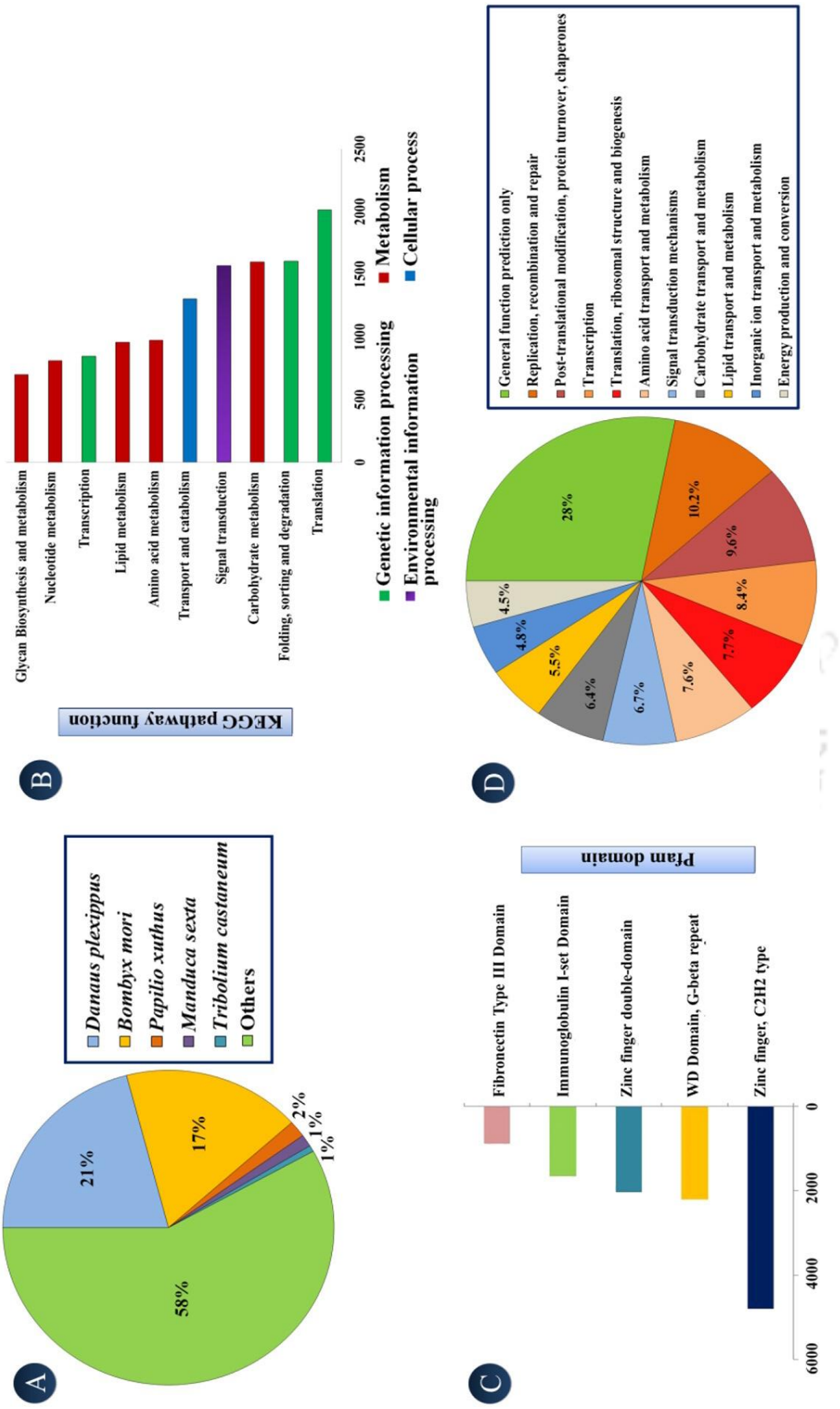


Fig. 2.2 Annotation of the collective transcriptome of *A. assamensis* [A– Top five species distribution, B- Frequency distribution of the top ten KEGG Pathway functions, C- Frequency distribution of the top five Pfam domains and D- Percentage distribution of the COG family functions within the annotated transcripts] [Figure re-use in compliance with publisher guidelines [14]

reducing/oxidizing conditions of the insect herbivores' gut are known to aid in the process of adaptation to its host plants chemical defences [42]. Since the silkwormsused in this study were in active feeding stage during sampling, the digestive condition of its gut may have promoted the up-regulation of the genes related to regulation of the gut's redox conditions. The pigmentation process, related to pigment synthesis and accumulation, was also significantly enriched in AC. Its corresponding transcripts were associated with pigment biosynthesis in insects such as anthocyanins and other related isoprenoids. However, we were not able to find protein homologues of known silk-pigmentation related genes among these transcripts.

Among the enriched molecular functions in AC and SG were “binding”, “catalytic”, “signal transducer”, “hydrolase” and “transferase” activity. Tracing back the “hydrolase binding” class in both AC and SG, majority of the up-regulated transcripts were found to be homologous to juvenile hormone epoxide hydrolase, which is involved in juvenile hormone inactivation and carboxyl ester hydrolase, whose role in silkworm silk glands is unknown, though it is speculated to play some important role in silk synthesis in major ampullate glands of black widow spiders [43]. Homologues of poly(ADP-ribose) glycohydrolase which is involved in a post-translational modification called poly(ADP-ribosyl)ation, implicated in chromatin modification process, were up-regulated only in SG. Similarly, homologue of digestive hydrolases like lipase, α-amylase, etc. were up-regulated only in AC. Among the transferases up-regulated in SG and RB, UDPglucosyltransferase homologues were dominantly expressed. These enzymes are known to be up-regulated in several tissues of *B. mori* including gut and silk gland, and are speculated to play important role in flavonoid metabolism

and detoxification [44]. These observations indicate that dietary flavonoids may have some important role to play in silk gland while detoxification of allelochemicals produced by plants or endoparasites may be relevant for the alimentary canal in this instance. The catabolic activities were also enriched among biological processes and molecular functions in both tissues along with binding and signal transduction processes. There is a possibility that binding and signal transduction are carried out in conjunction with other biological processes upregulated in these tissues.

As an alternative measure of mining genes related to pigmentation, another GO class that is associated with the pigments, i.e., the "pigment-binding" molecular function was analyzed. This functional class includes pigment-binding proteins for isoprenoids, for e.g. carotenoids. Despite the fact that it was not significantly enriched in AC or SG, we probed its associated up-regulated transcripts as they are a possible source of information on pigmentation in *A. assamensis*. We identified transcripts which shared homology with the two transcript variants (BmStart1 or transcript variant 1-NM_001043533.1 and transcript variant 2-NM_001110362.1) of the carotenoid binding-protein (Cbp) gene of *B. mori*. Their coding regions were significantly conserved and they were upregulated in SG. BmStart1 encodes a Cbp isoform which is considered to be extremely important for carotenoid-dependent cocoon pigmentation and is found in both white/colourless and yellow strains of *B. mori* [45]. *B. mori* Cbp transcript variant 2 is found only in yellow cocoons. Upregulation of both these isoforms in *A. assamensis* suggests that it may have an active carotenoid uptake process in silk gland. However, existing knowledge shows that muga silk colour varies from one

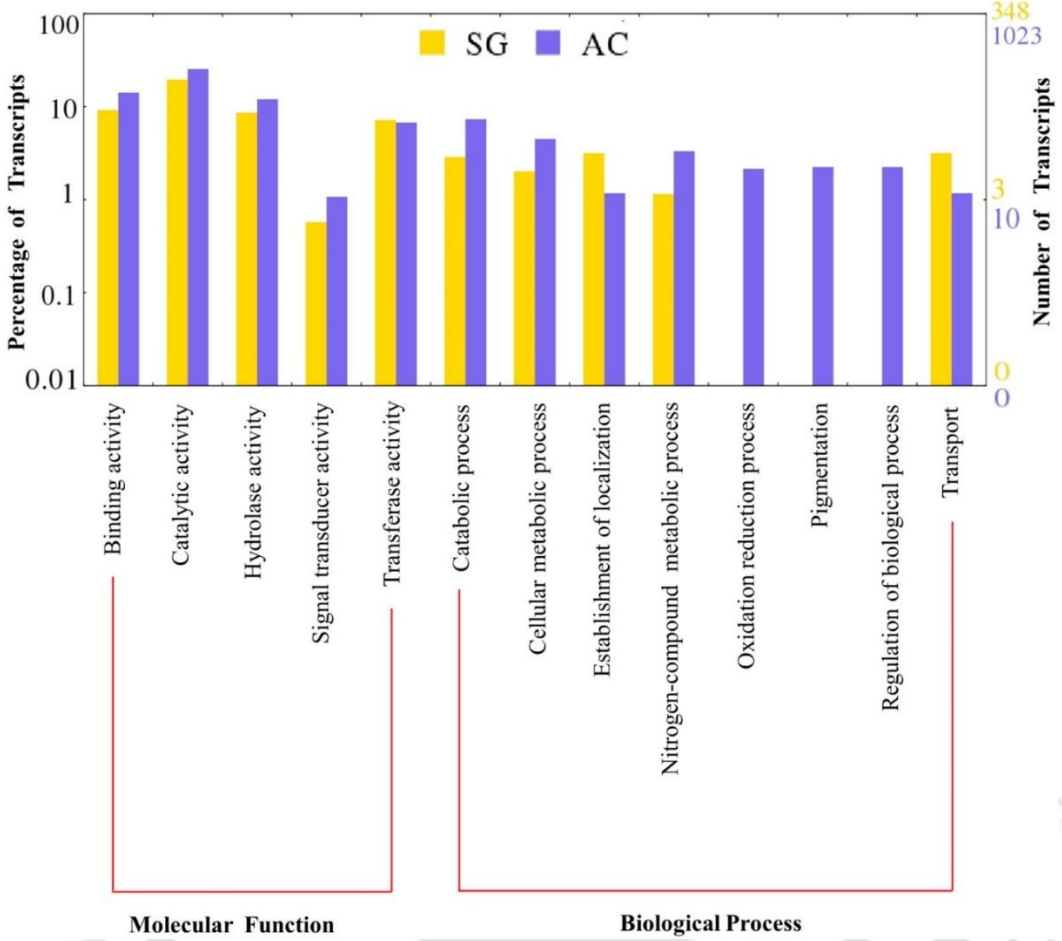


Fig. 2.3 Gene Ontology (GO) enrichment plot showing the enriched molecular functions and biological processes among the functionally annotated, up-regulated transcripts of Alimentary Canal (AC) and Silk Gland (SG) [Figure re-used in compliance with publisher guidelines [14]]

[Figure re-used in compliance with publisher guidelines [14]]

host plant to another and hence, further experimentation is necessary to remark upon the function of these candidate genes and their role in the muga silkworm [46,47]. Nevertheless, these candidate transcripts are a potential data source for further research.Various aspects of silk synthesis including experimental identification of involved functional elements have been well studied in Lepidopterans like *B. mori, G. mellonella*, etc. [48,49]. Identifying homologues of

these functional elements in *A. assamensis* will help in improving the existing knowledge on structural and regulatory aspects of its silk synthesis. While the sequence of fibroin gene of *A. assamensis* is known (Fibroin heavy chain or AaFhc, GenBank Accession-KJ862544.1), no information is available regarding other silk-related protein sequences such as sericin and silk gland factors [12]. Our differential expression studies showed that multiple candidate genes related to silk, namely, AaFHC, sericin and silk gland factor proteins were up-regulated. The AaFHC transcript assembled by us matched 100% with the experimentally determined one, thus testifying for our assembly process again. We also identified transcripts homologous to silk sericin MG-1 of *G. mellonella* (NCBI Accession-AGN03940.1) and silk gland factor-1 (NCBI Accession-NP_001037329), silk gland factor-3 (NCBI Accession- NP_001037456), homeobox protein homothorax-like (NCBI Accession - NP_001296493) and extradenticle (NCBI Accession-NP_001296565) of *B. mori* (Fig. 2.4). In *B. mori* and *G. mellonella*, sericin is expressed in mid-silk gland which binds fibroin filaments chaperoned by P25 on a cocoon wall or similar structure [50]. Silk sericin MG-1 homologue identified in *A. assamensis* SG transcriptome shared some common characteristics with that of *B. mori* and *G. mellonella* (Fig. 2.4A). Previous experiments had shown that sericin proteins of *B. mori* and *G. mellonella* had high proportions of Ser, Gly and Asn [49]. The conserved residues in our candidate transcript also constituted of a high percentage of Ser, Gly and Asn residues. Abundance of Ser (12.3%) and Pro (10.1%) residues and conservation of Gly-Ser repeats were also observed from analysis of amino acid composition. These facts outlined the similarities shared between the known sericin proteins and our transcript. A striking distinction between them was the

presence of a high proportion of Pro residues which is not commonly observed in another known silkworm sericin. However, evidence suggested that presence of more Pro residues can affect the intrinsic properties of a protein, for e.g. increase its thermal stability and differential calorimetric studies had shown sericin of *A. assamensis* to be more thermostable than *B. mori* or *A. mylitta* [51,52]. In light of these facts, we can presume that while functions carried out by *A. assamensis*

sericin may be similar to *B. mori* or *G. mellonella* sericin, their sequence and inherent properties might have differences. Presence of more than one sericin genes or alternative splice variants in muga silkworm is also possible.

In contrast to the presence of fibroin and sericin-like homologues in *A. assamensis*, P25 homologues were absent. Reported in earlier studies as well, its absence in muga silk is presumed to be compensated by the presence of a greater proportion of Ser and other polar residues in the C-terminal region of *A. assamensis* fibroin facilitating better solubility [12].

Other than the core silk proteins, we also found a set of candidate genes for silk gland factor-1, silk gland factor-3, homothorax and extradenticle of *A. assamensis*. The forkhead domain as well as the N- and C-terminal domain of putative silk gland factor-1 was conserved between *A. assamensis* and *B. mori,* including the DNA binding domain “KTYRRSYTHAKPPYSYISLITMAIQNNP SRMLTLSEIYQFIMDLFPFYRQNQQR W” and the two transactivation domains “LKQEPSGYAPAQHPFS” and “NYYQS PLYHHHHAHAQPPL” necessary for transcriptional stimulation (Fig. 2.4B) [53]. This suggests that *A. assamensis* silk gland factor-1 might be a part of the fork head/HNF-3 family like its *B. mori* counterpart and share similar functional attributes, i.e., transcriptional regulation

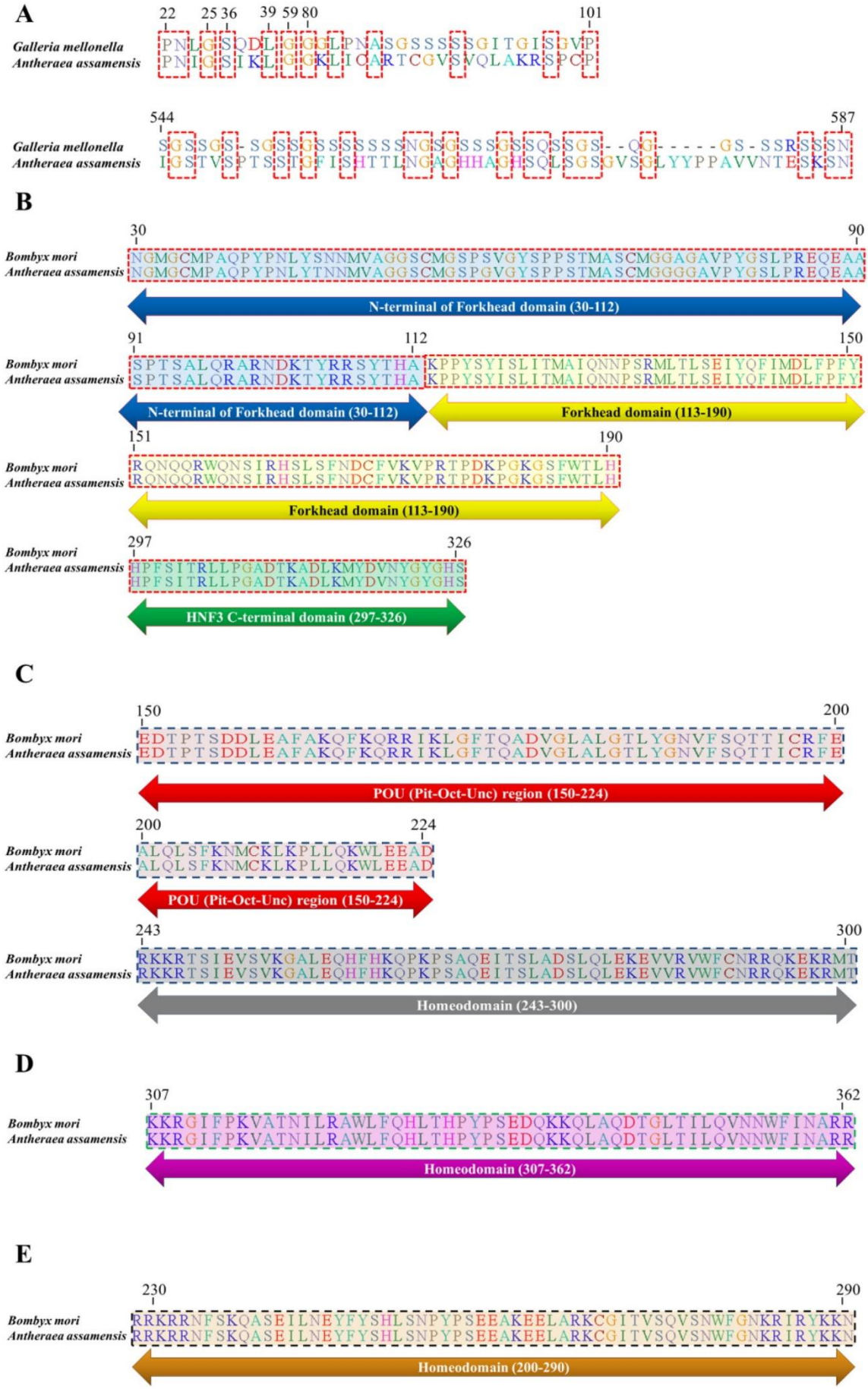

A
Galleria mellonella
Antheraea assamensis
B
Bombyx mori
Antheraea assamensis
N-terminal of Forkhead domain (30-112)
Forkhead domain (113-190)
HNF3 C-terminal domain (297-326)
C
POU (Pit-Oct-Unc) region (150-224)
Homeodomain (243-300)
D
Homeodomain (307-362)
E
Homeodomain (200-290)

Fig. 2.4 Conserved regions between the putative sericin-like transcript, silk gland factor-1, silk gland factor-3, homothorax and extradenticle in *A. assamensis* and the reference proteins from *Galleria mellonella* and *Bombyx mori*. NCBI Accession Numbers of the reference proteins; A- silk sericin MG-1 (NCBI Accession-AGN03940.1), B- silk gland factor-1 (NCBI Accession- NP_001037329), C- silk gland factor-3 (NCBI Accession- NP_001037456), D- homeobox protein homothorax-like (NCBI Accession- NP_001296493) and E- homeobox protein extradenticle (NCBI Accession- NP_001296565) [Figure re-used in compliance with publisher guidelines [14]]

of tissue-specific expression of sericin. Candidate silk gland factor-3 from *B. mori* was also identified with conserved POU (found in Pit-Oct-Unc transcription factors) and homeodomain, which contains conserved DNA binding sites involved in transcriptional regulation (Fig. 2.4C). *B. mori* silk gland factor-3 is reported to be differentially expressed for transcriptional regulation of ser-1, an event triggered by binding of the POU-M1 protein to silk gland factor-3 POU-domain which, in turn, interacts with one of the putative cis-acting regulatory elements of *B. mori* ser-1 gene [54]. Homologues of *B. mori* homothorax and extradenticle which act as putative co-factors for facilitative binding of HOX protein Antp for transcriptional regulation of ser1 gene, with conserved homeodomains were also identified (Fig. 2.4D-E) [55]. All these four candidates are known to be involved in regulation of ser-1 gene expression in *B. mori* suggesting the existence of a similar active transcriptomic network in *A. assamensis*. Complete gene sequence and RNA silencing or knock-out studies can yield more insights regarding these observations.

An important facet of the differential expression analysis was that some unannotated (PNTAa) transcripts were significantly differentially expressed in the tissues along with the annotated ones (1606 in AC and 1173 in SG). This dataset may include novel proteins, splice variants, microRNA precursors or long non-coding RNAs participating in tissue-specific or common processes of *A. assamensis* and can be better understood once muga silkworm's whole genome sequence becomes available.

2.3.4 Identification of candidate antimicrobial peptides in *A. assamensis* transcriptome

Other than silk synthesis, silkworms are also a good model for studying responses towards different immunogenic challenges. Silkworms rely heavily on a specific set of peptides, i.e., AMPs which are a vital part of the insect's immune system synthesized by fat bodies and circulated via haemolymph. A diverse range of AMPs are found in a silkworm's body to protect it against specific invaders such as bacteria, virus, fungi or protozoa. Muga silkworm continuously faces threat from invading micro-organisms; common pathogens are *Beauveria bassiana* (fungus), *Streptococcus* and *Staphylococcus* sps. (bacteria), nuclear polyhedrosis virus etc. which develop fatal diseases like flacherie, grasserie etc. (Patnaik, 2008). We identified several AMP homologues in the collective transcriptome. RB transcripts were a major source of these homologues as its library consisted of haemolymph and fat bodies along with other residual tissues. The types of AMPs identified were attacin, cecropin, defensin, gallerimycin, moricin and gloverin (Fig. 2.5). Attacin and gloverin are anti-bacterial peptides that tackle bacterial infection by inactivation of the synthesis of their vital outer

membrane proteins [56]. Attacin in *A. assamensis* was 81.5% identical to that in *S. cynthia ricini* (NCBI Accession-AB059394.1) with conserved regions in N-terminal and C-terminal (Fig. 2.5A). Attacin from S. cynthia ricini is considered to be a major inducible AMP which is effective against several strains of Gram-negative bacteria [57]. The isoelectric point (pI) of our putative attacin peptide is 8.95, close to the commonly observed range of 9.6-11.0 hinting that it may be similar to the basic attacin commonly found in Lepidoptera. Another class of insect AMP was cecropin which is homologous to that of *H. armigera* (NCBI Accession-ADR51154.1) (Fig. 2.5B). Named after *Hyalophora cecropia*, cecropins are a group of highly potent AMPs that inactivate microbes, including bacteria, fungi and parasites, within minutes by ionophoric mechanisms (Bulet et al., 1999). At present, there are three types of available *H. armigera* cecropins (HaCec 1, 2 and 3) [58]. Sequence comparison showed that our putative cecropin had a Lys-residue at position 61 similar to HaCec 1 and 3 while HaCec 2 had a Valresidue. HaCec 1 and 3 are known to be highly expressed during fungal infection in *H. armigera* and sequence characteristics indicate the same for our putative cecropin AMP. Defensin (homologous to that of *Trichoplusia ni* Defensin, NCBI Accession-ABV68852.1) and gallerimycin (homologous to that of *A. pernyi*, NCBI Accession-ACB45564.1) were also observed in our transcriptome; defensin usually acts against bacteria and sometimes against parasites; while gallerimycin confers anti-fungal resistance by binding to glucosylceramides in vulnerable fungi [56,59] (Fig. 2.5C-D). As for moricin, it is a highly basic AMP effective against both Gram-positive and Gram-negative bacteria which acts by increasing bacterial cytoplasmic membrane permeability causing cell death [60]. We identified two candidate moricins similar to that of *H. armigera* (NCBI AccessionADR51149.1)

and *B. mandarina* (NCBI Accession-AEM66431.1); both transcripts consisted of an amphipathic N-terminus and hydrophobic C-terminus that mediates the AMP's antibacterial activity (inferred by Kyte and Dolittle Plot in ProtScale (http://web.expasy.org/ protscale)) (Fig. 2.5E). Lastly, gloverin homologue of *A. assamensis* shared high sequence similarity with *A. mylitta* gloverin (GenBank AccessionABG72699.1); a few characteristics like Gly-richness (~18%) and basic pH (9.89) were also observed in our putative gloverin transcript adding to their similarities (Fig. 2.5F) [61].

An interesting aspect of this analysis was the identification of the same set of AMPs described above, excluding gallerimycin, in the SG transcriptome. Additionally, close homologues of hemolin and serpins were also found. The presence of AMPs and protease inhibitors in silkworm cocoons has been reported earlier; some novel peptides with unknown functions have been discovered in *B. mori* and *G. mellonella* silk glands and cocoons [62–64]. These studies suggested that silk gland could be a part of the insect immune system. Our observations also reflect this idea that silk glands may be a possible reservoir of AMPs and other immunogenic peptides which are released with the silk during spinning. Co-relating our observations with proteomic studies of the muga silk cocoon will yield more information regarding AMPs present in silk gland and if it carries out AMP synthesis itself or imports them to the gland.

Altogether, a repertoire of putative AMPs conferring immunity in *A. assamensis* against gram-positive and gram-negative bacteria, fungus and parasites was identified which can be probed further for gaining knowledge on silkworm immune system. Given that their mode of action is well-understood, these insect AMPs

may also find applications in disease resistance/treatment or vector control. Future experiments can be directed to learn about the AMP preferences of *A. assamensis* with respect to specific class of pathogens during immunecompromised situations and if they play synergistic roles in host defence.

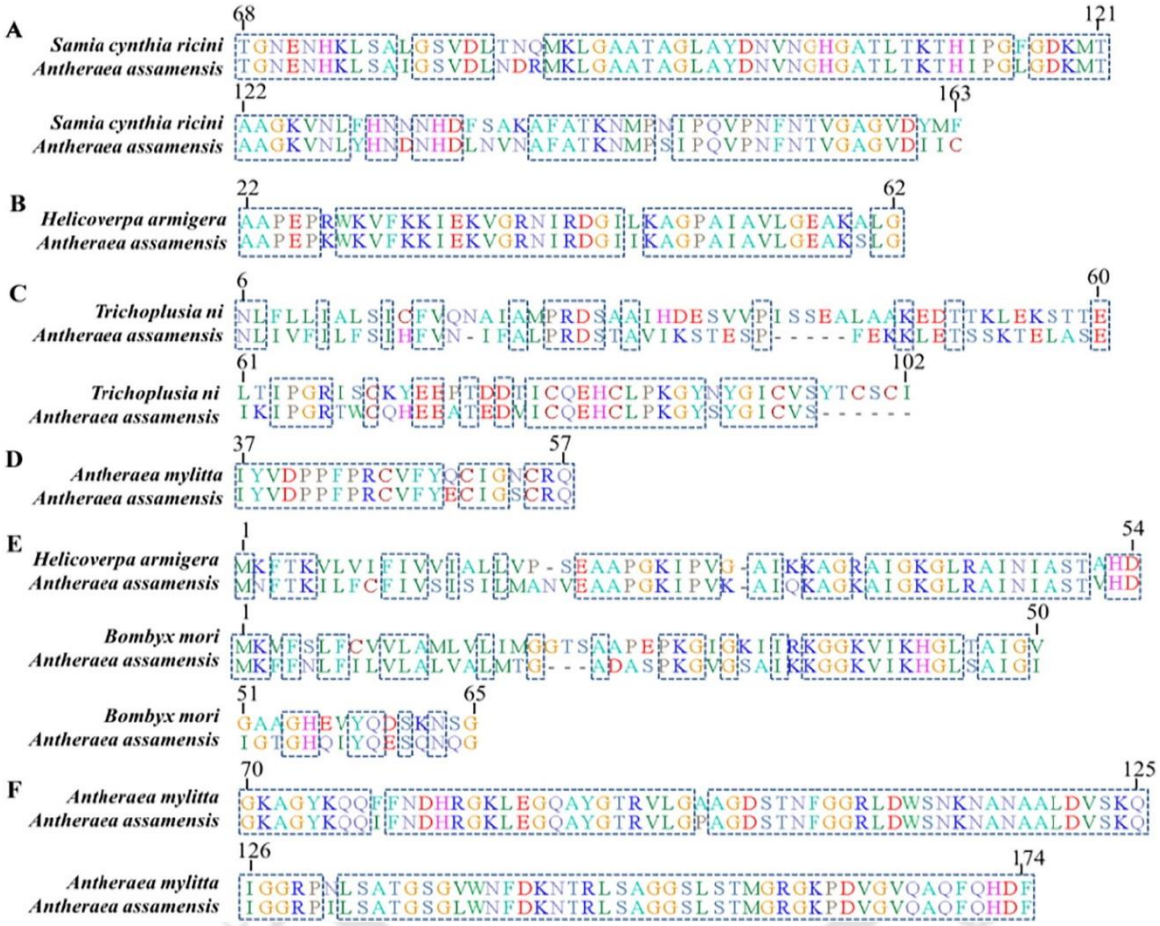


Fig. 2.5 Sequence conservation between the candidate antimicrobial peptides of *A. assamensis* and known antimicrobial peptides of different lepidopteran species: [A] Attacin, [B] Cecropin, [C] Defensin, [D] Gallerimycin, [E] Moricin and [F] Gloverin. The conserved residues are shown in blue boxes.

2.3.5 Experimental validation of *A. assamensis* transcripts by quantitative reverse-transcriptase PCR (RT-qPCR)-

We performed RT-qPCR of eight transcripts to experimentally validate our *A. assamensis* transcriptome assembly. Eight candidate transcripts, namely, actin, ecdysone receptor, juvenile hormone esterase, juvenile hormone acid o-methyl transferase, argonaute-2, calcium/calmodulin-dependent protein kinase as well as two putative novel transcripts from PNTAa dataset- PNTAa_1 and PNTAa_2 were selected for the experiment. The genes corresponding to the annotated transcripts are important for various molecular processes in insects. Actin is involved in formation of microtubules and microfilaments in the eukaryotic cytoskeleton. Juvenile hormone esterase and juvenile hormone acid O-methyl transferase is involved in the juvenile hormone biosynthetic pathway and their expression levels are closely related to metamorphosis and molting events in insects [65]. Calcium/Calmodulin-dependent protein kinase is involved in calcium signaling while argonaute is a highly conserved gene involved in RNA silencing [66,67]. Ecdysone receptor is a part of the nuclear receptor family and mediates the signalling of ecdysone, a steroid hormone which is involved numerous aspects of insect growth and development [68].

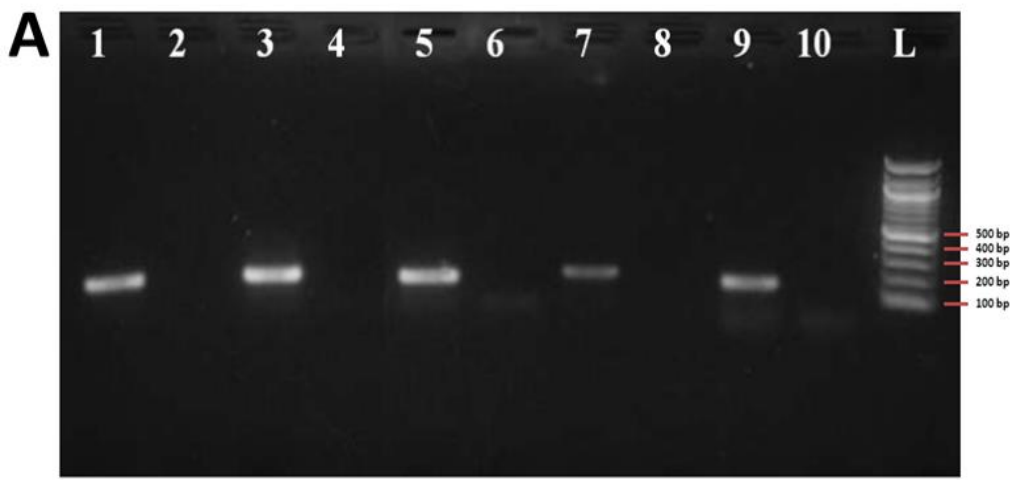
Forward and reverse primers for these candidate transcripts were designed followed by test for the primer's homology with non-specific transcripts (Table 2.4). These highly specific primers were used to perform RT-qPCR for these candidate genes followed by resolution on an agarose gel. Amplicons with single bands and expected amplicon sizes for each target transcript was observed indicating amplification of the target transcript (Fig. 2.6). PNTAa_2 displayed a single band with the expected amplicon size in AC sample.

All the six annotated transcripts were down-regulated in SG while in AC, all the transcripts, except actin and juvenile hormone esterase, were up-regulated (fold change (Fig. 2.7). Among the two putative novel transcripts, both were up-regulated in AC but PNTAa_2 was down-regulated in SG. The qPCR results were fairly consistent with the results of RNASeq experiment as analyzed by comparison of transcript abundance (FPKM values) generated via RSEM (Table 2.5). For e.g. abundance of actin transcript was lower in AC (FPKM= 69.16) and SG (FPKM= 35.27) relative to RB (FPKM= 366.52). Similarly, FPKM of argonaute-2 was greater in AC (FPKM= 5.92) and lower in SG (FPKM= 0.98) relative to RB (FPKM= 1.95) (Table 2.5). The RT-qPCR experiment results also reflected these observations in terms of differential expression. The relevance of differential expression of the putative transcripts is not discussed further as the validation experiment was carried out for random genes. Experimental proof of the existence and expression of the non-annotated transcripts affirmed their potential as a source of functional coding or non-coding transcripts. The reliability of this *de novo* transcriptome as a source of sequence and expression data for *A. assamensis* was also demonstrated.

Table 2.4 List of the eight transcripts for validation with their primers (F- Forward and R- Reverse) [Table re-used with modification in compliance with publisher guidelines [14]]

Candidate Gene	Primers	Expected amplicon size
Actin	F- CATCTACGAAGGTTACGCTC R- CCATCTCCTGCTCGAAGTC	191
Juvenile Hormone	F- CGGATCATGAGACCCAAGG	178

Esterase	R- CGAAAGCGAATCCTCCACC	
Juvenile hormone acid O-methyl transferase	F- TTTGTCACTCATATCGCTACC R- CACAGCACCAACGATCTTTC	192
Argonaute-2	F- TTCATGTCTGTAATCTCCACC R- GCTTTCACTCCGGATAAACC	185
Calcium/Calmodulin-dependent protein kinase	F- CAATCCGAATCGTGAGAGTG R- AGCCTCTCGTTCCAGTTTC	190
Ecdysone Receptor	F- CAGACAGAGGAAGACGAGG R- CTCGCAACATCATCACCTCG	181
PNTAa_1	F- CACTAAATTCCAGCGAACGA R- TTGTGCATTTGAGGACATGA	177
PNTAa_2	F- GCATTTGTTATCTATTTACGACGGT R- CAGGTGAACTTAAAGCGAGGT	165



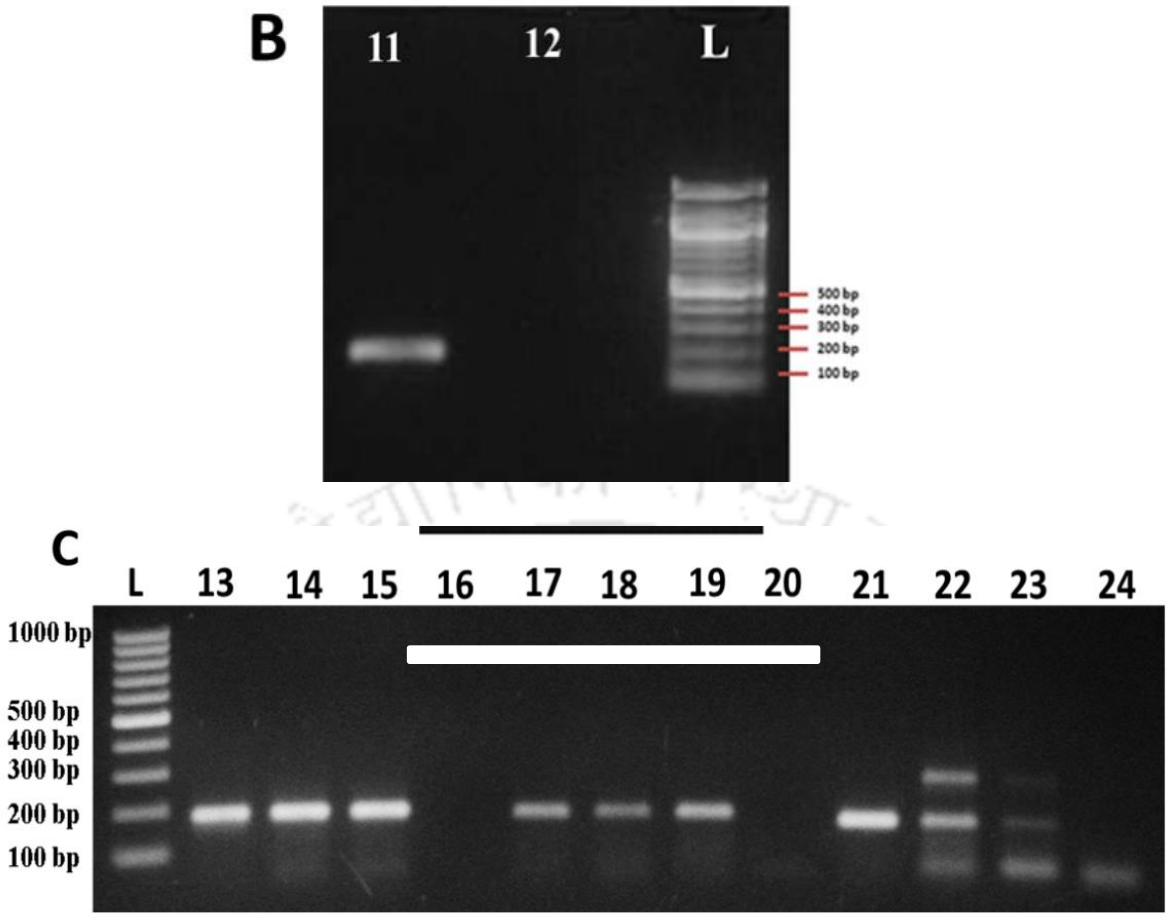


Fig. 2.6 Agarose gel electrophoresis of amplicons of the eight random transcripts selected for RT-qPCR validation. [Figure re-used in compliance with publisher guidelines [14]] Legends are tabulated below-

Lane No.	Figure No.	Sample Name
1	2.6A	Actin
2	2.6A	(No Template Control) Actin
3	2.6A	Juvenile Hormone Esterase
4	2.6A	(No Template Control) Juvenile Hormone Esterase
5	2.6A	Juvenile hormone acid O-methyl transferase
6	2.6A	(No Template Control) Juvenile hormone acid O-methyl transferase
7	2.6A	Argonaute

8	2.6A	(No Template Control) Argonaute
9	2.6A	Calcium/Calmodulin-dependent protein kinase
10	2.6A	(No Template Control) Calcium/Calmodulin-dependent protein kinase
L	2.6A	100 bp Ladder
11	2.6B	Ecdysone receptor
12	2.6B	(No Template Control) Ecdysone receptor
L	2.6B	100 bp Ladder
L	2.6C	100 bp Ladder
13	2.6C	Internal_Reference_Alpha_tubulin (AC)
14	2.6C	Internal_ Reference _Alpha_tubulin (RB)
15	2.6C	Internal_ Reference _Alpha_tubulin (SG)
16	2.6C	(No Template Control) Alpha_tubulin
17	2.6C	PNTAa_1(AC)
18	2.6C	PNTAa_1(RB)
19	2.6C	PNTAa_1(SG)
20	2.6C	(No Template Control) PNTAa_1
21	2.6C	PNTAa_2 (AC)
22	2.6C	PNTAa_2 (RB)
23	2.6C	PNTAa_2 (SG)
24	2.6C	(No Template Control) PNTAa_2

Table 2.5- Transcript abundance estimates for the transcripts targeted for transcriptome validation of *Antheraea assamensis* in terms of (Fragments Per Kilobase of transcript per Million mapped reads (FPKM) values [Table re-used with modification in compliance with publisher guidelines [14]]

TRANSCRIPT	**FPKM**		
	AC	**SG**	**RB**
Actin	69.16	35.27	366.52

Argonaute	5.92	0.98	1.95
Ca^{2+}/calmodulin-dependent protein kinase II	7.9	2.6	7.52
Ecdysone receptor	16.25	4.3	10.13
Juvenile hormone acid methyltransferase	22.87	3.05	3.11
Juvenile hormone esterase	1.43	1.89	33.86
PNTAa_1	5.98	6.63	0.9
PNTAa_2	6.07	2.9	1.112

2.4 Conclusion and future prospects

In this study, we reported the multi-tissue *de novo* transcriptome of the 5th instar larvae of *A. assamensis* reared on the leaves of *M. bombycina* for the first time. More than ~74% of the transcripts were functionally annotated and classified under broad functional categories of GO, KEGG, COG and Pfam. These transcripts, thus, constituted an exhaustive resource of candidate genes for *A. assamensis* which doesn't have a whole genome sequence till date. The study reported the sequences and differential expression of putative silk gland factor-1, silk gland factor-3, sericin-like transcript, homothorax, extradenticle and carotenoid transcripts providing molecular insights to the silk synthesis process of *A. assamensis.* The PNTAa dataset contributed novel candidates- new genes, splice variants or non-coding RNAs in *A. assamensis* which can be ascribed new identities once the whole genome sequence is available. The transcriptome data is now hosted in an online MugaSeqDB (http://mugaseqdb.in). Overall, the

resources generated by this study will enable further molecular or genomic studies on this commercially important organism.

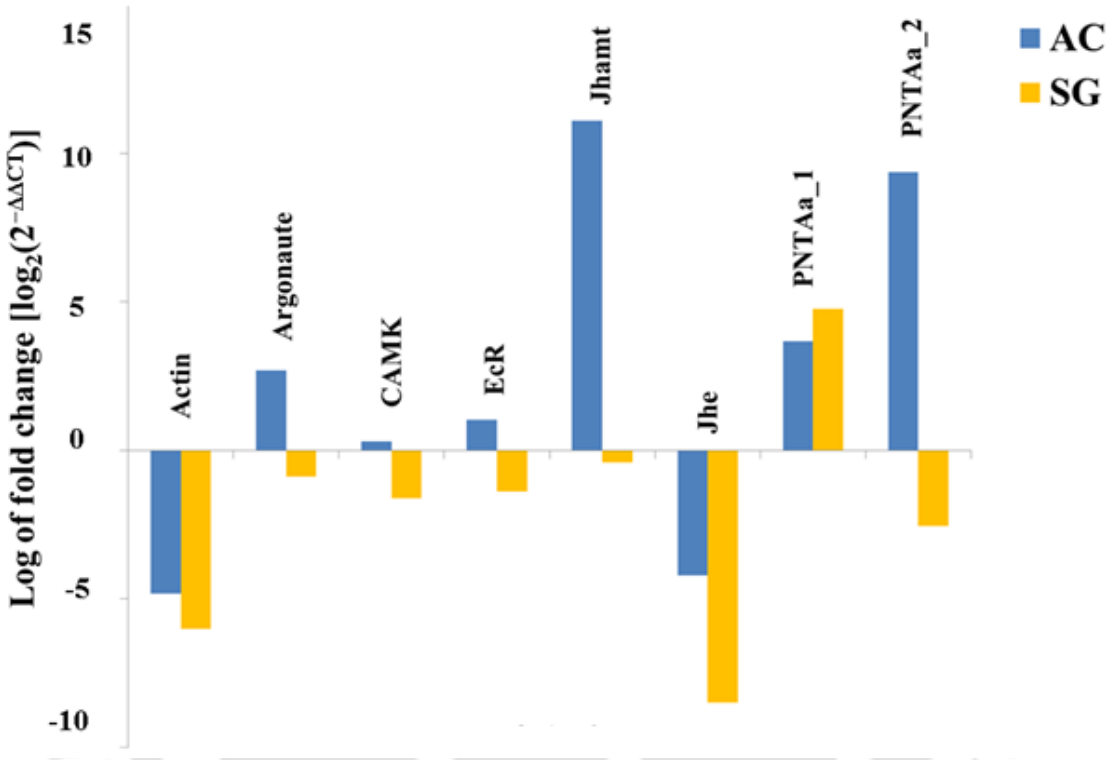


Figure 2.7- Log fold change values of expression of the eight random transcripts used for transcriptome validation in Alimentary Canal (AC) and Silk Gland (SG) relative to the control tissue (Residual Body, RB) using alpha-tubulin gene as internal standard [CAMK- Ca2+/calmodulin-dependent protein kinase II, EcR- Ecdysone receptor, Jhamt- Juvenile hormone acid methyltransferase, Jhe- Juvenile hormone esterase, PNTAa_1 and 2- Putative novel transcript of *A. assamensis* 1 and 2] [Figure re-used in compliance with publisher guidelines [14]]

REFERENCES

1. Li Y, Wang G, Tian J, et al. Transcriptome Analysis of the Silkworm (*Bombyx mori*) by High-Throughput RNA Sequencing. Gibas C, ed. PLoS One. 2012;7(8):e43713. doi:10.1371/journal.pone.0043713

2. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol. 2008;38(12):1036-1045. doi:10.1016/J.IBMB.2008.11.004

3. Tikader A, Vijayan K, Saratchandra B. Muga silkworm, *Antheraea assamensis* (Lepidoptera: Saturniidae) - an overview of distribution, biology and breeding. Eur J Entomol. 2013;110(2):293-300. doi:10.14411/eje.2013.096

4. Kundu S. Silk Biomaterials for Tissue Engineering and Regenerative Medicine. Woodhead Publishing; 2014.

5. Kasoju N, Bhonde RR, Bora U. Preparation and characterization of *Antheraea assama* silk fibroin based novel non-woven scaffold for tissue engineering applications. J Tissue Eng Regen Med. 2009;3(7):539-552. doi:10.1002/term.196

6. Kasoju N, Bora U. *Antheraea assama* silk fibroin-based functional scaffold with enhanced blood compatibility for tissue engineering applications. Adv Eng Mater. 2010;12(5):B139-B147. doi:10.1002/adem.200980055

7. Padamwar MN, Pawar AP. Silk sericin and its applications: a review. Vol 63.; 2004.

8. Ma M, Hussain M, Dong S, Zhou W. Characterization of the pigment in naturally yellow-colored domestic silk. Dye Pigment. 2016;124:6-11. doi:10.1016/J.DYEPIG.2015.08.003

9. Tabunoki H, Higurashi S, Ninagi O, et al. A carotenoid-binding protein (CBP) plays a crucial role in cocoon pigmentation of silkworm (*Bombyx mori*) larvae. FEBS Lett. 2004;567(2-3):175-178. doi:10.1016/

j.febslet.2004.04.067

10. Birol I, Behsaz B, Hammond SA, Kucuk E, Veldhoen N, Helbing CC. De novo transcriptome assemblies of *Rana* (*Lithobates*) *catesbeiana* and Xenopus laevis tadpole livers for comparative genomics without reference genomes. PLoS One. 2015;10(6):e0130720. doi:10.1371/journal.pone.0130720

11. Wang X-W, Luan J-B, Li J-M, Bao Y-Y, Zhang C-X, Liu S-S. De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics. 2010;11(1):400. doi:10.1186/1471-2164-11-400

12. Gupta K A, Mita K, Arunkumar KP, et al. Molecular architecture of silk fibroin of Indian golden silkmoth, *Antheraea assama*. Sci Rep. 2015;5(1):12706. doi:10.1038/srep12706

13. Bindroo BB, Singh NT, Sahu AK, Chakravorty R. Muga silkworm host plants. Indian Silk. 2006;44:13-17.

14. Chetia H, Kabiraj D, Singh D, et al. De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). Gene. 2017;611. doi:10.1016/j.gene.2017.02.021

15. Singh D, Kabiraj D, Sharma P, et al. The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. PLoS One. 2017;12(11). doi:10.1371/journal.pone.0188077

16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

17. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590-6. doi:10.1093/nar/gks1219

18. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat

Methods. 2012;9(4):357-359. doi:10.1038/nmeth.1923

19. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18(5):821-829. doi:10.1101/gr.074492.107

20. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086-1092. doi:10.1093/bioinformatics/bts094

21. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29(7):644-652. doi:10.1038/nbt.1883

22. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061-1067. doi:10.1093/bioinformatics/btm071

23. Arunkumar KP, Tomar A, Daimon T, Shimada T, Nagaraju J. WildSilkbase: an EST database of wild silkmoths. BMC Genomics. 2008;9(1):338. doi:10.1186/1471-2164-9-338

24. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150-3152. doi:10.1093/bioinformatics/bts565

25. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222-30. doi:10.1093/ nar/gkt1223

26. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277-80. doi:10.1093/nar/gkh063

27. Tatusov RL, Galperin MY, Natale DA, Koonin E V. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28(1):33-36. doi:10.1093/nar/28.1.33

28. Shimomura M, Minami H, Suetsugu Y, et al. KAIKObase: An integrated

silkworm genome database and data mining tool. BMC Genomics. 2009;10(1):486. doi:10.1186/1471-2164-10-486

29. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12(1):323. doi:10.1186/1471-2105-12-323

30. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. doi:10.1186/gb-2010-11-10-r106

31. Ye J, Fang L, Zheng H, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 2006;34(Web Server issue):W293-7. doi:10.1093/nar/gkl031

32. Ye J, Coulouris G, Zaretskaya I, et al. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics. 2012;13(1):134. doi:10.1186/1471-2105-13-134

33. Gschloessl B, Beyne E, Audiot P, et al. De novo transcriptomic resources for two sibling species of moths: *Ostrinia nubilalis* and *O. scapulalis*. BMC Res Notes. 2013;6(1):73. doi:10.1186/1756-0500-6-73

34. Patnaik BB, Wang TH, Kang SW, et al. Sequencing, De Novo Assembly, and Annotation of the transcriptome of the endangered freshwater pearl bivalve, *Cristaria plicata*, provides novel insights into functional genes and marker discovery. PLoS One. 2016;11(2):e0148622. doi:10.1371/journal.pone.0148622

35. Zhu J-Y, Li Y-H, Yang S, Li Q-W. De novo assembly and characterization of the global transcriptome for *Rhyacionia leptotubula* using Illumina paired-end sequencing. PLoS One. 2013;8(11):e81096. doi:10.1371/journal.pone.0081096

36. Ma K, Qiu G, Feng J, Li J. Transcriptome analysis of the oriental river prawn, *Macrobrachium nipponense* using 454 pyrosequencing for discovery of genes and markers. PLoS One. 2012;7(6):e39727. doi:10.1371/journal.pone.0039727

37. Vogel H, Altincicek B, Glöckner G, Vilcinskas A. A comprehensive transcriptome and immune-gene repertoire of the lepidopteran model host *Galleria mellonella*. BMC Genomics. 2011;12(1):308. doi:10.1186/1471-2164-12-308

38. Laity JH, Lee BM, Wright PE. Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol. 2001;11(1):39-46. doi:10.1016/S0959-440X(00)00167-6

39. Li D, Roberts R. WD-repeat proteins: structure characteristics, biological function, and their involvement in human diseases. Cell Mol Life Sci. 2001;58(14):2085-2097

40. Perrakis A, Ouzounis C, Wilson KS. Evolution of immunoglobulin-like modules in chitinases: their structural flexibility and functional implications. Fold Des. 1997;2(5):291-294. doi:10.1016/S1359-0278(97)00040-0

41. Xie W, Lei Y, Fu W, et al. Tissue-specific transcriptome profiling of *Plutella Xylostella* third instar larval midgut. Int J Biol Sci. 2012;8(8):1142-1155. doi:10.7150/ijbs.4588

42. Appel HM, Martin MM. Gut redox conditions in herbivorous lepidopteran larvae. J Chem Ecol. 1990;16(12):3277-3290. doi:10.1007/BF00982098

43. Chaw RC, Correa-Garhwal SM, Clarke TH, Ayoub NA, Hayashi CY. Proteomic evidence for components of spider silk synthesis from black widow silk glands and fibers. J Proteome Res. 2015;14(10):4223-4231. doi:10.1021/acs.jproteome.5b00353

44. Huang F-F, Chai C-L, Zhang Z, et al. The UDP-glucosyltransferase multigene family in *Bombyx mori*. BMC Genomics. 2008;9(1):563. doi:10.1186/1471-2164-9-563

45. Sakudoh T, Sezutsu H, Nakashima T, et al. Carotenoid Silk Coloration Is Controlled by a Carotenoid-Binding Protein, a Product of the Yellow Blood Gene.; PNAS 2007;104(21):8941-8946.

46. Landrum JT. Carotenoids: physical, chemical, and biological functions and

properties. CRC Press; 2010.

47. Mohanty PK. Tropical Tasar Culture in India. Daya Publishing House; 1998.

48. Borgohain A. Silk and its biosynthesis in silkworm *Bombyx mori*. J Acad Ind Res. 2015;3.

49. Zurovec M, Sehnal F. Unique molecular architecture of silk fibroin in the waxmoth, *Galleria mellonella*. J Biol Chem. 2002;277(25):22639-22647. doi:10.1074/jbc.M201622200

50. Xia Q, Li S, Feng Q. Advances in silkworm studies accelerated by the genome sequencing of *Bombyx mori*. Annu Rev Entomol. 2014;59(1):513-536. doi:10.1146/annurev-ento-011613-161940

51. Dutta S, Bharali R, Devi R, Devi D. Purification and Characterization of Glue Like Sericin Protein from a Wild Silkworm *Antheraea assamensis Helfer*. 2012;1(2):229-233.

52. Kumar S, Tsai CJ, Nussinov R. Factors enhancing protein thermostability. Protein Eng. 2000;13(3):179-191.

53. Mach V, Takiya S, Ohno K, Handa H, Imai T, Suzuki Y. Silk gland factor-1 involved in the regulation of *Bombyx* sericin-1 gene contains fork head motif. J Biol Chem. 1995;270(16):9340-9346. doi:10.1074/JBC.270.16.9340

54. Fukuta M, Matsuno K, Hui CC, et al. Molecular cloning of a POU domain-containing factor involved in the regulation of the *Bombyx* sericin-1 gene. J Biol Chem. 1993;268(26):19471-19475.

55. Kimoto M, Tsubota T, Uchino K, Sezutsu H, Takiya S. Hox transcription factor Antp regulates sericin-1 gene expression in the terminal differentiated silk gland of *Bombyx mori*. Dev Biol. 2014;386(1):64-71. doi:10.1016/j.ydbio.2013.12.002

56. Bulet P, Hetru C, Dimarcq JL, Hoffmann D. Antimicrobial peptides in insects; structure and function. Dev Comp Immunol. 23(4-5):329-344.

57. Kishimoto K, Fujimoto S, Matsumoto K, Yamano Y, Morishima I. Protein purification, cDNA cloning and gene expression of attacin, an antibacterial protein, from eri-silkworm, *Samia cynthia ricini*. Insect Biochem Mol Biol. 2002;32(8):881-887.

58. Wang Q, Liu Y, He H-J, et al. Immune responses of *Helicoverpa armigera* to different kinds of pathogens. BMC Immunol. 2010;11(1):9. doi:10.1186/1471-2172-11-9

59. Schuhmann B, Seitz V, Vilcinskas A, Podsiadlowski L. Cloning and expression of gallerimycin, an antifungal peptide expressed in immune response of greater wax moth larvae, *Galleria mellonella*. Arch Insect Biochem Physiol. 2003;53(3):125-133. doi:10.1002/arch.10091

60. Hara S, Yamakawa M. Moricin, a novel type of antibacterial peptide isolated from the silkworm, *Bombyx mori*. J Biol Chem. 1995;270(50):29923-29927. http://www.ncbi.nlm.nih.gov/pubmed/8530391.

61. Axén A, Carlsson A, Engström A, Bennich H. Gloverin, an antibacterial protein from the immune hemolymph of *Hyalophora pupae*. Eur J Biochem. 1997;247(2):614-619.

62. Guo X, Dong Z, Zhang Y, et al. Proteins in the cocoon of silkworm inhibit the growth of *Beauveria bassiana*. Ling E, ed. 2016;11(3):e0151764. doi:10.1371/journal.pone.0151764

63. Nirmala X, Kodrík D, Zurovec M, Sehnal F. Insect silk contains both a Kunitz-type and a unique Kazal-type proteinase inhibitor. Eur J Biochem. 2001;268(7):2064-2073.

64. Shaik HA, Sehnal F. Hemolin expression in the silk glands of *Galleria mellonella* in response to bacterial challenge and prior to cell disintegration. J Insect Physiol. 2009;55(9):781-787. doi:10.1016/j.jinsphys.2009.04.010

65. Noriega FG. Juvenile hormone biosynthesis in insects: what is new, what do we know, and what questions remain? Int Sch Res Not. 2014;2014:1-16. doi:10.1155/2014/967361

66. Braun AP, Schulman H. The multifunctional Calcium/Calmodulin-dependent protein kinase: from form to function. Annu Rev Physiol. 1995;57(1):417-445. doi:10.1146/annurev.ph.57.030195.002221

67. Wang G-H, Jiang L, Zhu L, et al. Characterization of Argonaute family members in the silkworm, *Bombyx mori*. Insect Sci. 2013;20(1):78-91. doi:10.1111/j.1744-7917.2012.01555.x

68. Jindra M, Palli SR, Riddiford LM. The juvenile hormone signaling pathway in insect development. Annu Rev Entomol. 2013;58(1):181-204. doi:10.1146/annurev-ento-120811-153700

CHAPTER 3

De novo transcriptome of *Antheraea assamensis* host plants: *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari)

CHAPTER 3

De novo transcriptome of *Antheraea assamensis* host plants: *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari)

ABSTRACT

Muga silkworm is an economically significant lepidopteran that is strictly phytophagous. Its physical properties like silk yield, size, etc. has are governed by the host plant it is reared upon. However, the amount of molecular data on its host plants is scarce. In this study, we attempted to address this existing gap by studying the two Lauraceae host plants, namely, Som (*Machilus bombycina*) and Mejankari (*Litsea citrata*). Som and Mejankari are two primary and secondary host plants of muga silkworm, respectively. Here, we sequenced and annotated the *de novo* transcriptome of both host plants. Som and Mejankari transcriptome consisted of 55,400 and 1,38,690 transcripts, respectively, of which ~50% transcripts in both species were annotated in this study. We also identified the putative genomic components of their chemical defense related to the glucosinolate-myrosinase system and antimicrobial peptides that tackle microbes as well as herbivorous pests. Our findings generated a novel resource of genomic data on the Lauraceae family. It also provided a new perspective to study silkworm host plants in terms of their immune system.

3.1 INTRODUCTION

Host plants are an integral part of any herbivore's life span as it is the solitary dietary source for the organism. Lepidopterans are one such order consisting of ~99% herbivorous insects which acquire nutrition from their host plants including nitrogen, carbon and other trace elements [1]. However, from a plant's perspective, its relationship with any herbivore is usually equivalent to that of a host and a pest. Hence, plants employ various direct and indirect defenses to confront herbivores. Direct defenses target a pest's host plant preference, survival and reproductive success while indirect defenses involve attracting predators of their existing pests [2]. Both modes involve various morphological as well as chemical effectors, some of which come handy in tackling pathogenic challenges as well [3]. Here, we shall discuss the molecular aspect of these chemical defense effectors which are usually diverse, has broad-spectrum action and are either constitutively produced or actively transcribed in response to plant hormone-triggered signalling related to physical damage, for e.g., larval feeding [2]. Chemical components of plant defense include but are not limited to antimicrobial peptides, toxic compounds (terpenoids, alkaloids, etc.) and volatile organic compounds (e.g. volicitin) [4].

Antimicrobial peptides are an integral part of the plant immune system and act as a part of chemical weaponry against biotic stresses. Their necessity can be inferred from the fact that around 2-3% of the total gene content in a plant species constitute AMPs. They're found ubiquitously across the kingdom and have a much broader activity spectrum than just microbes. They also play a role in plant defense against a wide range of multicellular pests like insects, molluscs,

nematodes, etc. These broad-spectrum activities are made possible by three fundamental features: heterogeneity in host peptide composition, existence of an effective signaling network that triggers the expression of appropriate AMPs with respect to biotic signals and remarkable structural stability rendered by presence of more Cys-residues (which forms intra-peptide disulfide bonds) and other biophysical features [5]. The first plant-derived antimicrobial peptide, purothionin, was isolated from wheat almost half a century back in 1972 [6]. Since then, there has been a steady growth in the number of plant AMPs which are classified as defensins, thionins, knottins, snakins, etc. They have a standard mode of action which relies upon the cationic nature of peptides to infiltrate negatively charged cell membranes of pathogens and form pores or disrupt the membrane structure which eventually leads to leakage of cell content and death [7]. Some can also interact with specific membrane components, for e.g. receptors, leading to internalization and subsequent disruption of crucial cellular functions like biomolecular biosynthesis, deactivation of host immune effectors and eliciting allergenic responses [5]. The existing diversity in available plant AMPs suggest that more variant modes of pest/pathogen neutralizing action will exist. Overall, these evidences bolster the role of plant AMPs as robust chemical defense against pests/pathogens and emphasizes on the need to study them.

Plants are also marvels of evolution due to their wide range of biosynthetic products, a major share of which interact with other organisms and are potentially defense compounds. They belong to different chemical classes of which terpenoids, steroids, alkaloids and phenolics are the most populous [4]. Their mechanism involves inhibition of metabolite transport, signaling pathways or metabolism as well as disruption of membrane [2].

Here, we have focused on one of these classes, glucosinolates, which are very effective in defense from herbivores. Glucosinolates are anionic thioglucodises which are well-studied in the Brassicaceae family [8]. Glucosinolates are relatively inert precursors of aglycones which are synthesized and compartmentalized until tissue damage. During such an event arising from herbivore chewing or equivalent, the glucosinolates in the plant tissues come in contact with myrosinase, their hydrolyzing enzyme, and form the unstable aglycones. These aglycones with aliphatic side chain are hydrolyzed at neutral pH to form isothiocyanates or nitriles in presence of Fe^{2+} or acidic pH [9]. Some of the resulting compounds are toxic to many insect species, including generalists like *Spodoptera eridania* or specialists like *Plutella xylostella* and plant pathogenic fungi [9–12]. These intermediates can covalently modify nucleic acid or proteins and have detrimental effect on the redox homeostasis of a target cell [4]. However, for some specialist herbivores, however, these may rather act as recognition cues or attractants. Overall, this Glucosinolate-Myrosinase (Glc-myr) system has been reported and well-studied in plants from Brassicaceae, Capparidaceae, and Tropaeolaceae families [4].

Extensive studies have been performed to study plant defense mediators and their biosynthetic pathways in the model plant, *Arabidopsis thaliana,* agricultural crops like rice, wheat, cabbage, mustard, etc. due to their commercial relevance as well as higher availability of molecular data [13,14]. However, similar studies on host plants reared commercially for sericulture is limited. Sericulture is one of the most anthropologically as well as ecologically viable agro-practices and very popular in India and China. Every year, these plants face a multitude of biotic challenges, from unicellular pathogens to invertebrate pests as well as

vertebrates. While the latter can be warded off, understanding the former two aspects is necessary to defend these host plants and promote sustainable sericulture.

The muga host plants are divided into primary, secondary and tertiary categories based on the feeding preference of muga silkworm [15]. All primary and secondary host plants of muga silkworm are from Laurales order and Lauraceae family. *A. assamensis,* our organism of interest, is a lepidopteran herbivore predominantly reared on primary host plants from Lauraceae for commercial needs [16]. ~20 additional plants from Lauraceae, Magnoliaceae, Rutaceae, etc. have been reported as its tertiary host plants [15,17]. However, availability of rearing data on most of the non-Lauraceae host plants is scarce to none and this lack of data on its dietary specialization and corresponding effects on muga silk makes it difficult to classify *A. assamensis* as a generalist or a specialist herbivore [18].

Table 3.1 Characteristics of *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari)

Vernacular Name	**Som**	**Mejankari**
Taxonomical Class		
Order	Laurales	Laurales
Family	Lauraceae	Lauraceae
Genus	*Machilus*	*Litsea*
Species	*bombycina*	*citrata*
Homotypic synonym	*Persea bombycina*	*Litsea cubeba*
NCBI Taxonomy ID	466588	1987498

Chromosome number (n)	12	12
Whole Genome	Not available	Not available
Transcriptome	Not available	Available
Mitochondrial genome	Not available	Not available
Chloroplast genome	Not available	Not available

Som (*M. bombycina)* is a primary host plant of muga silkworm which is evergreen and available almost all around the year (Table-3.1). It is predominantly used for commercial rearing of muga silkworm in North-East India. Muga silkworm reared on *M. bombycina* produces golden-yellow colored cocoon [15]. Mejankari (*L. citrata*), on the other hand, is a secondary host plant of the same insect, however, its usage for commercial rearing is extremely rare (Table-3.1). It is a deciduous shrub or tree, primarily cultivated for essential oils [16,19]. Muga silkworms reared on *L. citrata* leaves produce creamy white cocoons [15]. Overall, the Indian sericulture industry involves ~8 million people and earns ~2000-2500 crore every year. Muga silkworm has a significant contribution towards the farmers of Assam (http://texmin.nic.in/sites/default/files/note-on-sericulture2017-18-ThirdQtr.pdf).

Now, commercial benefits derived from muga silkworm indirectly depends upon its host plants. However, despite their significance of silkworm hosts, there is a massive disparity between the amount of studies performed on other commercial crops and these Lauraceae plants. A quick search in NCBI database reveals the presence of a few barcode sequences, but other forms of biomolecular data are infrequent [20]. In this study, we addressed this existing gap by studying the two Lauraceae host plants discussed above, namely, *M. bombycina* and *L. citrata*.

Som is always of commercial as well as biological interest due to its status as a primary host plant and long-term association with muga silkworm, respectively. Use of Mejankari as a host plant for rearing muga silkworm has diminished over the years; one of the reasons being the difficulties associated with its successful cultivation. Despite this, it has potential as a host plant and reports of alteration in silk coloration on Mejankari fed muga silkworms further stress upon this fact. The age-old cultural practice of producing muga silk on selected host plants also provides a possibility to examine the existing theory of co-adaptation arising from long-term insect-host plant association. This phenomenon has already been observed in the lepidopterans, *Pieris rapae* and *P. xylostella* [21]. Keeping these in mind, we sequenced and annotated the *de novo* transcriptome of both Lauraceae plants. We specifically identified putative genomic components of their chemical defense related to the glucosinolate-myrosinase system and AMPs that tackle microbes as well as herbivorous pests. Our findings generated a novel resource of genomic data on the Lauraceae family. It also provided a new perspective to study silkworm host plants and contributed towards a better understanding of plant immune system.

3.2 MATERIALS AND METHODS

3.2.1 Sample collection, RNA isolation, cDNA library preparation and sequencing:

M. bombycina and *L. citrata* leaves were collected from Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh (26.7844° N, 94.3443° E). The leaves were immediately washed, surface sterilized with sodium hypochlorite solution (1.0%) and dissected into very thin sections under sterile conditions and

stored in RNAlater stabilization solution (Ambion™) at -80˚C. RNeasy® plant mini kit (Qiagen, Hiden, Germany) was used for RNA isolation from the plant tissues. The concentration, purity and integrity of the isolated RNA (A260 /A280 ratio ≥1.8 and RIN number ≥8) were verified using Nanodrop spectrophotometer and High Sensitivity Bioanalyzer Chip (Agilent Technologies, CA, U.S.A.). Preparation of *L. citrata* and *M. bombycina* sequencing libraries was performed with Illumina-compatible SureSelect Strand-Specific RNA Library (Part # G9691-90010) and TruSeq RNA Library (Part # 15008136) preparation kits (Agilent Technologies, Santa Clara, CA, U.S.A.). The resulting cDNA libraries for *M. bombycina* and *L. citrata* were sequenced on Illumina HiSeq™ 2000 and 4000 sequencer platforms respectively, using the paired-end sequencing protocol at Genotypic Technology Pvt. Ltd., Bangalore, India.

3.2.2 Quality control of raw data and *de novo* assembly of transcriptome:

The quality metrices of the resulting raw reads per data set were examined using FastQC v0.11.5 [22]. Based on the report, both paired-end datasets were corrected for adapter contamination, over-represented sequences, erroneous k-mers, low quality reads as well as tiles (Phred quality score cutoff was ≥30) using a combination of Trimmomatic 0.35, Trim Galore, rCorrector tools as well as in-house shell scripts [23–25]. The resulting raw reads were mapped to ribosomal rRNA reads (entitled SSUParc and LSUParc files) from the SILVA rRNA database project using bowtie 2 and unmapped read pairs were retained [26]. The resultant reads were re-examined using FastQC to check for attainment of desirable quality features.

Following this, the final set of paired-end reads were utilized as inputs for de novo assembly of *M. bombycina* and *L. citrata* transcriptomes via Trinity tool (v2.6.5) utilizing the in-built parameters for normalization by read set and k-mer size of 32 [27]. The quality of transcriptome assemblies was assessed on the basis of parameters like N50 value, mean transcript length, percentage of reads mapped to the transcriptome, etc. using the scripts provided under Trinity package. Transcriptome completeness was quantitatively examined via BUSCO (Benchmarking Universal Single-Copy Orthologs) tool v3 using the dataset "Embryophyta odb10" curated by OrthoDB in transcriptome mode [28,29].

3.2.3 Annotation and functional classification of the transcriptome:

The assembled transcriptomes of *M. bombycina* and *L. citrata* were aligned to a combination of protein databases- NCBI Non-redundant Protein and UniProt databases for Viridiplantae (NCBI & UniProt Taxon ID- 33090) downloaded in April 2018 using the blastx program of NCBI BLAST package v 2.6.0+ with the e-value cut-off of 1e-05 [30]. The top-most hits with best bit-score followed by query coverage and percentage identity (in that order) were sorted and de-duplicated. Additional information on annotation, namely, Gene ontology (Biological process, Molecular function and Cellular compartment) were obtained from UniProt database using the respective reference proteins. The transcriptomes were translated into putative peptides using TransDecoder v5.5.0. The functional domains in these hypothetical proteins were identified by matching HMM profiles described in Pfam-A database using the hmmscan command of hmmer tool v3.1b2 with e-value threshold (E) for model reporting as 1e-03 [31,32]. The

remaining set of transcripts without annotation or general identifier were classified as putative novel transcripts for *L. citrata* (PNT_Lc) *or M. bombycina* (PNT_Mb).

3.2.4 Identification of candidate defense-related genes:

A. Antimicrobial peptides:

The Antimicrobial Peptide Database (APD) (http://aps.unmc.edu/AP/main.php) version 3 was downloaded and matched with the translated proteomes of both plant transcriptomes using blastp (e-value cut-off 1e-02) [33]. APDv3 consists of manually curated antimicrobial peptides from multiple species including plants. We also downloaded the reference plant AMP sequences from PhytAMP database and probed our transcriptomes using blastp similarly [34]. The existing annotations for these transcripts from Section 3.3.3 were compared to these hits and only the ones with better bit-score were retained. Physico-chemical parameters of the candidate peptides were generated using Prot-PARAM [35].

B. Glucosinolate-myrosinase biosynthetic system:

Reference gene identifiers for the glucosinolate biosynthetic pathway and glucosinolate activation pathway were downloaded from MetaCyc (under BioCyc) database and their protein sequences were retrieved from UniProtKB [36]. These genes were then matched with the translated proteomes of both plant transcriptomes using blastp (e-value cut-off 1e-02). Top hits per transcript were retained after sorting by query coverage, percentage identity and bit score. Expression of the transcripts were represented as TPM (Transcript per million bases) and the method of TPM derivation is described below.

3.2.5 Analysis of expression values (TPM) for transcripts:

Abundance of each transcript was estimated using an alignment-based mapping tool, RSEM, which aligns the quality control processed paired-end raw reads to the bowtie-indexed reference of a transcriptome [27,37]. This experiment generates normalized values of the count-based expression metric for each transcript (also termed as isoforms), namely, 'transcripts per million' (TPM). Normalization removes technical biases inherent in sequencing approaches, most notably the length of the RNA species and the sequencing depth of a sample. The formula for TPM is-

$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T}$$ where $$T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

Here, r_g is the number of reads mapped to a particular transcript g;

fl_g is feature length, the number of nucleotides in a mappable region of the transcript;

rl is the read length, i.e., the average number of nucleotides mapped per read

T represents the total number of transcripts sampled in a sequencing run.

3.3 RESULTS AND DISCUSSION

3.3.1 *De novo* transcriptome assembly

3.3.1.1 *M. bombycina*: Sequencing of the *M. bombycina* transcriptome generated 73,798,878 paired-end raw reads of which 69,820,812 were retained post-quality control steps for the de novo assembly. N50 statistic is popularly used to assess a transcriptome assembly's quality, with a higher value indicating a better assembly. Our observed N50 value (1930) for *M. bombycina* transcriptome was higher than those obtained in multiple other plant

transcriptome sequencing studies which indicated the robustness of the assembly (Table-3.2) [38–40]. Read coverage of the assembly was estimated to be 97.89% indicating that majority of the reads were utilized for the assembly. We also estimated the percentage of BUSCO orthologues from embryophyta present in *M. bombycina* transcriptome and found it to be satisfactory (>80% of complete BUSCOs) (Fig. 3.1).

Table 3.2 Transcriptome assembly statistics for *M. bombycina*

Number of Transcripts	55400
Percent GC	43.59
Contig N50 value	1930
Median contig length	876
Average contig	1228.55
Total assembled bases	68061887
Percentage of reads covered in the assembly	97.89

3.3.1.2 *L. citrata:* Sequencing of the *L. citrata* transcriptome generated 25,015,009 paired-end raw reads of which 23,286,733 were retained post-quality control steps for the de novo assembly. These high-quality reads were taken as input for transcriptome assembly using Trinity. The resulting assembly statistics are provided in Table 3.3. Our observed N50 value (1722) for *L. citrata* transcriptome was higher than those obtained in multiple other plant transcriptome sequencing studies which indicated the robustness of the assembly [38–40]. Read coverage of the assembly was very high (96.73%). We also estimated the percentage of BUSCO orthologues from embryophyta present in *L.*

citrata transcriptome and found it to be satisfactory (>80% of complete BUSCOs) (Fig. 3.1).

Table 3.3 Transcriptome assembly statistics for *L. citrata*

Number of Transcripts	138690
Percent GC	41.78
Contig N50 value	1722
Median contig length	526
Average contig	968.19
Total assembled bases	134278470
Percentage of reads mapped	96.73 %

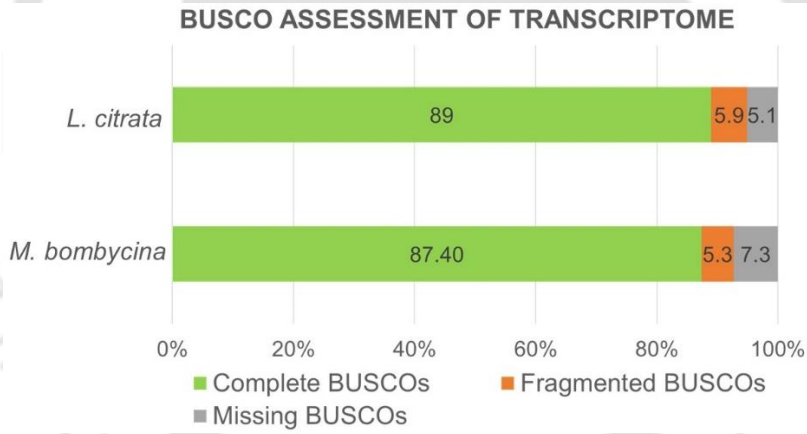


Fig. 3.1 Percentage of BUSCOs present in the transcriptomes of *M. bombycina* and *L. citrata*

3.3.2 Functional annotation and classification of the transcriptome

3.3.2.1 *M. bombycina*

The purpose of annotating any de novo transcriptome is to assign an identity to each transcript. Here, a combination of databases (NCBI Protein, UniProt, Pfam,

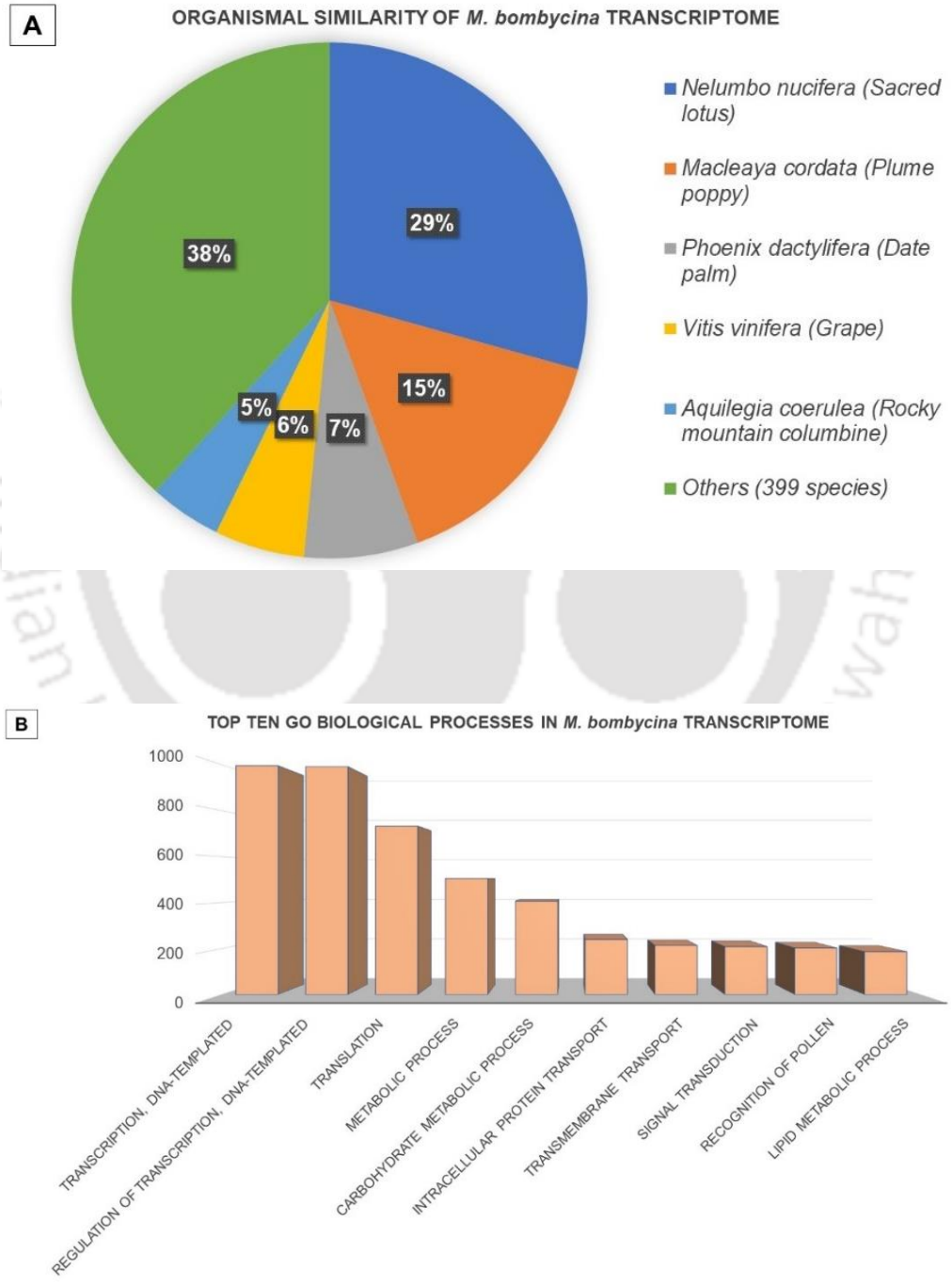
GO and KEGG) were strategically selected to ensure annotation of maximum number of transcripts from the *M. bombycina* transcriptome. We were able to annotate 39,198 out of 55,400 transcripts in totality while 16,202 transcripts remained unannotated and were classified as PNTMb (Annotation summary in Table 3.4). Presence of a reference genome is usually helpful in achieving a high percentage of transcriptome annotation. However, given the lack of genomic information in the current scenario, this study succeeded in assigning putative identifications to a significant proportion (70.75%) of *M. bombycina* transcripts.

Table 3.4 Annotation summary for *M. bombycina*

Total number of transcripts annotated	39198
Number of transcripts with GO classification	39063
Number of transcripts with Pfam domains	33801
Number of transcripts with KO (KEGG Orthology) identifier	7144
Putative novel transcript of *M. bombycina* (PNTMb)	16202
Total number of transcripts	55,400

Top species distribution of the blast output of *M. bombycina* transcriptome identified *Nelumbo nucifera* (Sacred lotus) from the Nelumbacaea family as the species with largest number of blast hits (Top five similar organisms shown in Fig. 3.2A). We further demarcated the Gene Ontology (GO) categories, namely, biological processes (BP) and molecular functions (MF), with the highest frequency. BPs represent large cellular processes like DNA repair or signal transduction which are accomplished by multiple molecular activities while MFs represent these molecular activities which correspond to entities (individual or

complexes) like “protein kinase activity”. Among GO_BPs, the transcriptome of *M. bombycina* had a higher abundance of transcription, translation and metabolic processes than others (Fig. 3.2B). In GO_MFs, ATP-binding activity consisted of more than four-fold transcripts than the successive over-represented functions like protein kinase, DNA or metal ion binding activity, etc.



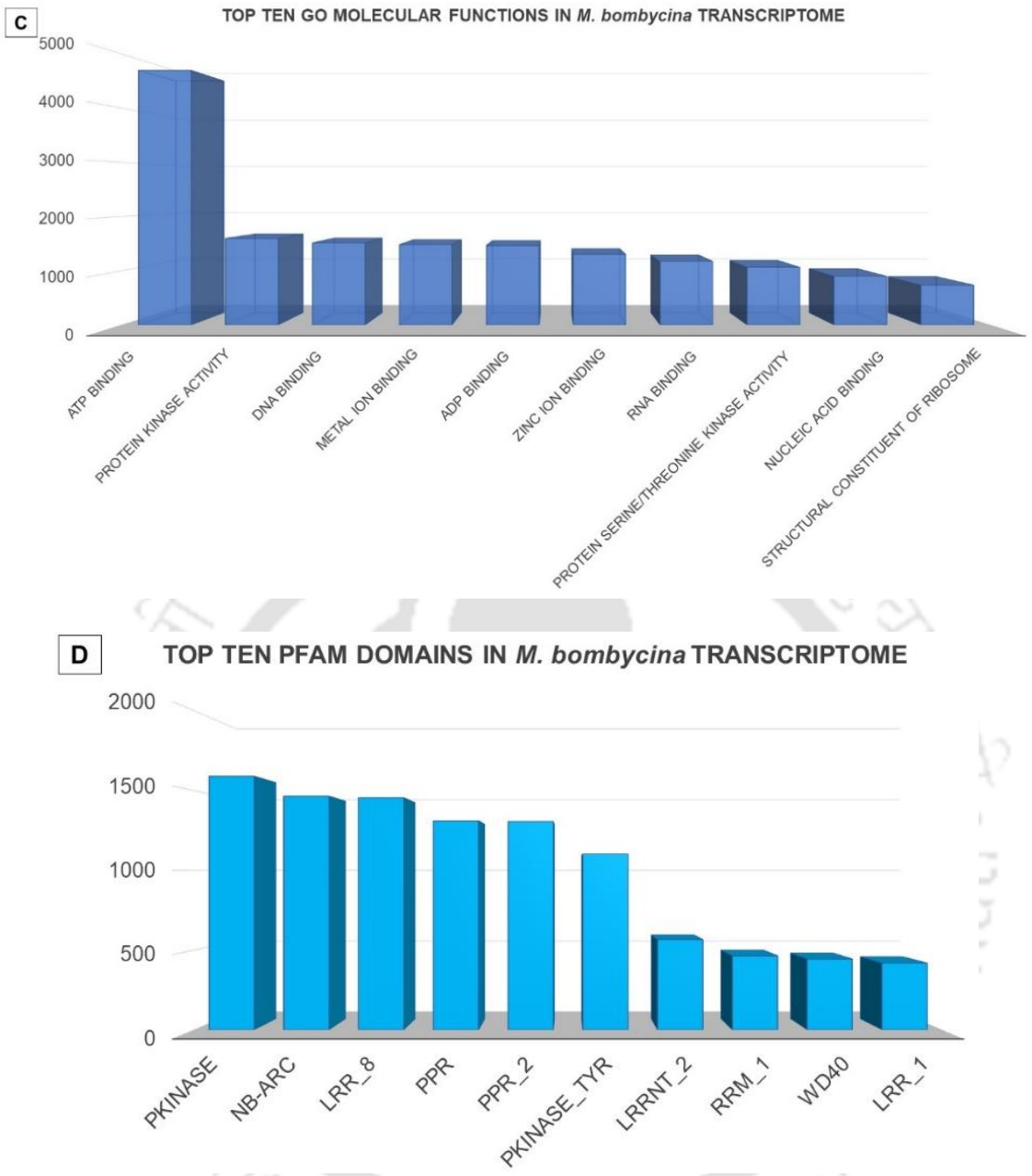


Fig 3.2 Annotation of *M. bombycina* transcriptome [A- Top five organismal similarities of transcripts; B- Top ten Gene Ontology (GO) Biological Processes (BP); C- Top ten GO Molecular Functions (MF); D- Distribution of top ten protein families in PFAM; Abbreviations- Protein Kinase (PKINASE), Nucleotide-binding ARC (NB-ARC), Leucine-rich repeat families (LRR 1, 2 & 8), Pentatricopeptide repeat families (PPR and PPR_2), Tyrosine kinase (Pkinase_Tyr) and RNA recognition motif (RRM_1)]

Pfam-based annotation of *M. bombycina* transcriptome revealed the relative abundance of Protein kinase (Pkinase), Nucleotide-binding ARC (NB-ARC), Leucine-rich repeat families (LRR 1, 2 & 8), Pentatricopeptide repeat families (PPR and PPR_2), Tyrosine kinase (Pkinase_Tyr), RNA recognition motif (RRM_1) and WD40 above others (Fig. 3.2D). Pkinase activity phosphorylates proteins at Serine (S), Threonine (T) and Tyrosine (Y) residues, a modification that acts as a regulatory switch for a multitude of molecular processes. Pkinase_Tyr transfers phosphate-residues from donors like ATP to S-residues on proteins. The protein kinase-domain containing proteins are widely implicated to play the role of switches in cellular mitosis, cytokinesis, metabolic signaling, stress response and cellular death in plant cells [41]. Their relatively greater abundance in plants have also been related to the higher rate of genome duplications and polyploidy [41]. LRR motifs orchestrate formation of solenoid domains by folding of hydrophobic Leucine (L) repeats and are often reported to be associated with protein-protein interactions related to host defense, for e.g. toll-like receptors (TLRs) and binding with pathogen and danger associated molecular patterns (PAMPs and DAMPs) respectively [42]. PPR proteins are a class of RNA-binding proteins found in relatively higher abundance in plant kingdom. These were identified while screening for proteins targeted at sub-cellular organelles of *Arabidopsis thaliana* and are presumed to be involved in organellar post-transcriptional processes like splicing, editing, processing and translation of RNA [43]. RRM_1 domain containing proteins are the largest class of eukaryotic ssRNA-binding proteins and are also involved in regulating post-transcriptional modifications, for e.g. alternative splicing, among other putative function in plants [44].

Finally, WD40 domains are structural domains rich in Tryptophan (W) and Aspartic acid (D) residues. They are involved in a wide range of biological activities like regulation of cell cycle, transcription, apoptosis, etc. in eukaryotic cells and owing to their underlying specialty to coordinate the assembly of multi-protein complexes [45]. Comparison of the top GO_BP and Pfam domains in *M. bombycina* revealed that a greater share of the genomic resources is dedicated towards regulation of transcription and post-transcriptional processes. Our observation reverberates the evolving idea of non-linearity of the central dogma with a significant role of post-transcriptional regulation in maintaining the flow of accurate genetic information.

Other than the annotated transcripts, a percentage of transcripts from *M. bombycina* (PNT_Mb) remained unannotated. These transcripts may consist of novel proteins, non-coding RNAs, chimeric transcripts or simply, products of spurious transcription. A high-coverage whole genome assembly of *M. bombycina* will be better suited in resolving their identities.

3.3.2.2 *L. citrata*

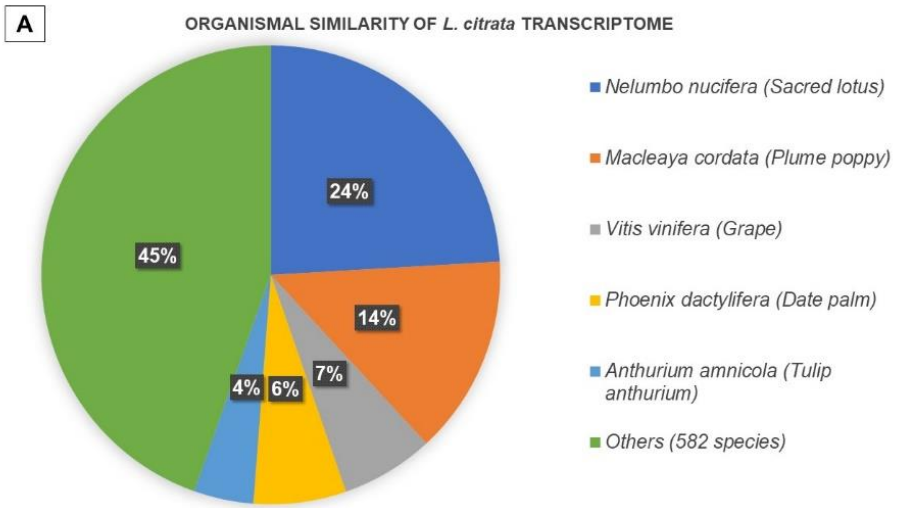
L. citrata transcriptome was similarly annotated using a combination of databases (NCBI Protein, UniProt, Pfam, GO and KEGG). We were able to annotate 61,586 out of 1,38,690 transcripts while 77,104 of the transcripts remained unannotated and were classified as PNT_Lc (Annotation summary in Table 3.5). Similar to *M. bombycina,* we were able to assign putative identities to a significant proportion (~45%) of *L. citrata* transcripts.

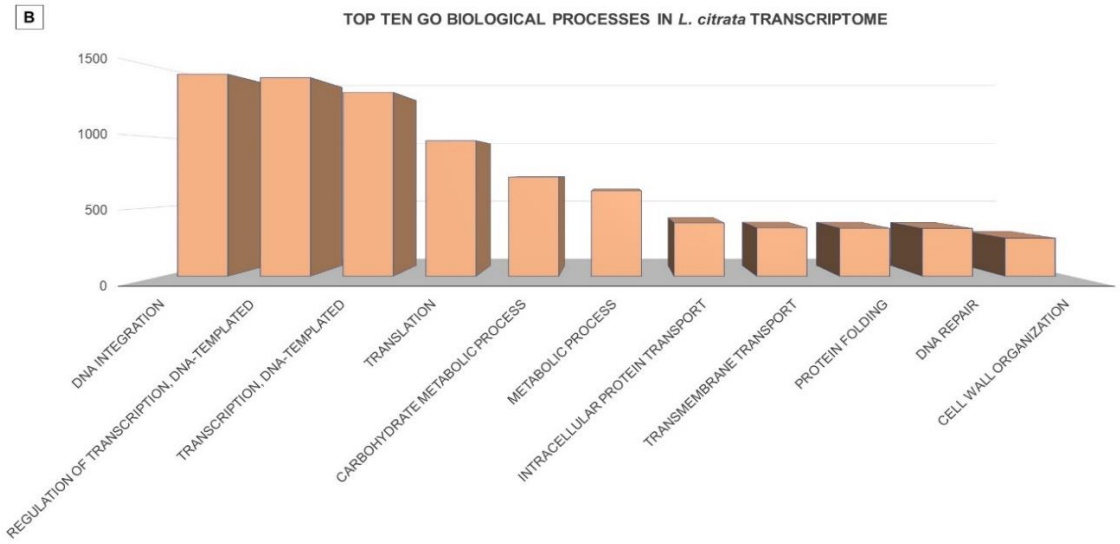
Table 3.5 Annotation summary for *L. citrata*

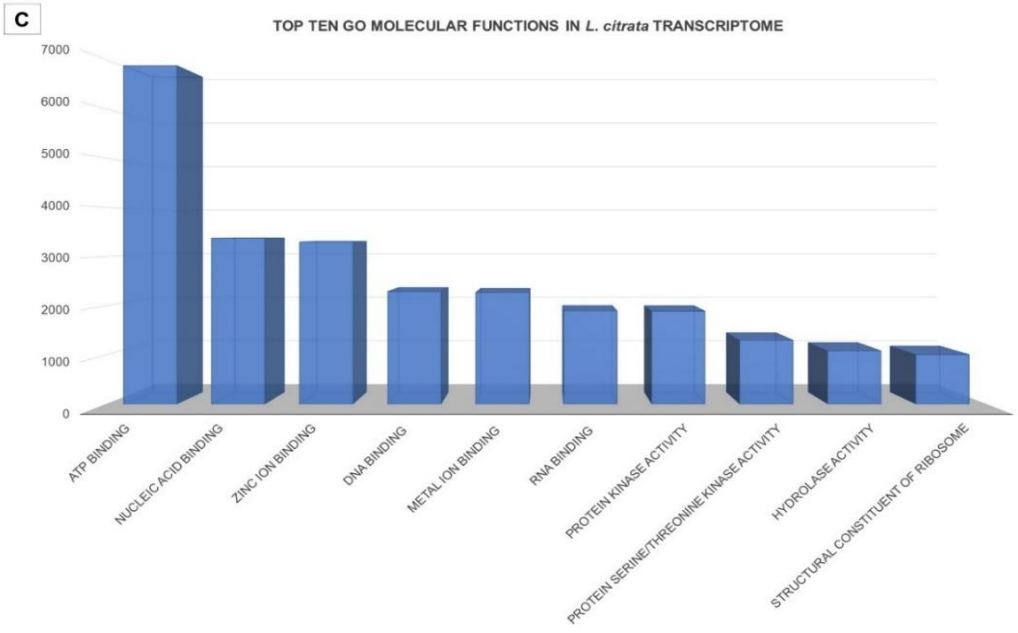
Total number of transcripts annotated	61586
Number of transcripts with GO classification	45282
Number of transcripts with Pfam domains	52587
Number of transcripts with KO (KEGG Orthology) identifier	9832
Putative novel transcript in *L. citrata*	77104
Total number of transcripts	138690

Top species distribution of the blast output of *L. citrata* transcriptome also identified *Nelumbo nucifera* (Sacred lotus) from the Nelumbacaea family as the species with largest number of blast hits (Top five similar organisms shown in Fig. 3.3A). Further demarcation of annotated transcripts into GO_BP and GO_MF categories was carried out. Among GO_BPs, the transcriptome of *L. citrata* had a higher abundance of primary information processing pathways, viz. transcription and translation, DNA integration as well as molecular metabolism (Fig. 3.3B). Among top GO_MFs, ATP-binding activity was again the most abundant function in *L. citrata* (Fig. 3.3C). However, the following molecular functions were more abundant in this host plant in comparison to the Som plant, namely, nucleic acid, DNA, metal ion, zinc ion and RNA binding activity etc.

PKINASE, PPR, Pkinase_Tyr, LRR and RRm_1 were amongst the top ten most abundant functional domains in the *L. citrata* transcriptome (Fig. 3.3D). Their functional significance has been discussed in the previous section pertaining to *M. bombycina* (Section 3.3.2A). However, three domains, namely, reverse transc-


A
ORGANISMAL SIMILARITY OF L. citrata TRANSCRIPTOME
24%
14%
7%
6%
4%
45%
Nelumbo nucifera (Sacred lotus)
Macleaya cordata (Plume poppy)
Vitis vinifera (Grape)
Phoenix dactylifera (Date palm)
Anthurium amnicola (Tulip anthurium)
Others (582 species)


B
TOP TEN GO BIOLOGICAL PROCESSES IN L. citrata TRANSCRIPTOME
1500
1000
500
0
DNA INTEGRATION
REGULATION OF TRANSCRIPTION, DNA-TEMPLATED
TRANSCRIPTION, DNA-TEMPLATED
TRANSLATION
CARBOHYDRATE METABOLIC PROCESS
METABOLIC PROCESS
INTRACELLULAR PROTEIN TRANSPORT
TRANSMEMBRANE TRANSPORT
PROTEIN FOLDING
DNA REPAIR
CELL WALL ORGANIZATION


C
TOP TEN GO MOLECULAR FUNCTIONS IN L. citrata TRANSCRIPTOME
7000
6000
5000
4000
3000
2000
1000
0
ATP BINDING
NUCLEIC ACID BINDING
ZINC ION BINDING
DNA BINDING
METAL ION BINDING
RNA BINDING
PROTEIN KINASE ACTIVITY
PROTEIN SERINE/THREONINE KINASE ACTIVITY
HYDROLASE ACTIVITY
STRUCTURAL CONSTITUENT OF RIBOSOME

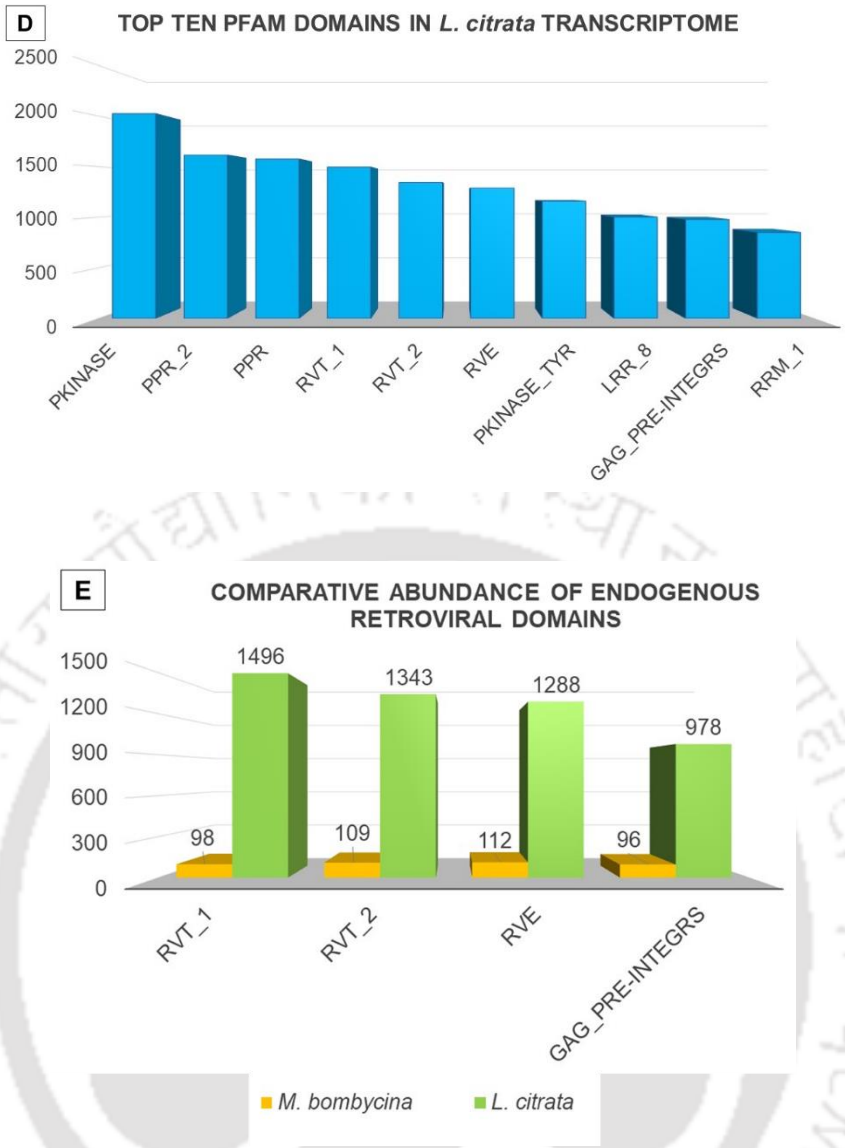


Fig. 3.3 Annotation of *Litsea citrata* transcriptome [A- Top five organismal similarities of transcripts; B- Top ten GO Biological Processes (BP); C- Top ten GO Molecular Functions (MF); D-Distribution of top ten Pfam domains (Abbreviations- Protein Kinase (PKINASE), Pentatricopeptide repeat families (PPR and PPR_2), Reverse transcriptase (RVT 1 & 2), Integrase (RVE), Tyrosine kinase (Pkinase_Tyr), Leucine-rich repeat families (LRR 8), Gag pre-integrase (GAG_PRE-INTEGRS) and RNA recognition motif (RRM_1)); E- Comparative abundance of the endogenous retroviral domains in *M. bombycina* and *L. citrata*]

-riptase (RVT_1 & 2), integrase (RVE) and GAG-pre-integrase (GAG_PRE-INTEGRS) were found to be abundant in *L. citrata* and not in *M. bombycina.* RVT, as its full name suggests, are domains that help in reverse transcribing RNA into cDNA. While these domains are predominantly present in retroviral polymerases, they are also not uncommon for a eukaryotic genome. One example is the eukaryotic enzyme, telomerase, which itself is a reverse transcriptase and is highly expressed during flowering stages [46]. Another fine example of eukaryotic RVT domain bearer is the group of the mobile self-replicating elements, retrotransposons, which can jump from one position of the genome to another via RNA transposition intermediates [47]. Our result indicated that *L. citrata* genome are probably enriched in retroelements, an observation bolstered by the relative abundance of two other retroelement-related functional domains, namely, GAG_PRE-INTEGRS and RVE. Both these domains are generally translatable elements of the retrotransposons, with the former domain contributing to formation of virus-like particles and the latter forming integrase enzyme that helps in the transposable cDNA integration into chromosome [47,48]. We further compared the abundance of these four domains of retroviral origin in both the muga silkworm host plants and found that *L. citrata* is relatively more enriched in these domains (Fig. 3.3E). Endogenous retroviral elements are usually extinct retroviruses that integrated into host chromosome post-infection and thereby, vertically transmitted with host genome [49]. While earlier plant retrotransposons were not considered to be descendants of infectious viral ancestors, characterization of the endogenous retroelements, *Gypsy* in *D. melanogaster, Athila4* in *A. thaliana* and *Calypso* in *Glycine max* (soybean) blurred this idea [49,50]. Presence of these elements in angiosperms including our

two species hints at their ubiquitous nature and possible role in plant genome evolution, thereby providing an interesting avenue for future evolutionary and functional studies [50]. Based on the position in the host genome, these elements can either be beneficial or harmful and mapping our putative retroviral elements into the whole genome of *L. citrata* and *M. bombycina* will provide better answers to their organismal role or lack thereof in near future. These whole genomes can also be used to map and annotate the putative PNT_LC transcripts identified here.

3.3.3 Antimicrobial peptides

Plant AMPs are a group of promising antimicrobial compounds that form a component of plant immunity. As discussed in the introductory section, these are present in all anatomical parts of the plant including leaves. We, hereby, sequenced the leaves of muga silkworm's host plants and identified a group of candidate AMPs in each species using the latest version of APD database (APD3) and PhytAMP [33,34].

3.3.3.1 *M. bombycina*

Blast-based alignment of known AMPs of APD3 and PhytAMP with *M. bombycina* transcriptome refined by applying the following cut-offs- percentage identity (≥30%) and bit-score (≥30) followed by de-duplicating the transcripts with more than single hit by retaining the best hit. We were, thereby, able to identify 224 candidate AMPs of 58 AMP types in *M. bombycina* (Fig. 3.4, Table 3.6). Out of these, 32 types were common between the two host plants while 26 were present in *M. bombycina* only.

Among the shared classes, anti-bacterial AMPs were present in higher frequency than AMPs with other activity. The most abundant class of AMPs was cgUbiquitin. It was originally reported in the pacific oyster, *Crassostrea gigas* and now, in other invertebrates, molluscs and animals. Ubiquitin and ubiquitin-like peptides usually regulate the signal transduction processes in immunity. This peptide, on the other hand, has shown bacteriostatic activity effective against both gram-positive and negative bacteria [51]. YFGAP or yellowfin tuna GAPDH-related AMP, identified in both *M. bombycina* and *L. citrata,* was another bacteriostatic AMP [52]. Four types of defensin were also shared among the host plants. Three of them were anti-bacterial, namely, BDEF_TACTR (also called Big defensin), Smd1 (also called *S. calcitrans* defensin 1) and e-NAP1 (or alpha defensin). All three have been reported in arthropods as well as mammals and have demonstrated activity against both gram-positive and negative bacteria [53–55]. The fourth type of defensin, namely, Fa-AMP1 is a hevein-type plant defensin isolated from buckwheat (*Fagopyrum esculentum Moench*) which has broad spectrum activity, both anti-bacterial and anti-fungal [56]. One more plant AMP, Hevein Pn-AMP1 was identified in both host plants. This AMP is also broad spectrum like Fa-AMP1 and works via membrane perturbation. Its expression has been related to pathogen-resistance in transgenic plants [57]. Penaeidin-3a is a Pro- and Cys-rich peptide with anti-bacterial (gram positive) and antifungal activity isolated from *Penaeus vannamei* [58]. Among the other shared multi-copy anti-bacterial peptides were saposin SP-BN (an anionic peptide isolated from *Mus musculus*), ubiquicidin (anti-bacterial as well as anti-MRSA (multi-drug resistant *Staphylococcus aureus*) peptide of mammalian origin and crustin (found in crustaceans and anti-gram positive bacteria only), [59–61]. Kinocidin CCL27, was

another shared AMP of mammalian origin with anti-fungal and chemotactic activity [62]. On the other hand, Griffithsin, an anti-viral and anti-parasitic lectin (named after its source- the red alga *Griffithsia* sp.) was the second most abundant class of AMPs in Som [63]. It is a potent inhibitor of viral entry and its activity has been demonstrated against parasites as well as multiple virus strains including HIV [64]. Using PhytAMP, we were also able to identify additional plant-specific AMPs, namely, vicilin-like (isolated from *Petunia hybrida;* has antibacterial and antifungal activity), Kalata-B2 (isolated from *Euonymus europaeus;* can form membrane pores and has insecticidal as well as pharmaceutical activity), thionin (isolated from *E. europaeus;* cystine-rich, cationic small peptides) and unnamed AMPs from *Triticum aestivum* and *Oryza sativa* [65–67].

TABLE 3.6 Distribution of the classes of Antimicrobial peptides (AMPs) in *M. bombycina* (Som) *(Classes common between Som and Mejankari are depicted in italics; UniProt IDs are shown for PhytoAMP peptides)*

ID	AMP CLASS	COUNT	ID	AMP CLASS	COUNT
AP02030	*cgUbiquitin*	*83*	*AP02146*	*ALFpm3*	*1*
AP02133	*Griffithsin*	*7*	*AP01587*	*Defensin*	*1*
AP02012	*YFGAP*	*7*	*AP01365*	*Defensin*	*1*
AP01238	*Defensin*	*5*	*AP01245*	*Cy-AMP3*	*1*
AP00270	*Hevein*	*5*	*AP01157*	*Ixodidin*	*1*
AP01576	*Crustin*	*4*	*AP01068*	*Cycloviolacin*	*1*
AP01540	*Saposin*	*4*	*AP00997*	*Bacteriocin*	*1*
AP00394	*Penaeidin-3a*	*4*	*AP00489*	*Hipposin*	*1*

ID	AMP CLASS	COUNT	ID	AMP CLASS	COUNT
AP02187	*Kinocidin*	*2*	*AP00294*	*Defensin*	*1*
AP02096	*Ubiquicidin*	*2*	*O24006*	*Antimicrobial peptides*	*1*
P82010	*Defensin-like protein AX2*	*1*	*Q9SPL5*	*Vicilin-like antimicrobial peptides 2-1*	*2*
Q8H6Q1	*Floral defensin-like protein*	*3*	*Q9SPL4*	*Vicilin-like antimicrobial peptides 2-2*	*4*
Q7Y238	*Hevein-like*	*9*	AP00308	Buforin II	10
Q7X9R9	*Hevein-like*	*2*	AP01520	Stylicin	4
Q8H950	*Hevein-like protein OS*	*2*	AP00915	Ee-CBP	3
P58454	*Kalata-B2 OS*	*2*	AP01234	Lumbricin	3
Q8LT03	*Leaf thionin Asthi1 OS*	*1*	AP00023	Antiviral	2
Q42952	*Non-specific lipid-transfer protein 1*	*1*	AP00395	Penaeidin-4a	2
Q9ATG4	*Non-specific lipid-transfer protein*	*1*	AP00807	Enterocin	2
Q5Z5V1	*Putative thionin Osthi1*	*2*	AP01341	Naegleriapore	2

ID	AMP CLASS	COUNT	ID	AMP CLASS	COUNT
AP02244	PaSn	2	P80915	Antimicrobial peptide 1	2
AP02330	Abaecin	2	Q9FR52	Antimicrobial peptide shep-GRP	3
AP00621	Palustrin-3a	1	P83399	Defensin-like protein 1	1
AP00806	Microplusin	1	P82784	Defensin-like protein 7	1
AP01164	EAFP1	1	P82782	Defensin-like protein 8	1
AP01469	Hevein	1	Q8LT02	Leaf thionin Asthi2	1
AP02157	Gloverin	1	A0AT29	Non-specific lipid-transfer protein 2	2
P32032	Alpha-2-purothionin	7	P10973	Non-specific lipid-transfer protein A	4
Q5UNP2	Non-specific lipid-transfer protein	1	P19656	Non-specific lipid-transfer protein	1

26 AMP types were exclusive to *M. bombycina* of which four were broad-spectrum. First of them is Buforin II which was isolated from *Bufo bufo gargarizans* and has a strong affinity for nucleic acids which confers it the ability

to inhibit bacterial and fungal cells [68]. Stylicin or Ls-Stylicin1, Penaeidin-4a or PEN4a also have both antibacterial and antifungal properties and were found in Som only [58,69]. Ee-CBP (*Euonymus europaeus* chitin-binding protein) is another anti-fungal, anti-gram positive peptide isolated from spindle tree *E. europaeus* with multiple disulfide bridged three-dimensional structure [70]. Lumbricin I is broad-spectrum and constitutively expressed in some invertebrates [71]. Other than these, anti-bacterial AMP classes like Enterocin E-760 (also bacteriocin), abaecin and PaSn (*Persea Americana* snakin) were also present [72–74]. Two candidate Naegleriapore B (saposin-like) AMPs, originally from a highly virulent protozoon pathogen *Naegleria fowleri,* were also identified. These peptides have membrane permeabilizing capability, thus making it cytolytic and tissue-lytic, active against human and bacterial cells [75]. Finally, homolog of the antiviral peptide, Antiviral protein Y3 isolated from edible fungi, *Coprinus comatus*, was identified. AP-Y3 has demonstrated both antiviral (against Tobacco mosaic virus) as well as anticancer (stomach tumor cells *in vitro*) activity [76]. Some non-specific lipid-transfer proteins were also exclusively found in Som. Plant nsLTPs are basic peptides involved in key cytological processes, stress resistance as well as antifungal, antiviral and antibacterial properties [77]. There were other single copy AMPs that were either common or unique to *M. bombycina*, for e.g. bacteriocin, hipposin, gloverin, etc. which were predominantly anti-bacterial.

3.3.3.2 *L. citrata*

Blast-based alignment of known AMPs of APD3 with *L. citrata* transcriptome were refined by applying the following cut-offs- percentage identity (≥30%) and bit-score (≥30) followed by de-duplicating the transcripts with more than single hit by

retaining the best hit. We were, thereby, able to identify 249 candidate AMPs of 53 AMP types in *L. citrata* (Fig. 3.5, Table 3.7). Most of the common AMP classes have been discussed in the previous section other than ALFpm3, Cy-AMP3 and Ixodidin. ALFpm3 (Anti-lipopolysaccharide factor 3 of *Penaeus monodon* or the back-tiger shrimp) is the versatile AMP discussed till now with activities ranging from bacteria (both gram-positive and negative), antiviral as well as antifungal [78]. Mejankari has two homologs of this peptide while Som has one. The other AMP, Cy-AMP3 is an antibacterial as well as antifungal AMP originally isolated from cycad seeds (*Cycas revoluta)* [79]. Finally, Ixodidin was originally isolated from Asian blue cattle tick (*Boophilus microplus)* and is effective against both gram-positive and negative bacteria [80].

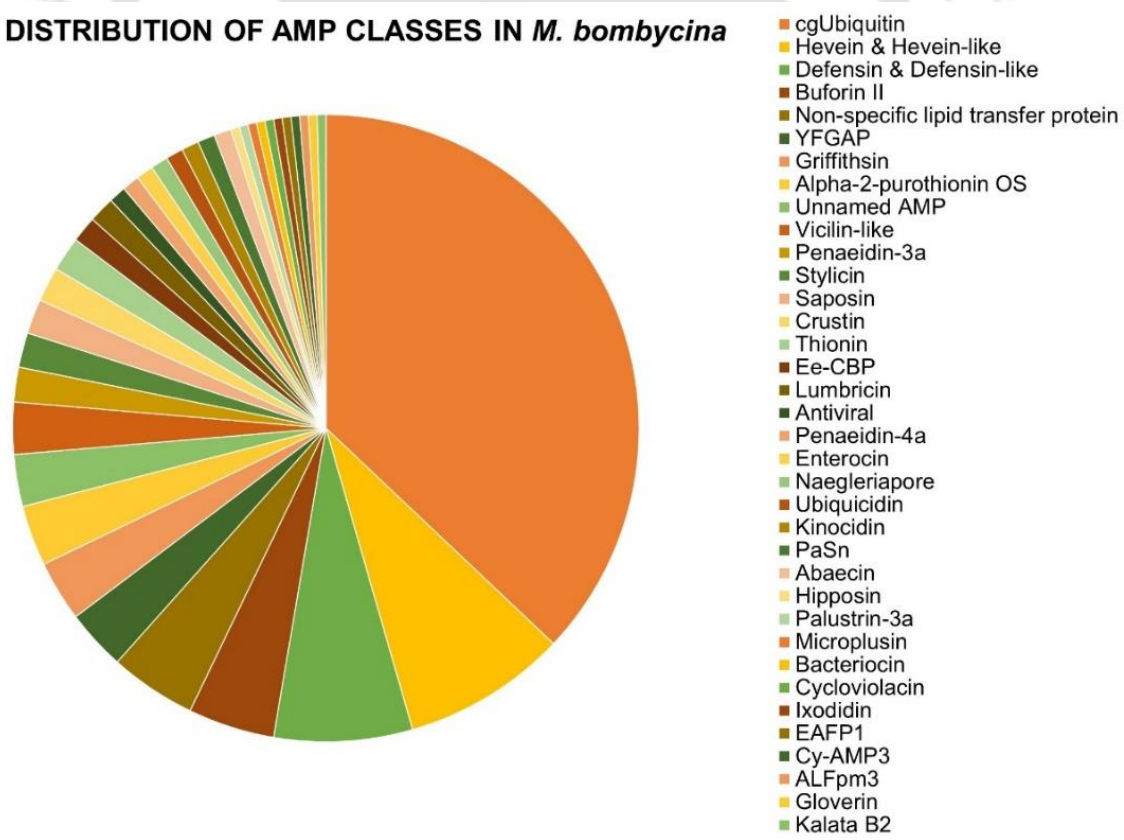


Fig. 3.4 Percentage distribution of the Antimicrobial peptide (AMP) classes in *Machilus bombycina* (Som)

TABLE 3.7 Distribution of the classes of Antimicrobial peptides (AMPs) in *Litsea citrata* (Mejankari) *(Classes common between Som and Mejankari are depicted in italics; UniProt IDs are shown for PhytoAMP peptides)*

ID	AMP CLASS	COUNT	ID	AMP CLASS	COUNT
AP02030	*cgUbiquitin*	*100*	*AP00294*	*Defensin*	*2*
AP00489	*Hipposin*	*13*	*AP01157*	*Ixodidin*	*2*
AP00394	*Penaeidin-3a*	*12*	*AP02146*	*ALFpm3*	*2*
AP02012	*YFGAP*	*12*	*AP02187*	*Kinocidin*	*2*
AP01540	*Saposin*	*6*	*AP00997*	*Bacteriocin*	*1*
AP01245	*Cy-AMP3*	*5*	*AP01068*	*Cycloviolacin*	*1*
AP00270	*Hevein*	*4*	*AP01238*	*Defensin*	*1*
AP02133	*Griffithsin*	*4*	*AP01365*	*Defensin*	*1*
AP01576	*Crustin*	*3*	*AP02096*	*Ubiquicidin*	*1*
AP01587	*Defensin*	*3*	*Q5Z5V1*	*Putative thionin Osthi1 OS*	*1*
O24006	*Antimicrobial peptide*	*1*	*Q7X9R9*	*Hevein-like antimicrobial peptide*	*1*
P58454	*Kalata-B2*	*4*	*Q7Y238*	*Hevein-like*	*9*
P82010	*Defensin-like protein AX2*	*3*	*Q9ATG4*	*Non-specific lipid-transfer protein*	*1*
Q42952	*Non-specific lipid-transfer protein 1*	*1*	*Q9SPL4*	*Vicilin-like*	*5*

ID	AMP CLASS	COUNT	ID	AMP CLASS	COUNT
Q8H6Q1	*Floral defensin-like protein 1*	*2*	*Q9SPL5*	*Vicilin-like*	*3*
Q8H950	*Hevein-like*	*5*	*Q8LT03*	*Leaf thionin Asthi1*	*4*
AP01153	Tachycitin	2	AP02094	Kinocidin	1
AP01282	Viscotoxin	2	AP02120	Nabaecin-3	1
AP02171	MrCrs	2	AP00753	Apolipophoricin	1
AP02070	RegIIIgamma	1	AP01819	Ac-AFP4	1
AP00355	Ginkobilobin	1	Q40901	Defensin-like protein	6
AP01186	Acidocin	1	Q43748	Non-specific lipid-transfer protein	4
AP01975	Silkworm AMP	1	Q8LSZ9	Thionin Asthi5	2
AP00495	Pseudo-hevein	1	A9NP30	Uncharacterized protein	2
AP01215	Abaecin	1	A0AT31	Non-specific lipid-transfer protein 5	1
AP02080	Kinocidin	1	Q9SBK8	Thionin	1
Q5USN7	Varv peptide A/Kalata-B1	1			

Mejankari transcriptome also had an exclusive set of AMPs (n=18) absent in Som. These included three different kinocidins; firstly, CXCL10 (a chemokine)

isolated from animals with a wide range of activities (anti-parasitic, anti-fungal, anti-bacterial and chemotactic) and secondly, CCL22 (also a chemokine) isolated from humans and with anti-bacterial and chemotactic activity [81,82]. Other putative *L. citrata* AMPs consisted of Tachytin (isolated from *Tachypleus tridentatus* with anti-bacterial and antifungal properties), MrCrs (*Macrobrachium rosenbergii* crustin) isolated from crustaceans which carries out bactericidal effect via agglutination) and Viscotoxin B or VtB (a type of thionin isolated from *Viscum album* with anti-cancer activity) [83–85]. Other than these, there were single copy AMPs either unique to *L. citrata*, for e.g. gingkobilobin, acidocin, abaecin etc. with predominantly anti-bacterial or anti-fungal activity [86]. We were also able to identify homologs of some nsLTPs, defensin-like, hevein-like, Kalata and thionins in Mejankari that were different than those present in Som.

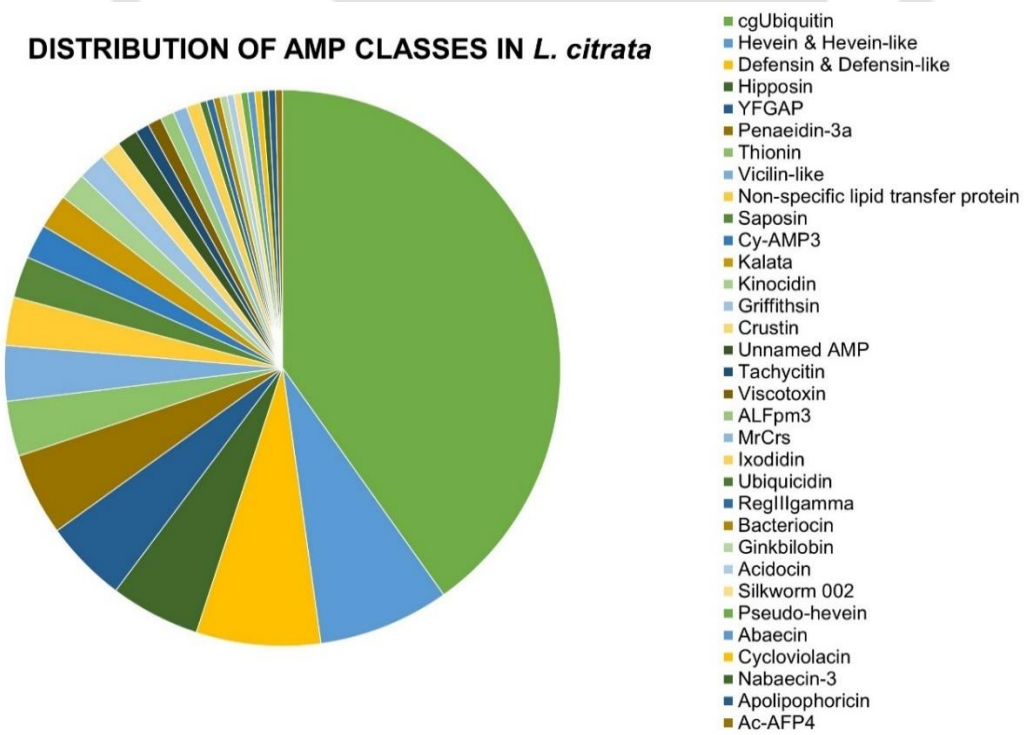


Fig. 3.5 Percentage distribution of the antimicrobial peptide (AMP) types in *Litsea citrata* (Mejankari)

3.3.3.3 General physicochemical properties of the candidate antimicrobial peptides:

The general physicochemical parameters of AMPs, like hydropathicity, determines their structural stability and mechanism of action. For instance, presence of Cys-residues ensures better stability of the three-dimensional structure of the peptide. AMPs are also classified with respect to most of these features. Here, we also generated the measurable physicochemical properties of the AMPs based on their primary information, i.e., sequence. We analyzed the stretches of peptides from our predicted proteome that aligned with the reference AMPs from APD3 and PhytAMP for this procedure. Based on the results provided by Prot-PARAM, we garnered the following information of the putative AMPs in *M. bombycina* and *L. citrata*- total peptide length, number of Cys-, Pro- and Arg-residues, number of positively and negatively charged residues as well as the grand average of hydropathicity (GRAVY) value [35].

Our analysis showed that the average length of AMPs in our both species were ~51 respectively which is in the range of conventional AMPs reported earlier [87]. Roughly 3 residues of Cys, Pro and Arg-residues were present on an average. Presence of more than 2 Cys-residues draws the possibility of at least one 1 disulfide bridge per peptide and as discussed above, presence of a S-S bridge reflects upon the peptide's structural stability. Here, the putative defensins and saponins had much higher ratio of Cys than the other AMPs. On the other hand, Pro- and Arg-residues also have a role in action mechanism of the peptides. Pro-richness, often with Arg-Pro-repeat motifs can entail bacterial membrane penetration via a non-lytic mechanism and act on cytoplasmic targets intracellularly [88]. Again, 5 and 9 positively and negatively-charged residues were

present on the peptides on an average, respectively. Distribution of positively and negatively charged residues in the AMPs are also necessary to have an idea about its mode of action, like cell entry and intracellular disruption or pore formation. This also impacts the average hydropathicity of any peptide. Finally, the average GRAVY score of the cumulative set of predicted AMPs in both host plants was -0.33. The negative score of GRAVY indicates that most of the candidate plant peptides of this study are cationic in nature.

In summary, our study identified many AMPs not only of plant-origin but also homologs of AMPs originating from other higher taxa like Arthropoda or Mammalia in *M. bombycina* and *L. citrata*. AMPs with anti-bacterial activity were the most abundant in both species followed by anti-fungal, anti-viral as well as other activities like cytotoxicity or chemotactic. Both species shared majority of the AMP classes, but still had an enriched and exclusive set of AMPs as well. These peptides will be further hosted at MugaSeqDB, a database developed for muga silkworm and its host plants, in near future (discussed in a chapter later). Overall, identification of AMPs in these two plants generated a resource for comprehensive investigation of their advanced physical, structural as well activity-related features and explore their potential uses. This will have implications on novel drug identification as well as genetic improvement of plants for pathogen resistance. It is pertinent to note that all the AMPs identified from *L. citrata* or *M. bombycina* are putative in nature and more biochemical as well as in vitro tests are necessary to establish to their predicted roles.

3.3.4 Glucosinolate-Myrosinase (Glc-Myr) induced herbivore defense

Glc-Myr system is a well-studied system of chemical defense in host plants, especially in crucifers. The plants synthesize glucosinolates which remain spatially confined and protected from their respective hydrolase, namely, myrosinase. When these two components come in contact due to physical damage, myrosinase catalyzes the hydrolysis of the glucosinolates into an array of secondary derivates of this parent compound [89]. One pathway, namely, the indole glucosinolate activation pathway, utilizes 7 additional enzymes to synthesize herbivore defense compounds (isothiocyanates and nitriles). These compounds are toxic to the herbivores and even lethal to some non-specialists.

The MetaCyc database enlists ten pathways for glucosinolate biosynthesis in plants out of which six originates with homomethionine and its polymers (di, tri, tetra, penta and hexa) (*MetaCyc Pathways Class: Glucosinolate Biosynthesis*). One pathway starts off with L-homomethionine and is termed as aliphatic glucosinolate biosynthesis with side-chain elongation cycle (Fig. 3.6A). The remaining three pathways have the aromatic amino acids (Phe, Tyr and Trp) as the starting material (Fig. 3.6B). Despite segregation into ten different pathways, these pathways share the enzymes catalyzing the intermediate reactions. We enlisted all the enzymes involved in these pathways and de-duplicated to create a reference pathway set consisting primarily of experimentally verified proteins (33 no.s) from *A. thaliana* and one each from *Sinapis alba* and *Brassica napus.* We used this reference sequence set consisting of 35 proteins to create a blastdb and blasted against the predicted proteomes of our target host plants, *M. bombycina and L. citrata.* Our blast hits were further sorted by query coverage (cut-off 30) and percentage identity (cut-off 30) as well as de-duplicated query-wise to retain the best hits. We identified 384 and 433 putative transcripts with

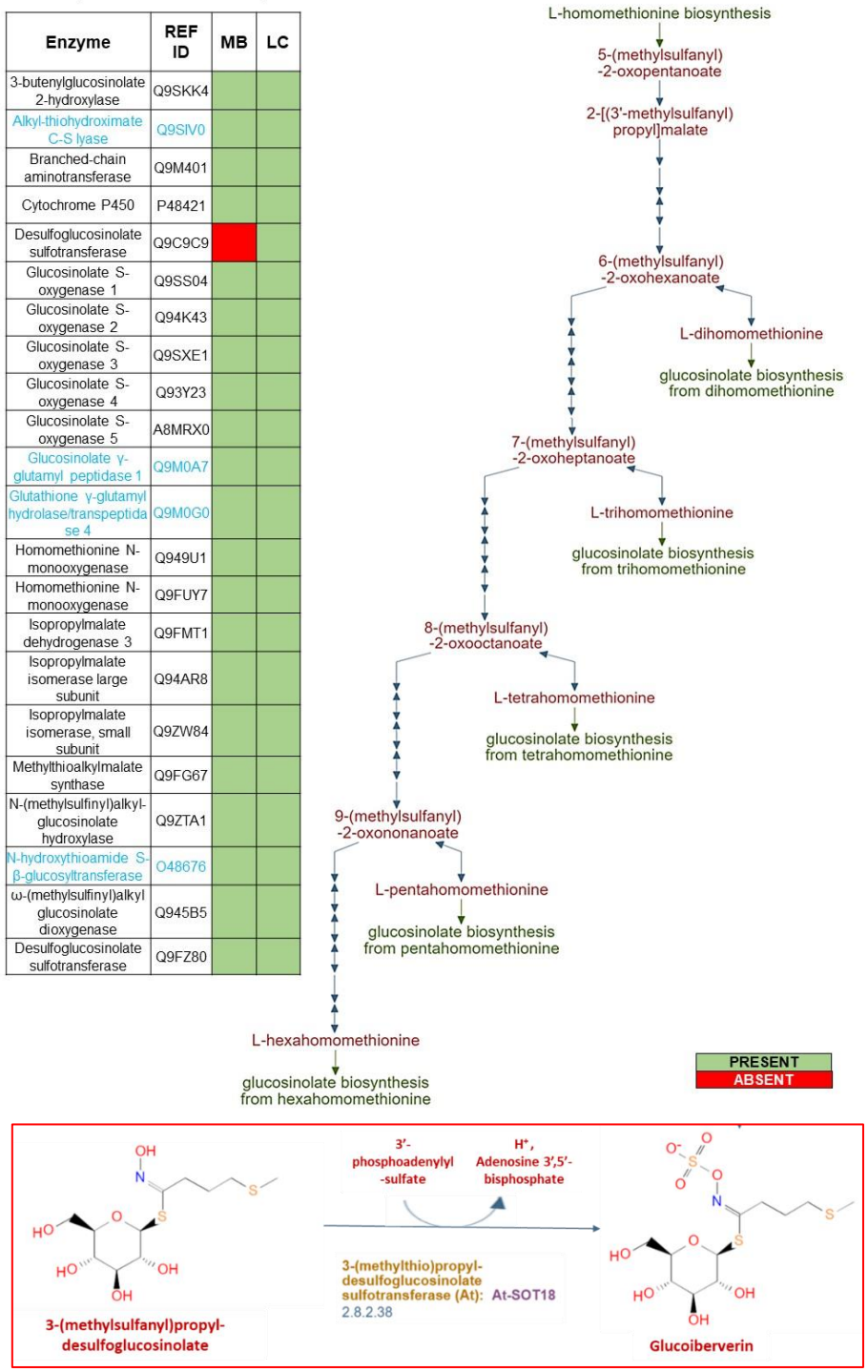
homology to these reference enzymes in Som and Mejankari. Similarly, we successfully identified myrosinase homologs in both plant transcriptomes (25 in Som and 19 in Mejankari) (Fig. 3.6C). Our analysis also indicated the presence of homologs of the seven genes involved in indole glucosinolate activation pathway (54 in Som and 98 in Mejankari) (Fig. 3.6C). We further calculated the relative expression of the top homologs from Som and Mejankari transcripts of the total set of reference genes (44) using RSEM using the methodology described earlier [37]. A heat map was drawn using the expression values (represented as log(TPM)) where higher the TPM value, greater was the expression of that transcript (Fig. 3.6D).

Overall, our study identified homologs for the 44-reference enzymes with the exception of Desulfoglucosinolate sulfotransferase SOT18 transcript in Som plant (Fig. 3.6A). In vitro studies conducted in *A. thaliana* has indicated that SOT17 and 18 preferably catalyze biosynthesis of long-chain desulfoglucosinolates derived from methionine while SOT16 prefer aromatic amino acid-derived desulfoglucosinolates [90]. Som has candidate transcripts for both SOT16 and SOT17. The presence of both SOT16 and SOT17 alongside other enzymes indicates that aromatic amino acid and long-chain aliphatic derived glucosinolate pathways may be active in the plant. In Mejankari, transcripts of all three isozymes (SOT16, 17 and 18) are present. Som SOT transcripts had greater TPM value than Mejankari (Fig. 3.6D). One must keep in mind that the comparison of transcript expressions here is a strictly preliminary estimation of a possible Glucosinolate-myrosinase pathway in both host plants. This is due to two reasons- firstly, both are outdoor-reared plants and secondly, this pathway is activated during herbivore attack and only in such a controlled experimental

scenario, we can gain a better comparative idea of the expression patterns. We were also obstructed by the fact these host plants whose whole genomes are not reported yet and so, the functionality of our candidate transcripts has to be experimentally tested.

A. BIOSYNTHESIS FROM L-HOMOMETHIONINE & ITS HOMOPOLYMERS (DI-, TRI-, TETRA-, PENTA- & HEXA-)

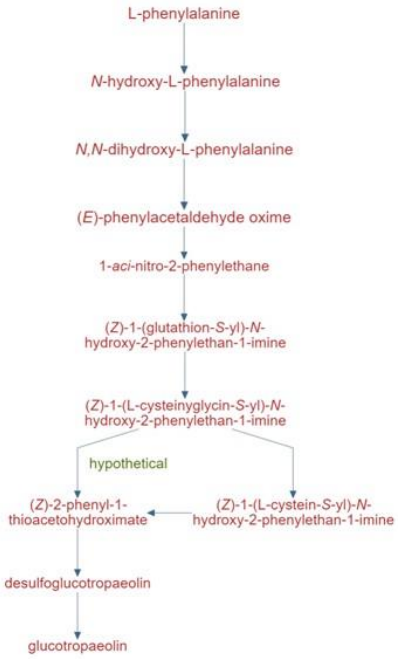
Enzyme	REF ID	MB	LC
3-butenylglucosinolate 2-hydroxylase	Q9SKK4		
Alkyl-thiohydroximate C-S lyase	Q9SIV0		
Branched-chain aminotransferase	Q9M401		
Cytochrome P450	P48421		
Desulfoglucosinolate sulfotransferase	Q9C9C9		
Glucosinolate S-oxygenase 1	Q9SS04		
Glucosinolate S-oxygenase 2	Q94K43		
Glucosinolate S-oxygenase 3	Q9SXE1		
Glucosinolate S-oxygenase 4	Q93Y23		
Glucosinolate S-oxygenase 5	A8MRX0		
Glucosinolate γ-glutamyl peptidase 1	Q9M0A7		
Glutathione γ-glutamyl hydrolase/transpeptidase 4	Q9M0G0		
Homomethionine N-monooxygenase	Q949U1		
Homomethionine N-monooxygenase	Q9FUY7		
Isopropylmalate dehydrogenase 3	Q9FMT1		
Isopropylmalate isomerase large subunit	Q94AR8		
Isopropylmalate isomerase, small subunit	Q9ZW84		
Methylthioalkylmalate synthase	Q9FG67		
N-(methylsulfinyl)alkyl-glucosinolate hydroxylase	Q9ZTA1		
N-hydroxythioamide S-β-glucosyltransferase	O48676		
ω-(methylsulfinyl)alkyl glucosinolate dioxygenase	Q945B5		
Desulfoglucosinolate sulfotransferase	Q9FZ80		



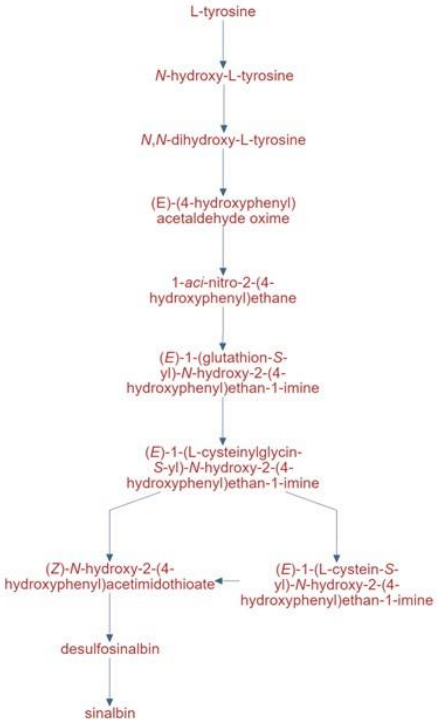
B. BIOSYNTHESIS FROM AROMATIC AMINO ACIDS

PRESENT
ABSENT

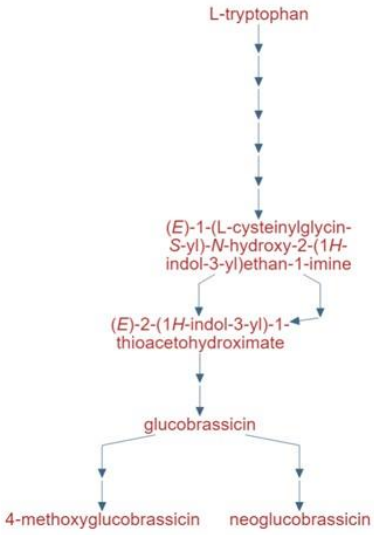
Phenyl Alanine



Tyrosine



Tryptophan

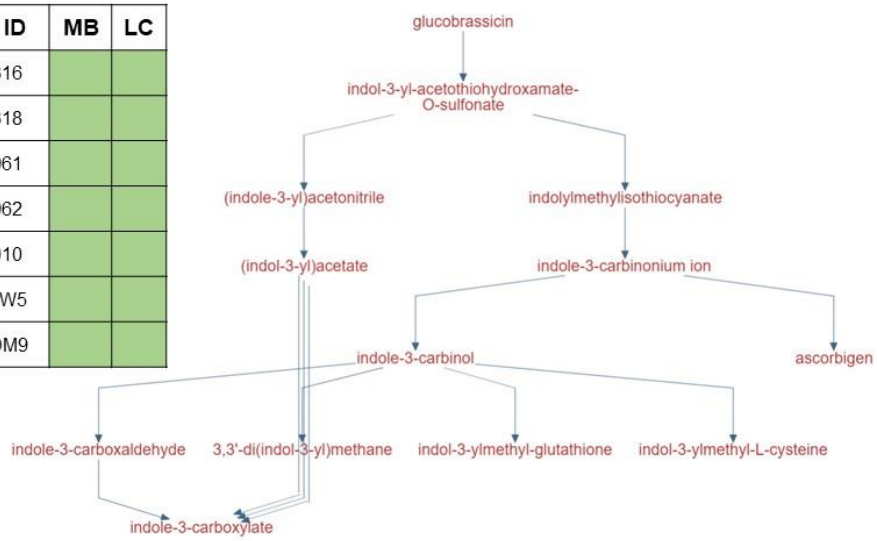


Enzyme	REF ID	MB	LC
Cytochrome p450	O65782		
Desulfoglucosinolate sulfotransferase SOT16	Q9C9D0		
Glucobrassicin monooxygenase	Q9LVD6		
Glucosinolate γ-glutamyl peptidase 3	Q9M0A5		
L-phenylalanine N-monooxygenase	Q9FLC8		
L-tryptophan N-monooxygenase	Q501D8		
L-tryptophan N-monooxygenase	O81346		
N-hydroxythioamide S-β-glucosyltransferase	Q947K4		
Tyrosine N-monooxygenase	O81345		
2-oxoglutarate-dependent dioxygenase AOP3	Q944X7		
Methylthioalkylmalate synthase 2	Q8VX04		
Methylthioalkylmalate synthase 3	Q9FN52		
2-oxoglutarate-dependent dioxygenase AOP2	Q9ZTA2		

C. INDOLE GLUCOSINOLATE ACTIVATION VIA MYROSINASE

Enzyme	REF ID	MB	LC
Nitrile-specifier protein 4	O04316		
Nitrile-specifier protein 3	O04318		
Nitrilase 1	P32961		
Nitrilase 2	P32962		
Nitrilase 3	P46010		
Nitrile-specifier protein 5	Q93XW5		
Nitrile-specifier protein 1	Q9SDM9		

PRESENT
ABSENT



(Fig. 3.6D and title on next page)

Despite these hurdles, our study sheds light into an herbivore-defense pathway in two Lauraceae family plants which are the hosts for an economical asset, the muga silkworm, but also source of naturally-derived oils and secondary metabolites of significance. Till date, this chemical defense pathway has been studied predominantly in Capparales and its family, Brassicaceae. The *de novo* transcriptomes of both host plants, hereby, provided a large set of candidate transcripts which can be probed experimentally in future.

3.4 CONCLUSION AND FUTURE PERSPECTIVES

In the present study, we assembled the transcriptomes of Som (*M. bombycina)* and Mejankari (*L. citrata)* for the first time, to the best of our knowledge. Despite the lack of a reference genome, we were able to assemble two robust transcriptomes and annotate as well as characterize ~50% of transcripts in both species. Putative transcripts homologous to key proteins involved in two

predominant modes of chemical defense- AMPs and Glc-myr system, were also intensively analyzed. Future studies can verify the function of these candidate proteins and apply this knowledge to develop pest-resistant muga host plants. Overall, these transcriptomes of these non-model plants also will be beneficial for basic and in-depth research on aspects relevant to sericulture like leaf age, yield, etc. for future studies.

D. HEAT MAP OF TOP TRANSCRIPTS HOMOLOGOUS TO THE PATHWAY GENES

	GENE_UNIPROT_ID	SOM (logTPM)	MEJANKARI (logTPM)
GLUCOSINOLATE BIOSYNTHESIS GENES	FMOGS-OX5_A8MRX0	1.093071306	0.424881637
	UGT74B1_O48676	1.692670699	1.442479769
	CYP83B1_O65782	0.57054294	1.303627976
	CYP79B1_O81345	1.160468531	-0.522878745
	CYP79B2_O81346	0.45331834	-0.795880017
	CYP83A1_P48421	0.57054294	1.303627976
	CYP79B3_Q501D8	1.160468531	-0.795880017
	MAM2_Q8VX04	1.581494542	0.271841607
	FMOGS-OX4_Q93Y23	1.093071306	-0.48148606
	AOP3_Q944X7	0.574031268	1.183839037
	AOP2_Q945B5	1.040997692	-0.657577319
	CYP79F1_Q949U1	1.160468531	-0.552841969
	IIL1_Q94AR8	1.862727528	1.16879202
	FMOGS-OX2_Q94K43	1.093071306	-0.48148606
	SOT18_Q9C9C9	NA	-0.494850022
	SOT16_Q9C9D0	0.662757832	0.365487985
	MAM1_Q9FG67	1.581494542	0.271841607
	CYP79A2_Q9FLC8	1.773493892	-0.522878745
	IMDH1_Q9FMT1	0.688419822	0.526339277
	MAM3_Q9FN52	1.581494542	0.271841607
	CYP79F2_Q9FUY7	1.180125875	0.699837726
	SOT17_Q9FZ80	0.662757832	-0.494850022
	CYP81F2_Q9LVD6	1.733678656	0.827369273
	GGP3_Q9M0A5	-0.031517051	1.113943352
	GGP1_Q9M0A7	-0.13667714	-0.397940009
	GGT3_Q9M0G0	1.02201574	0.311753861
	BCAT3_Q9M401	1.835500328	0.86923172
	SUR1_Q9SIV0	1.67614468	-0.387216143
	GSL-OH_Q9SKK4	1.048441804	-0.522878745
	FMOGS-OX1_Q9SS04	1.093071306	0.424881637
	FMOGS-OX3_Q9SXE1	0.703291378	-0.27572413
	AOP3_Q9ZTA1	0.574031268	-0.657577319
	AOP2_Q9ZTA2	0.489958479	0.303196057
	IPMI2_Q9ZW84	1.773493892	1.018284308
	Unrev_GlcTransf_Q947K4	1.692670699	1.442479769
INDOLE GLUCOSINOLATE ACTIVATION	BGLU26_O64883	0.691081492	-0.251811973
	NSP4_O04316	1.621902961	1.260786669
	NSP3_O04318	-0.229147988	-0.568636236
	NIT1_P32961	0.193124598	-0.602059991
	NIT2_P32962	0.491361694	0.255272505
	NIT3_P46010	-0.366531544	0.004321374
	NSP5_Q93XW5	0.57054294	-1
	NSP1_Q9SDM9	1.24723655	-0.48148606

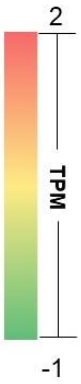


Fig. 3.6 Glucosinolate biosynthetic and activation pathway in *M. bombycina* and *L. citrata.* [A] Biosynthesis from aliphatic amino acids (the step catalyzed by SOT18 is shown in red border) and [B] Aromatic amino acids; [C] Indole glucosinolate activation via myrosinase enzyme; [D] Heatmap depicting log(TPM) values of the best candidate transcripts for each enzyme [Reference pathway diagram is sourced from MetaCyc; the enzymes shared between aliphatic and aromatic pathways are named in blue; Abbreviations used MB- *Machilus bombycina,* LC- *Litsea citrata,* FMOGS-OX5_A8MRX0 (Glucosinolate S-oxygenase 5), GT74B1_O48676 (N-hydroxythioamide S-β-glucosyltransferase), CYP83B1_O65782 Cytochrome p450 (CYP79B1_O81345 Tyrosine N-monooxygenase), CYP79B2_O81346 (L-tryptophan N-monooxygenase), CYP83A1_P48421 (Cytochrome P450), CYP79B3_Q501D8 (L-tryptophan N-monooxygenase), MAM2_Q8VX04 (Methylthioalkylmalate synthase 2), FMOGS-OX4_Q93Y23 (Glucosinolate S-oxygenase 4), AOP3_Q944X7 (2-oxoglutarate-dependent dioxygenase AOP3), AOP2_Q945B5 (ω-(methylsulfinyl)alkyl glucosinolate dioxygenase), CYP79F1_Q949U1 (Homomethionine N-monooxygenase), IIL1_Q94AR8 (Isopropylmalate isomerase large subunit), FMOGS-OX2_Q94K43 (Glucosinolate S-oxygenase 2), SOT18_Q9C9C9 (Desulfoglucosinolate sulfotransferase), SOT16_Q9C9D0 (Desulfoglucosinolate sulfotransferase SOT16), MAM1_Q9FG67 (Methylthioalkylmalate synthase), CYP79A2_Q9FLC8 (L-phenylalanine N-monooxygenase), IMDH1_Q9FMT1 (Isopropylmalate dehydrogenase 3) MAM3_Q9FN52 (Methylthioalkylmalate synthase 3), CYP79F2_Q9FUY7 (Homomethionine N-monooxygenase), SOT17_Q9FZ80 (Desulfoglucosinolate sulfotransferase), CYP81F2_Q9LVD6 (Glucobrassicin monooxygenase), GGP3_Q9M0A5 (Glucosinolate γ-glutamyl peptidase 3), GGP1_Q9M0A7 (Glucosinolate γ-glutamyl peptidase 1), GGT3_Q9M0G0 (Glutathione γ-glutamyl hydrolase/transpeptidase 4), BCAT3_Q9M401 (Branched-chain aminotransferase), SUR1_Q9SIV0 (Alkyl-thiohydroximate C-S lyase), GSL-OH_Q9SKK4 (3-butenylglucosinolate 2-hydroxylase), FMOGS-OX1_Q9SS04 (Glucosinolate S-

oxygenase 1), FMOGS-OX3_Q9SXE1 (Glucosinolate S-oxygenase 3), AOP3_Q9ZTA1 (N-(methylsulfinyl)alkyl-glucosinolate hydroxylase), AOP2_Q9ZTA2 (2-oxoglutarate-dependent dioxygenase AOP2), IPMI2_Q9ZW84 (Isopropylmalate isomerase, small subunit), Unrev_GlcTransf_Q947K4 (N-hydroxythioamide S-β-glucosyltransferase), BGLU26_O64883 (Myrosinase), NSP4_O04316 (Nitrile-specifier protein 4), NSP3_O04318 (Nitrile-specifier protein 3), NIT1_P32961 (Nitrilase 1), NIT2_P32962 (Nitrilase 2), NIT3_P46010 (Nitrilase 3), NSP5_Q93XW5 (Nitrile-specifier protein 5) and NSP1_Q9SDM9 (Nitrile-specifier protein 1)]

REFERENCES

1. Awmack CS, Leather SR. Host plant quality and fecundity in herbivorous insects. *Annu Rev Entomol*. 2002;47(1):817-844. doi:10.1146/annurev.ento.47.091201.145300.
2. War AR, Paulraj MG, Ahmad T, et al. Mechanisms of plant defense against insect herbivores. *Plant Signal Behav*. 2012;7(10):1306-1320. doi:10.4161/psb.21663.
3. Zhang L, Gallo RL. Antimicrobial peptides. Vol 26.; 2016. doi:10.1016/j.cub.2015.11.017.
4. Mithöfer A, Mithöfer M, Boland W. Plant defense against herbivores: chemical aspects. *Annu Rev Plant Biol*. 2012;63:431-450. doi:10.1146/annurev-arplant-042110-103854.
5. Campos ML, Lião LM, Alves ESF, Migliolo L, Dias SC, Franco OL. A structural perspective of plant antimicrobial peptides. *Biochem J*. 2018;475(21):3359-3375. doi:10.1042/BCJ20180213.
6. De Caleya RF, Gonzalez-Pascual B, García-Olmedo F, Carbonero P. Susceptibility of phytopathogenic bacteria to wheat purothionins in vitro. *Appl Environ Microbiol*. 1972;23(5).
7. Hancock RE., Lehrer R. Cationic peptides: a new source of antibiotics. *Trends Biotechnol*. 1998;16(2):82-88. doi:10.1016/S0167-7799(97)01156-6
8. Rashid War A, Kumar Taggar G, Hussain B, Sachdeva Taggar M, Nair RM, Sharma HC. Plant defense against herbivory and insect adaptations. *AoB Plants*. 2018;10(4). doi:10.1093/aobpla/ply037
9. Lambrix V, Reichelt M, Mitchell-Olds T, Kliebenstein DJ, Gershenzon J.

The Arabidopsis epithiospecifier protein promotes the hydrolysis of glucosinolates to nitriles and influences *Trichoplusia ni* herbivory. *Plant Cell*. 2001;13(12):2793-2807. doi:10.1105/tpc.010261.

10. Borek V, Elberson LR, McCaffrey JP, Morra MJ. Toxicity of rapeseed meal and methyl isothiocyanate to larvae of the black vine weevil (Coleoptera: Curculionidae). *J Econ Entomol*. 1997;90(1):109-112. doi:10.1093/jee/90.1.109.
11. Li Q, Eigenbrode SD, Stringam GR, Thiagarajah MR. Feeding and growth of *Plutella xylostella* and *Spodoptera eridania* on *Brassica juncea* with varying glucosinolate concentrations and myrosinase activities. *J Chem Ecol*. 2000;26(10):2401-2419. doi:10.1023/A:1005535129399.
12. Olivier C, Vaughn SF, Mizubuti ESG, Loria R. Variation in allyl isothiocyanate production within brassica species and correlation with fungicidal activity. *J Chem Ecol*. 1999;25(12):2687-2701. doi:10.1023/A:1020895306588.
13. Qi J, Malook S ul, Shen G, et al. Current understanding of maize and rice defense against insect herbivores. *Plant Divers*. 2018;40(4):189-195. doi:10.1016/J.PLD.2018.06.006.
14. Schenk PM, Kazan K, Wilson I, et al. Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. *Proc Natl Acad Sci U S A*. 2000;97(21):11655-11660. doi:10.1073/pnas.97.21.11655.
15. Bindroo BB, Singh NT, Sahu AK, Chakravorty R. Muga silkworm host plants. *Indian Silk*. 2006;44:13-17. http://mugadbase.com/pdf/Bindroo_et_al_2006.pdf.
16. Tikader A, Vijayan K, Saratchandra B. Muga silkworm, *Antheraea*

assamensis (Lepidoptera: Saturniidae) - an overview of distribution, biology and breeding. *Eur J Entomol*. 2013;110(2):293-300. doi:10.14411/eje.2013.096.

17. Arunkumar KP, Kifayathullah L, Nagaraju J. Microsatellite markers for the Indian golden silkmoth, *Antheraea assama* (Saturniidae: Lepidoptera). *Mol Ecol Resour*. 2009;9(1):268-270. doi:10.1111/j.1755-0998.2008.02414.x.
18. Ali JG, Agrawal AA. Specialist versus generalist insect herbivores and plant defense. *Trends Plant Sci*. 2012;17(5):293-302. doi:10.1016/j.tplants.2012.02.006.
19. Han X-J, Wang Y-D, Chen Y-C, Lin L-Y, Wu Q-K. Transcriptome sequencing and expression analysis of terpenoid biosynthesis genes in *Litsea cubeba*. Schönbach C, ed. *PLoS One*. 2013;8(10):e76890. doi:10.1371/journal.pone.0076890.
20. Singh D, Chetia H, Kabiraj D, et al. A comprehensive view of the web-resources related to sericulture. 2016;2016. https://academic.oup.com/database/article/doi/10.1093/database/baw086/2630457#87283070.
21. Wheat CW, Vogel H, Wittstock U, Braby MF, Underwood D, Mitchell-Olds T. The genetic basis of a plant-insect coevolutionary key innovation.; *PNAS.* 2007. www.pnas.org/cgi/content/full/.
22. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
23. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.
24. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina

sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170.

25. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience*. 2015;4(1):48. doi:10.1186/s13742-015-0089-y.
26. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590-6. doi:10.1093/nar/gks1219.
27. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644-652. doi:10.1038/nbt.1883.
28. Zdobnov EM, Tegenfeldt F, Kuznetsov D, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 2017;45(D1):D744-D749. doi:10.1093/nar/gkw1119.
29. Waterhouse RM, Seppey M, Simão FA, et al. BUSCO Applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 2018;35(3):543-548. doi:10.1093/molbev/msx319.
30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2.
31. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42(Database issue):D222-30. doi:10.1093/nar/gkt1223.
32. Eddy SR. Accelerated profile HMM searches. Pearson WR, ed. *PLoS*

Comput Biol. 2011;7(10):e1002195. doi:10.1371/journal.pcbi.1002195.

33. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res*. 2016;44(D1):D1087-D1093. doi:10.1093/nar/gkv1278.

34. Hammami R, Ben Hamida J, Vergoten G, Fliss I. PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res*. 2009;37(Database issue):D963-8. doi:10.1093/nar/gkn655.

35. Gasteiger E, Hoogland C, Gattiker A, et al. Protein analysis tools on the expasy server 571 571 from: the proteomics protocols handbook protein identification and analysis tools on the expasy server. http://www.expasy.org/tools/.

36. Caspi R, Altman T, Billington R, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2014;42(D1):D459-D471. doi:10.1093/nar/gkt1103.

37. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):323. doi:10.1186/1471-2105-12-323.

38. Annadurai RS, Jayakumar V, Mugasimangalam RC, et al. Next generation sequencing and de novo transcriptome analysis of *Costus pictus* D. Don, a non-model plant with potent anti-diabetic properties. *BMC Genomics*. 2012;13(1):663. doi:10.1186/1471-2164-13-663.

39. Annadurai RS, Neethiraj R, Jayakumar V, et al. De Novo transcriptome assembly (NGS) of *Curcuma longa* l. rhizome reveals novel transcripts related to anticancer and antimalarial terpenoids. *PLoS One*.

2013;8(2):e56217. doi:10.1371/journal.pone.0056217.

40. Sudheesh S, Sawbridge TI, Cogan NO, Kennedy P, Forster JW, Kaur S. De novo assembly and characterisation of the field pea transcriptome using RNA-Seq. *BMC Genomics*. 2015;16(1):611. doi:10.1186/s12864-015-1815-7.

41. Lehti-Shiu MD, Shiu S-H. Diversity, classification and function of the plant protein kinase superfamily. *Philos Trans R Soc Lond B Biol Sci*. 2012;367(1602):2619-2639. doi:10.1098/rstb.2012.0003.

42. Kobe B, Deisenhofer J. The leucine-rich repeat: a versatile binding motif. *Trends Biochem Sci*. 1994;19(10):415-421.

43. Delannoy E, Stanley WA, Bond CS, Small ID. Pentatricopeptide repeat (PPR) proteins as sequence-specificity factors in post-transcriptional processes in organelles. *Biochem Soc Trans*. 2007;35(Pt 6):1643-1647. doi:10.1042/BST0351643.

44. Lorkovim ZJ, Barta A. Genome analysis: RNA recognition motif (RRM) and K homology (KH) domain RNA-binding proteins from the flowering plant *Arabidopsis thaliana*. *Nucleic Acids Res*. 2002;30(3):623-635. doi:10.1093/nar/30.3.623.

45. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci*. 1999;24(5):181-185.

46. McKnight TD, Fitzgerald MS, Shippen DE. Plant telomeres and telomerases. A review. *Biochemistry (Mosc)*. 1997;62(11):1224-1231.

47. Finnegan DJ. Retrotransposons. *Curr Biol*. 2012;22(11):R432-R437. doi:10.1016/J.CUB.2012.04.025.

48. Chaparro C, Gayraud T, de Souza RF, et al. Terminal-repeat retrotransposons with GAG domain in plant genomes: a new testimony on the complex world of transposable elements. *Genome Biol Evol*. 2015;7(2):493-504. doi:10.1093/gbe/evv001.

49. Laten HM, Gaston GD. Plant endogenous retroviruses? a case of mysterious ORFs. Springer, Berlin, Heidelberg; 2012:89-112. doi:10.1007/978-3-642-31842-9_6.

50. Wright DA, Voytas DF. Athila4 of Arabidopsis and Calypso of soybean define a lineage of endogenous plant retroviruses. *Genome Res*. 2002;12(1):122-131. doi:10.1101/gr.196001.

51. Seo J-K, Lee MJ, Go H-J, et al. Purification and antimicrobial function of ubiquitin isolated from the gill of Pacific oyster, *Crassostrea gigas*. *Mol Immunol*. 2013;53(1-2):88-98. doi:10.1016/j.molimm.2012.07.003.

52. Seo J-K, Lee MJ, Go H-J, Park TH, Park NG. Purification and characterization of YFGAP, a GAPDH-related novel antimicrobial peptide, from the skin of yellowfin tuna, *Thunnus albacares*. *Fish Shellfish Immunol*. 2012;33(4):743-752. doi:10.1016/j.fsi.2012.06.023.

53. Saito T, Kawabata S, Shigenaga T, et al. A novel big defensin identified in horseshoe crab hemocytes: isolation, amino acid sequence, and antibacterial activity. *J Biochem*. 1995;117(5):1131-1137. doi:10.1093/oxfordjournals.jbchem.a124818.

54. Lehane MJ, Wu D, Lehane SM. Midgut-specific immune molecules are produced by the blood-sucking insect *Stomoxys calcitrans*. *Proc Natl Acad Sci*. 1997;94(21):11502-11507. doi:10.1073/pnas.94.21.11502.

55. Couto MA, Harwig SS, Cullor JS, Hughes JP, Lehrer RI. Identification of

eNAP-1, an antimicrobial peptide from equine neutrophils. *Infect Immun*. 1992;60(8):3065-3071.

56. Fujimura M, Minami Y, Watanabe K, Tadera K. Purification, Characterization, and sequencing of a novel type of antimicrobial peptides, *fa* -amp1 and *fa* -amp2, from seeds of buckwheat (*Fagopyrum esculentum* Moench.). *Biosci Biotechnol Biochem*. 2003;67(8):1636-1642. doi:10.1271/bbb.67.1636.

57. Koo JC, Lee SY, Chun HJ, et al. Two hevein homologs isolated from the seed of *Pharbitis nil* L. exhibit potent antifungal activity. *Biochim Biophys Acta - Protein Struct Mol Enzymol*. 1998;1382(1):80-90. doi:10.1016/S0167-4838(97)00148-9.

58. Destoumieux D, Munoz M, Bulet P, Bachère E. Penaeidins, a family of antimicrobial peptides from penaeid shrimp (Crustacea, Decapoda). *Cell Mol Life Sci*. 2000;57(8-9):1260-1271.

59. Yang L, Johansson J, Ridsdale R, et al. Surfactant protein b propeptide contains a saposin-like protein domain with antimicrobial activity at low pH. *J Immunol*. 2010;184(2):975-983. doi:10.4049/jimmunol.0900650.

60. Imjongjirak C, Amparyup P, Tassanakajon A, Sittipraneed S. Molecular cloning and characterization of crustin from mud crab *Scylla paramamosain*. *Mol Biol Rep*. 2009;36(5):841-850. doi:10.1007/s11033-008-9253-0.

61. Hiemstra PS, van den Barselaar MT, Roest M, Nibbering PH, van Furth R. Ubiquicidin, a novel murine microbicidal protein present in the cytosolic fraction of macrophages. *J Leukoc Biol*. 1999;66(3):423-428. doi:10.1002/jlb.66.3.423.

62. Hieshima K, Ohtani H, Shibano M, et al. CCL28 has dual roles in mucosal immunity as a chemokine with broad-spectrum antimicrobial activity. *J Immunol*. 2003;170(3):1452-1461. doi:10.4049/jimmunol.170.3.1452.

63. Lusvarghi S, Bewley C, Lusvarghi S, Bewley CA. Griffithsin: an antiviral lectin with outstanding therapeutic potential. *Viruses*. 2016;8(10):296. doi:10.3390/v8100296.

64. Alexandre KB, Moore PL, Nonyane M, et al. Mechanisms of HIV-1 subtype C resistance to GRFT, CV-N and SVN. *Virology*. 2013;446(1-2):66-76. doi:10.1016/j.virol.2013.07.019.

65. Cranfield CG, Henriques ST, Martinac B, Duckworth P, Craik DJ, Cornell B. Kalata B1 and Kalata B2 have a surfactant-like activity in phosphatidylethanolomine-containing lipid membranes. *Langmuir*. 2017;33(26):6630-6637. doi:10.1021/acs.langmuir.7b01642.

66. Marcus JP, Green JL, Goulter KC, Manners JM. A family of antimicrobial peptides is produced by processing of a 7S globulin protein in *Macadamia integrifolia* kernels. *Plant J*. 1999;19(6):699-710.

67. Westermann AJ, Vogel J. Host-Pathogen transcriptomics by dual RNA-Seq. In: *Methods in Molecular Biology (Clifton, N.J.)*. Vol 1737. ; 2018:59-75. doi:10.1007/978-1-4939-7634-8_4.

68. Park CB, Kim MS, Kim SC. A novel antimicrobial peptide from *Bufo bufo gargarizans*. *Biochem Biophys Res Commun*. 1996;218(1):408-413. doi:10.1006/bbrc.1996.0071.

69. Rolland JL, Abdelouahab M, Dupont J, Lefevre F, Bachère E, Romestand B. Stylicins, a new family of antimicrobial peptides from the Pacific blue shrimp *Litopenaeus stylirostris*. *Mol Immunol*. 2010;47(6):1269-1277.

doi:10.1016/j.molimm.2009.12.007.

70. Van den Bergh KPB, Proost P, Van Damme J, Coosemans J, Van Damme EJM, Peumans WJ. Five disulfide bridges stabilize a hevein-type antimicrobial peptide from the bark of spindle tree (*Euonymus europaeus* L.). *FEBS Lett*. 2002;530(1-3):181-185. doi:10.1016/s0014-5793(02)03474-9.

71. Cho JH, Park CB, Yoon YG, Kim SC. Lumbricin I, a novel proline-rich antimicrobial peptide from the earthworm: purification, cDNA cloning and molecular characterization. *Biochim Biophys Acta*. 1998;1408(1):67-76. doi:10.1016/s0925-4439(98)00058-1.

72. Line JE, Svetoch EA, Eruslanov B V., et al. Isolation and purification of enterocin e-760 with broad antimicrobial activity against gram-positive and gram-negative bacteria. *Antimicrob Agents Chemother*. 2008;52(3):1094-1100. doi:10.1128/AAC.01569-06.

73. Rees JA, Moniatte M, Bulet P. Novel antibacterial peptides isolated from a European bumblebee, *Bombus pascuorum* (Hymenoptera, Apoidea). *Insect Biochem Mol Biol*. 1997;27(5):413-422.

74. Guzmán-Rodríguez JJ, Ibarra-Laclette E, Herrera-Estrella L, et al. Analysis of expressed sequence tags (ESTs) from avocado seed (*Persea americana* var. drymifolia) reveals abundant expression of the gene encoding the antimicrobial peptide snakin. *Plant Physiol Biochem*. 2013;70:318-324. doi:10.1016/j.plaphy.2013.05.045.

75. Herbst R, Ott C, Jacobs T, Marti T, Marciano-Cabral F, Leippe M. Pore-forming polypeptides of the pathogenic protozoon *Naegleria fowleri*. *J Biol Chem*. 2002;277(25):22353-22360. doi:10.1074/jbc.M201475200.

76. Wu,L.-P., Wu,Z.-J., Lin,D, Fang F, Lin Q.-Y., Xie L-H. Characterization and amino acid sequence of y3, an antiviral protein from mushroom *Coprinus comatus. Chinese J Biochem Mol Biol*. 2008;24(07):597-603.

77. Liu F, Zhang X, Lu C, et al. Non-specific lipid transfer proteins in plants: presenting new advances and an integrated functional analysis. *J Exp Bot*. 2015;66(19):5663-5681. doi:10.1093/jxb/erv313.

78. Somboonwiwat K, Marcos M, Tassanakajon A, et al. Recombinant expression and anti-microbial activity of anti-lipopolysaccharide factor (ALF) from the black tiger shrimp. *Dev Comp Immunol*. 2005;29(10):841-851. doi:10.1016/j.dci.2005.02.004.

79. Yokoyama S, Kato K, Koba A, Minami Y, Watanabe K, Yagi F. Purification, characterization, and sequencing of antimicrobial peptides, Cy-AMP1, Cy-AMP2, and Cy-AMP3, from the Cycad (*Cycas revoluta*) seeds. *Peptides*. 2008;29(12):2110-2117. doi:10.1016/j.peptides.2008.08.007.

80. Fogaça AC, Almeida IC, Eberlin MN, Tanaka AS, Bulet P, Daffre S. Ixodidin, a novel antimicrobial peptide from the hemocytes of the cattle tick *Boophilus microplus* with inhibitory activity against serine proteinases. *Peptides*. 2006;27(4):667-674. doi:10.1016/j.peptides.2005.07.013

81. Cole AM, Ganz T, Liese AM, Burdick MD, Liu L, Strieter RM. Cutting Edge: IFN-Inducible ELR ⁻ CXC chemokines display defensin-like antimicrobial activity. *J Immunol*. 2001;167(2):623-627. doi:10.4049/jimmunol.167.2.623.

82. Yang D, Chen Q, Hoover DM, et al. Many chemokines including CCL20/MIP-3α display antimicrobial activity. *J Leukoc Biol*. 2003;74(3):448-455. doi:10.1189/jlb.0103024.

83. Schrader G, Apel K. Isolation and characterization of cDNAs encoding

viscotoxins of mistletoe (*Viscum album*). *Eur J Biochem*. 1991;198(3):549-553. doi:10.1111/j.1432-1033.1991.tb16049.x.

84. Kawabata S -i., Nagayama R, Hirata M, et al. Tachycitin, a small granular component in horseshoe crab hemocytes, is an antimicrobial protein with chitin-binding activity. *J Biochem*. 1996;120(6):1253-1260.
85. Arockiaraj J, Gnanam AJ, Muthukrishnan D, et al. Crustin, a WAP domain containing antimicrobial peptide from freshwater prawn *Macrobrachium rosenbergii*:immune characterization. *Fish Shellfish Immunol*. 2013;34(1):109-118. doi:10.1016/j.fsi.2012.10.009.
86. Gao N, Wadhwani P, Mühlhäuser P, et al. An antifungal protein from *Ginkgo biloba* binds actin and can trigger cell death. *Protoplasma*. 2016;253(4):1159-1174. doi:10.1007/s00709-015-0876-4.
87. Kang S-J, Kim D-H, Mishig-Ochir T, Lee B-J. Antimicrobial peptides: Their physicochemical properties and therapeutic application. *Arch Pharm Res*. 2012;35(3):409-413. doi:10.1007/s12272-012-0302-9.
88. Scocchi M, Tossi A, Gennaro R. Proline-rich antimicrobial peptides: converging to a non-lytic mechanism of action. *Cell Mol Life Sci*. 2011;68(13):2317-2330. doi:10.1007/s00018-011-0721-7.
89. Kazana E, Pope TW, Tibbles L, et al. The cabbage aphid: a walking mustard oil bomb. *Proceedings Biol Sci*. 2007;274(1623):2271-2277.
90. Klein M, Reichelt M, Gershenzon J, Papenbrock J. The three desulfoglucosinolate sulfotransferase proteins in Arabidopsis have different substrate specificities and are differentially expressed. *FEBS J*. 2006;273(1):122-136. doi:10.1111/j.1742-4658.2005.05048.x

CHAPTER 4

Transcriptome profile of *Antheraea assamensis* with respect to host plant and development

CHAPTER 4

Transcriptome profile of *Antheraea assamensis* with respect to host plant and development

---

ABSTRACT

A. assamensis is an economically significant silkworm due to its ability to produce golden silk and its endemic nature. Dearth of sequence information on this species has hindered the scientists and indigenous seri-rearer communities of India for long. In this study, we sequenced the *de novo* transcriptomes of 5th instar larvae of *A. assamensis* reared on two of its host plants, *Litsea citrata* and *Machilus bombycina*. We also sequenced the *de novo* transcriptomes of 4th instar larvae of *A. assamensis* reared on *M. bombycina.* Using the data generated in this study, we reconstructed the transcriptome for *A. assamensis,* identified the top most expressed transcripts in each tissue and observed how biological processes associated with each tissue varies with respect to host plant and larval development (4th instar and 5th instar). We found that translation was the most unanimous process expressed in each tissue of silk gland of *A. assamensis* regardless of the developmental stage or host plant. Other than this process, other processes like oxidative stress management, redox homeostasis, transcriptional regulation, etc. had variable representation across different stages. Analysis of these patterns showed how the transcriptional profile of *A. assamensis* can vary in different anatomical sections and variation in host plants.

4.1 INTRODUCTION

The adaptability of any organism to any change is usually termed as plasticity. The same terminology is used in biology, to describe the ability of expressing variable genotype (genotypic plasticity) and phenotype (phenotypic plasticity). An organisms' gene expression profile also shows variability as (i) the collective transcriptome where the pattern of functional categories of genes change or (ii) in specific pathways or genes where one isoform is preferred over another. This phenomenon is generally described as transcriptomic or gene expression plasticity [1,2]. This phenomenon is applicable for the complete repertoire of functional genomic content ranging from protein-coding genes to non-coding RNAs.

Gene expression ideally links genotype to phenotype and plays a central role in determining a cell's adaptation to the changing environment. It is common for regulatory elements of genes to be highly conserved, however, the patterns of gene expression changes with response to biotic and abiotic factors around it. This is especially true in the case of insects. They strongly respond to changes in diet or environmental cues like temperature rise or fall and manifests discernible changes in their physical characteristics [3]. Long-term interactions of such manner lead to adaptive evolution and heritable changes in genotype as well as phenotype. A remarkable example of such a relationship is that between the yucca moth and its host, the yucca plant which are obligately co-dependent for host plant pollination and moth diet respectively [4]. Below, we discuss two scenarios where gene expression plasticity is usually manifested in herbivorous insects with an emphasis on Lepidopterans.

4.1.1 Gene expression variations in response to host plant:

Changes in host plant can impact multiple aspects of a herbivorous insect's life starting from its reproductive strategy to higher trophic level interactions like predators, pests etc [5]. The herbivores feed on a wide variety of plants and are exposed to phytochemical diversities in terms of defense compounds. The general notion is that polyphagous insects are capable of displaying variations in their gene expression profiles, especially in their alimentary canal or gut, as it is the primary site for digestion, nutritional absorption and detoxification, followed by the feeding apparatus [6,7]. Their interactions begin when the salivary secretory components of the invading insect comes in contact with the plant leading to a cascade of defensive compound biosynthesis like terpenoids, cyanogenic glycosides and so forth [8]. This encounter triggers the anti-phytochemical defense mechanism in the herbivore's gut to avoid trauma. For example, *Helicoverpa zea* larvae suppresses host plant defense in tobacco owing to the presence of glucose oxidase in its saliva and its expression varies with respect to host plant [9]. Similarly, *Pieris brassicae* and *Spodoptera littoris* suppresses defenses in a non-host plant, *Arabidopsis thaliana*, leading to increased larval weights [10]. Now, these reciprocal exchanges over time leads to evolution of novel mechanisms for dismantling plant defenses, staring from avoidance of strongly resistant plants to suppression.

Changes in host plant may also lead to changes in the changes in nutrition uptake and digestion patterns in any herbivorous insect. The components that represent any host plants value such as carbon or nitrogen content, can clearly affect herbivore's fecundity [5]. Changes in the phytochemical content can have impact on the expression of regular biosynthetic processes leading to apparent

changes in their secretory products, such as silk. An example for this type of change are the reports on colour variations in eri silk cocoons when reared on different host plants [11]. Another example can be the increase in larval weight of *P. brassicae* and *S. littoris* described above. These types of phenotypic changes are accompanied or preceded by changes in the usual expression profile of selected genes. Based on the existing studies of these nature, we can anticipate the significance of profiling the gene expression variations to gain deeper understanding of insect-host plant relationship.

4.1.2 Gene expression variation in response to development

The typical life cycle of Lepidopterans can be divided into egg, caterpillar (larvae), pupa and adult. The larval stage of this holometabolan sequence is usually the most actively feeding stage. Larval growth rate is strongly affected by the environment and host plant allelochemicals. The larval stages are divided into different instars based on the number of molting (shedding of skin) in the species. For eg. silkworms generally molt five times and therefore, have five instars. There is a drastic difference in the gene expression profile of a fifth instar larvae than any other instar and this phenomenon can be best explained by using silkworm, say *Bombyx mori,* as an example. Transition of silkworms into the fifth instar stage is accompanied by a more voracious feeding behavior which ceases right before it starts spinning a cocoon for its impending metamorphosis to pupa. This transition is accompanied by various changes in physiology of the silk gland. The silk gland is a typical exocrine gland specialized in production of large amounts of silk proteins- fibroin and sericin. By the end of 5th instar, these glands reach the size of ~25 cm and constitute ~40% of the body weight [12]. This increase is made possible by genome amplification, cell enlargement and co-

ordinated changes in gene activity, showing marked increase in DNA, RNA as well as proteosynthesis of proteins like fibroin, sericin, etc [13]. Similar changes are also observed during instar-level developmental transitions.

Insects have a complex endocrine system coordinated by sesquiterpenoids like methyl farnesoate (MF) and juvenile hormone (JH) and ecdysteroid (like 20-hydroxyecdysone or 20E) which regulate their development. A high JH titer is known to arrest molting or ecdysis process (one larval stage to another) while a high titer of 20E triggers molting or ecdysis [14,15]. Understanding how developmental changes are accompanied by changes in gene expression profiles will be an interesting way to learn more about the organismal biology.

A. assamensis is an economically significant silkworm and the dearth of basic biological knowledge on it has long hindered the scientists and indigenous seri-rearer communities of India. In Chapters-2 and 3, we had assembled and annotated the *de novo* transcriptomes of *A. assamensis* and its two host plants, *Litsea citrata* and *Machilus bombycina.* In this study, we studied changes in transcriptomic profiles that accompany the muga silkworm (silk gland) when fed upon both of these host plants. Similarly, we studied how these patterns are affected by development (4th instar and 5th instar).

4.2 MATERIALS AND METHOD

4.2.1 Sample collection, RNA isolation, cDNA library preparation and sequencing:

Muga silkworms were reared from egg to 5th instar larval stage at Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh (26.7844° N, 94.3443°

E). The rearing was carried during winter season (January-February) on two separate host plants, namely, *M. bombycina* and *L. citrata,* in outdoor conditions. 10 no.s of 4th and 5th instar larvae of *A. assamensis* were collected from Som plant. Similarly, 10 no.s of 5th instar larvae were collected from Mejankari plant. These larvae were dissected under sterile conditions to obtain the following sections- silk gland, alimentary canal or gut and residual body. The silk glands were further dissected into anterior, middle and posterior silk glands. These tissues were further dissected into very thin sections and stored in RNAlater stabilization solution (Ambion^{TM}) at -80˚C. These samples were further processed at Genotypic Technology, Bangalore.

Total RNA was isolated from these harvested tissues using RNAEasy mini kit (Qiagen). The concentration, purity and integrity of the isolated RNA (A260 /A280 ratio ≥1.8 and RIN number ≥8) were verified using Nanodrop spectrophotometer and High Sensitivity Bioanalyzer Chip (Agilent Technologies, CA, U.S.A.). Preparation of each tissue library was performed using Illumina-compatible SureSelect Strand-Specific RNA Library (Part # G9691-90010) except for alimentary canal and residual body of, See Chapter 2) (Agilent Technologies, Santa Clara, CA, U.S.A.). The resulting cDNA libraries were sequenced on Illumina HiSeq^{TM} 4000 sequencer platforms using the paired-end sequencing protocol at Genotypic Technology Pvt. Ltd., Bangalore, India. Library preparation and sequencing protocol for alimentary canal and residual body of 5th instar muga larvae on *M. bombycina* (TruSeq RNA Library, Part #15008136; Illumina HiSeq^{TM} 2000) have been discussed in Chapter 2. A table with the sequenced samples with their acronyms and sequencing statistics have been provided below.

Table 4.1 Information on *Antheraea assamensis* samples of this study

HOST PLANT	INSTAR	TISSUE/ GLAND	LIBRARY METHOD	SEQUENCING PLATFORM	ACRONYM
Machilus bombycina	4th	Anterior silk gland	SureSelect Strand-Specific RNA library kit	Illumina HiSeq™ 4000	Som 4ASG
-do-	4th	Middle silk gland	-do-	-do-	Som 4MSG
-do-	4th	Posterior silk gland	-do-	-do-	Som 4PSG
-do-	5th	Anterior silk gland	SureSelect Strand-Specific RNA library kit	-do-	Som 5ASG
-do-	5th	Middle silk gland	-do-	-do-	Som 5MSG
-do-	5th	Posterior silk gland	-do-	-do-	Som 5PSG
-do-	5th	Alimentary Canal	TruSeq RNA Library	Illumina HiSeq™ 2000	Som 5AC
-do-	5th	Residual body	TruSeq RNA Library	Illumina HiSeq™ 2000	Som 5RB

Litsea citrata	5th	Anterior silk gland	SureSelect Strand-Specific RNA library kit	Illumina HiSeq™ 4000	Mej5 ASG
-do-	5th	Middle silk gland	-do-	-do-	Mej5 MSG
-do-	5th	Posterior silk gland	-do-	-do-	Mej5 PSG
-do-	5th	Alimentary Canal	-do-	-do-	Mej5 AC
-do-	5th	Residual Body	-do-	-do-	Mej5 AC

Table 4.2 Experimental matrix for comparison of [A] host-plant induced and [B] development-induced biological processes in *A. assamensis*

A. *Set X vs Set Y*		
Host plant →	**Set X** ***Machilus bombycina* (Som)**	**Set Y** ***Litsea citrata* (Mejankari)**
Developmental stage → Tissue↓	**5th instar**	**5th instar**
ASG	Som5ASG	Mej5ASG
MSG	Som5MSG	Mej5MSG
PSG	Som5PSG	Mej5PSG
AC	Som5AC	Mej5AC

B. Set A vs Set B		
Host plant- *Machilus bombycina* (Som)		
Developmental Stage→ / Tissue↓	Set A 4th instar	Set B 5th instar
ASG	Som4ASG	Som5ASG
MSG	Som4MSG	Som5MSG
PSG	Som4PSG	Som5PSG

4.2.2 Quality control of raw data and *de novo* assembly of a consolidated transcriptome for *A. assamensis*

The quality metrices of the resulting raw reads per data set were examined using FastQC v0.11.5 [16]. Based on the report, both paired-end datasets were corrected for adapter contamination, over-represented sequences, erroneous k-mers, low quality reads as well as tiles (Phred quality score cutoff was ≥30) using a combination of Trimmomatic 0.35, Trim Galore, rCorrector tools as well as in-house shell scripts [17–19]. The resulting raw reads were mapped to ribosomal rRNA reads (entitled SSUParc and LSUParc files) from the SILVA rRNA database project using bowtie 2 and unmapped read pairs were retained [20]. The resultant reads were re-examined using FastQC to check for attainment of desirable quality features. Following this, the final thirteen sets of paired-end reads were utilized as inputs for de novo assembly of a consolidated transcriptome for *A. assamensis* via Trinity tool (v2.6.5) utilizing the in-built parameters for normalization by read set and k-mer size of 32 [21]. The quality of

transcriptome assemblies was assessed on the basis of parameters like N50 value, mean transcript length, percentage of reads mapped to the transcriptome, etc. using the scripts provided under Trinity package. Transcriptome completeness was quantitatively examined via BUSCO (Benchmarking Universal Single-Copy Orthologs) tool v3 using the dataset "Insecta odb10" curated by OrthoDB in transcriptome mode [22,23].

4.2.3 Annotation, identification of candidate genes and comparative gene expression

The consolidated transcriptome was annotated using a combination of protein sequence databases specific for insects (UniProt). The respective gene ontology annotations were also retrieved from the same database. A bowtie-indexed reference map was created using the newly assembled transcriptome of *A. assamensis* followed by estimation of abundance by a tool, RSEM, which aligns the quality control processed paired-end raw reads per sample to the bowtie-indexed reference of a transcriptome [21,24]. RSEM generated a normalized count-based expression values for each transcript (termed 'transcripts per million' or TPM). The transcripts were sorted based on TPM values into the top 1000 most-expressed genes followed by identification of the top fifty GO Biological (GO_BP) among those transcripts.

4.3 RESULTS AND DISCUSSION

4.3.1 Assembly and annotation of the transcriptome

The transcriptome of *A. assamensis* was e-assembled using the newly sequenced tissues which consisted of 2,58,367 transcripts. The overall

completeness estimated using BUSCO showed that 98.4% of single copy orthologues from Insecta were present in the transcriptome; this indicated the robust nature of our assembly. We further annotated these transcripts using the UniProt database and were able to assign gene names to 1,03,914 transcripts (~40% of the transcripts). The list of SRA IDs for the samples sequenced for this study are listed in Table 4.3.

Table 4.3 Identifiers of the samples sequenced for this study and submitted to NCBI Short read archive (SRA) database are shown below-

Transcriptome Sample Name	**SRA ID**	**Transcriptome Sample Name**	**SRA ID**
Muga 4th instar Anterior Silk Gland on Som	SRR8208773	Muga 5th instar Anterior Silk Gland on Mejankari	SRR8208764
Muga 4th instar Middle Silk Gland on Som	SRR8208772	Muga 5th instar Middle Silk Gland on Mejankari	SRR8208777
Muga 4th instar Posterior Silk Gland on Som	SRR8208769	Muga 5th instar Posterior Silk Gland on Mejankari	SRR8208776
Muga 5th instar Anterior Silk Gland on Som	SRR8208768	Muga 5th instar Alimentary Canal on Mejankari	SRR8208775
Muga 5th instar Middle Silk Gland on Som	SRR8208771	Muga 5th instar Residual Body on Mejankari	SRR8208774
Muga 5th instar	SRR8208770		

Posterior Silk Gland on Som		
Muga 5th instar Alimentary Canal on Som	SRR2532163	
Muga 5th instar Residual Body on Som	SRR2532165	

4.3.2 Comparison of fourth and fifth instar silk gland for development-induced changes

For observing how development changes accompany changes in the transcriptome profiles, we compared the 4th instar silk gland of muga silkworm reared on Som plant with that of 5th instar silk gland. As mentioned in the materials and methods, the transcripts per sample were first ranked in descending order of their TPM values. The top 1000 genes with the highest TPM values were extracted and their gene ontological information was retrieved. The ontology information for biological processes (GO_BP) was further analyzed to identify the top 50 GO classes that were over-represented among those highly expressed transcripts. We compared these top 50 GO_BP classes across anterior, middle and posterior silk gland of 4th and 5th instar larvae (Fig. 4.1-3).

Comparison of the top biological processes across these tissues showed that translation was one of the most active processes for the silk gland as it was

unanimously the most expressed biological process across all the tissues sequenced (Fig. 4.1-3). A close observation of the top processes showed that after translation, response to oxidative stress, ATP synthesis and hydrolysis coupled proton transport, redox homeostasis, aromatic compound catabolism, carbohydrate metabolism, translation initiation complex, protein folding and ribosome biogenesis were among the most expressed processes in the anterior, middle and posterior tissues of 4th instar silk gland. While all these processes were also present in the 5th instar silk gland, a few other processes like chitin metabolic process, transcriptional regulation and intracellular protein transport were more prominent among the most over-represented GO processes. These processes indicate that the biological needs for the 4th and the 5th instar silk gland varies. For e.g. oxidative stress management is very important for the silk gland as it is a specialized gland for silk synthesis and oxidative stress heightens during such translationally and energetically stressful biological processes like nearly constitutive translation of silk proteins. Cell division and cell death are usually more heightened during the processes of silk gland development, thus generating more free radicals which is counterbalanced by the redox homeostasis processes. Feeding of antioxidants has been shown to increase the rate of silk synthesis by 31% [25]. This is an indication that redox homeostasis is a probable energetic challenge for the silk gland and alleviating this process can be useful in increasing silk output. Chitin metabolic process was distinctly more prevalent in the fifth instar ASG than the fourth instar ASG. Chitin is a part of insect cuticle and peritrophic matrix and presence of chitinous matrices provides properties like elasticity [26]. Anterior silk gland consists of a thick duct wall of chitin that narrows the lumen and chitin also forms an extracellular matrix to protect the silk gland

cell from high mechanical shear of spinning [27]. This biological process is a clear indication of how the silk gland's transcriptome profile adapts to the biological needs of the organism.

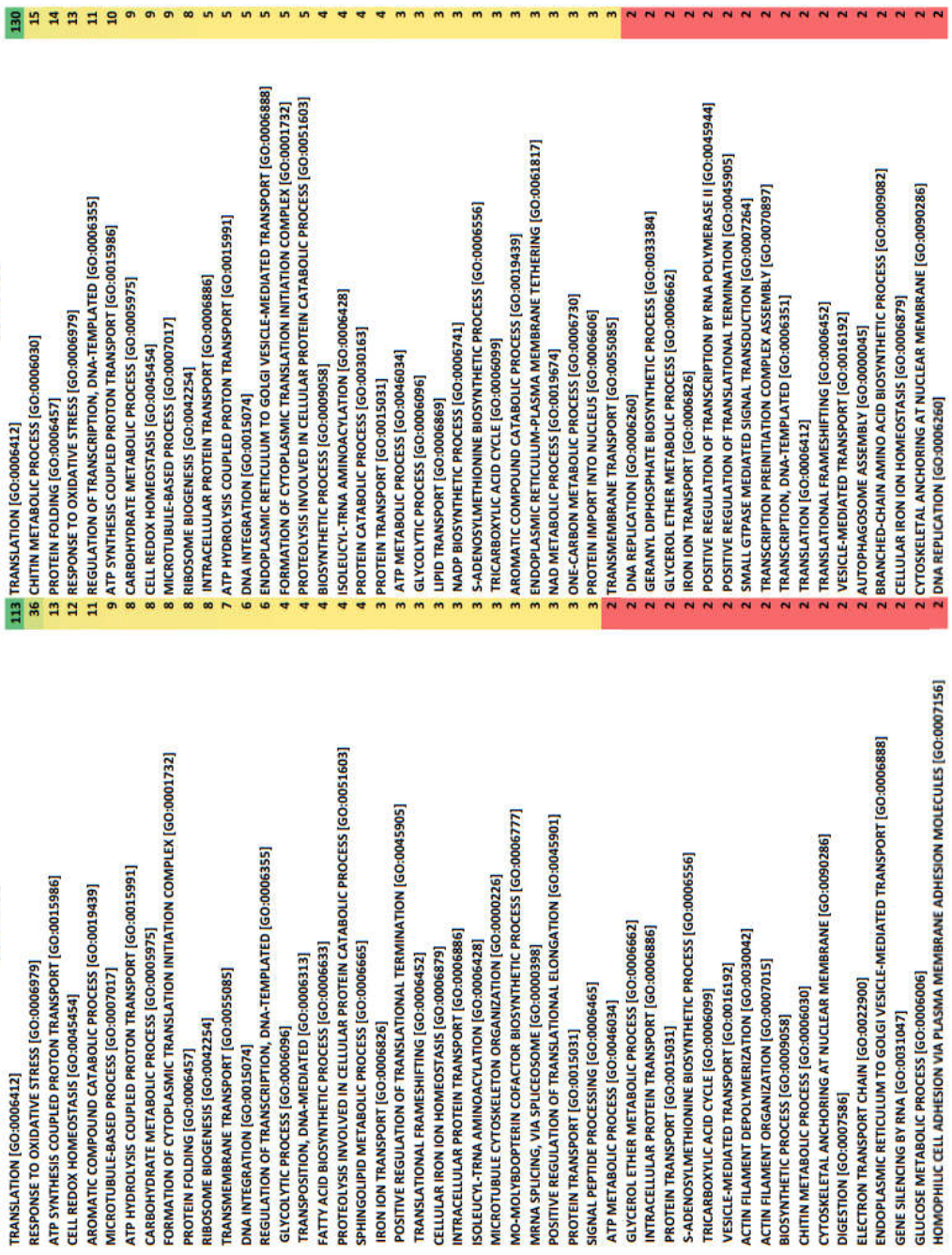


Fig. 4.1 A comparative heat map of the top fifty over-represented biological processes in anterior silk gland of 4th and 5th instar larvae of *A. assamensis*

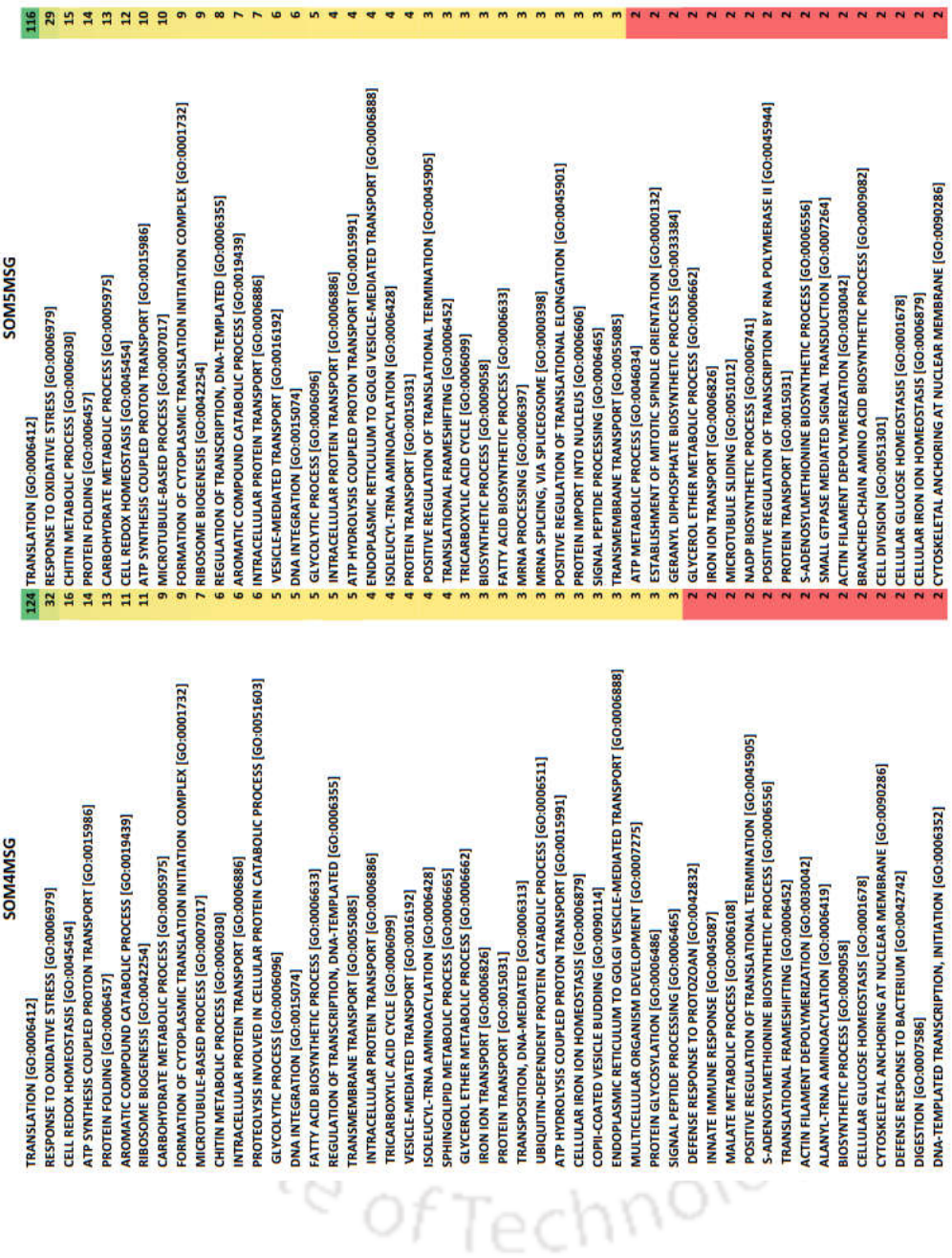


Fig. 4.2 A comparative heat map of the top fifty over-represented biological processes in middle silk gland of 4th and 5th instar larvae of *A. assamensis*

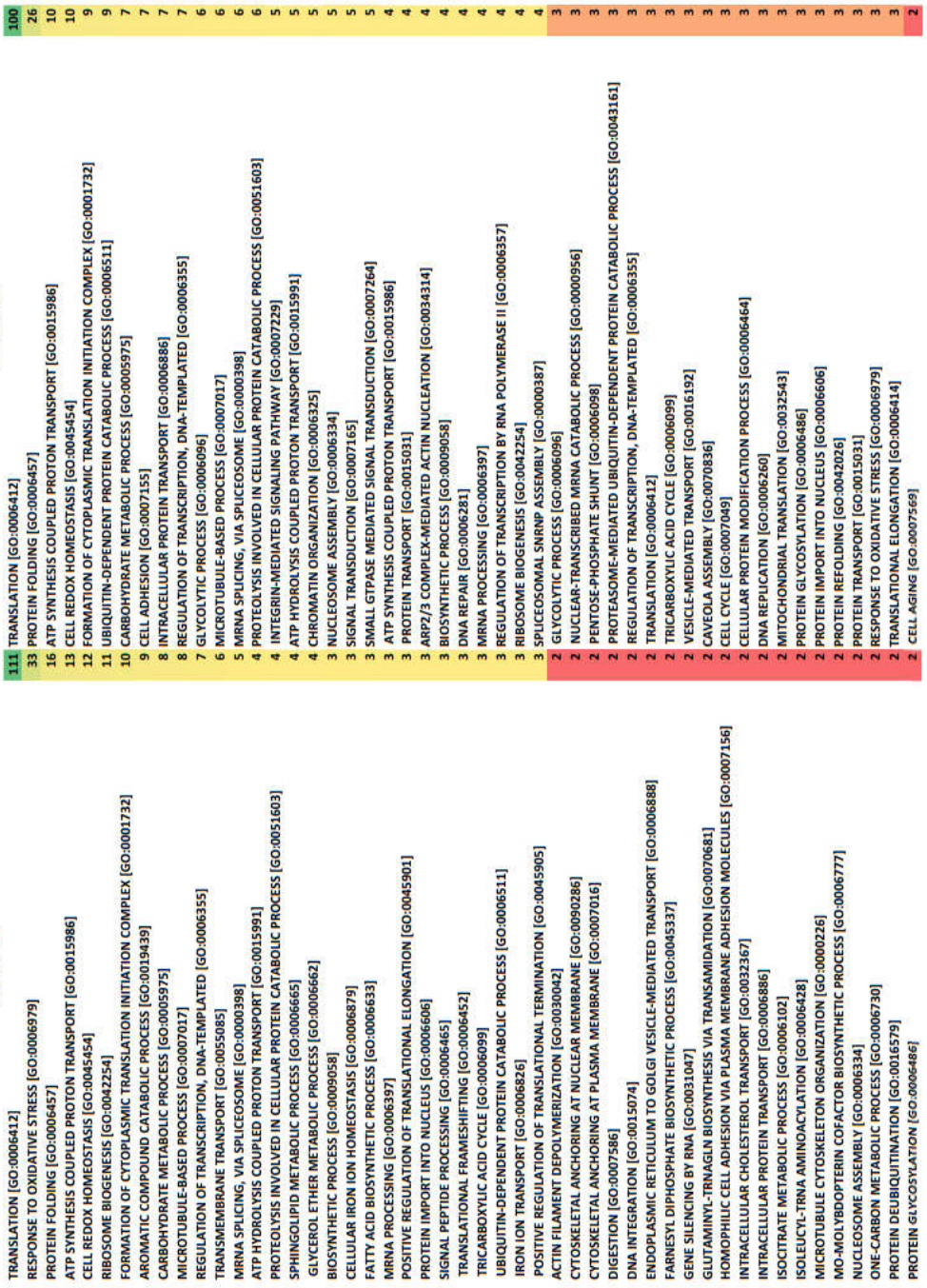


Fig. 4.3 A comparative heat map of the top fifty over-represented biological processes in posterior silk gland of 4th and 5th instar larvae of *A. assamensis*

4.3.3 Comparison of fifth instar *A. assamensis* silk gland for host plant-

For observing how variation in host plant accompany transcriptome dynamics of the silk gland, we compared the 5th instar silk gland of muga silkworm reared on *M. bombycina* (som) to that reared on *L. citrata* (Mejankari). Som plant with that reared on of 5th instar silk gland. TPM for the transcripts were calculated using RSEm and was followed by sorting them in descending order of these values. The top 1000 genes with the highest TPM values were extracted and their gene ontological information was retrieved. The ontology information for biological processes (GO_BP) was further analyzed to identify the top 50 GO classes that were over-represented among those highly expressed transcripts.

Again, translation was the most active process in all the transcriptomes for both host plants (Fig. 4.4-6). Silk glands reared on both host plants shared some common biological processes like carbohydrate metabolism, ATP hydrolysis coupled proton transport, cell redox homeostasis etc. For the Som-reared muga silkworms, one of the prominent variable process was aromatic compound catabolic process. For Mejankari-reared muga silkworm, it was transmembrane transport. Another process related to chitin metabolism was dominant in the middle silk gland of som-reared muga silkworm and posterior silk gland of mejankari-reared muga silkworm. From the ontologies of these biological processes, one can assume that transcriptome scenario of muga silkworm is very responsive towards the host plant. These variations usually rise due to the nutritional and allelochemical content of the respective host plant and the ability of the dependent organism to metabolize or detoxify such a content.

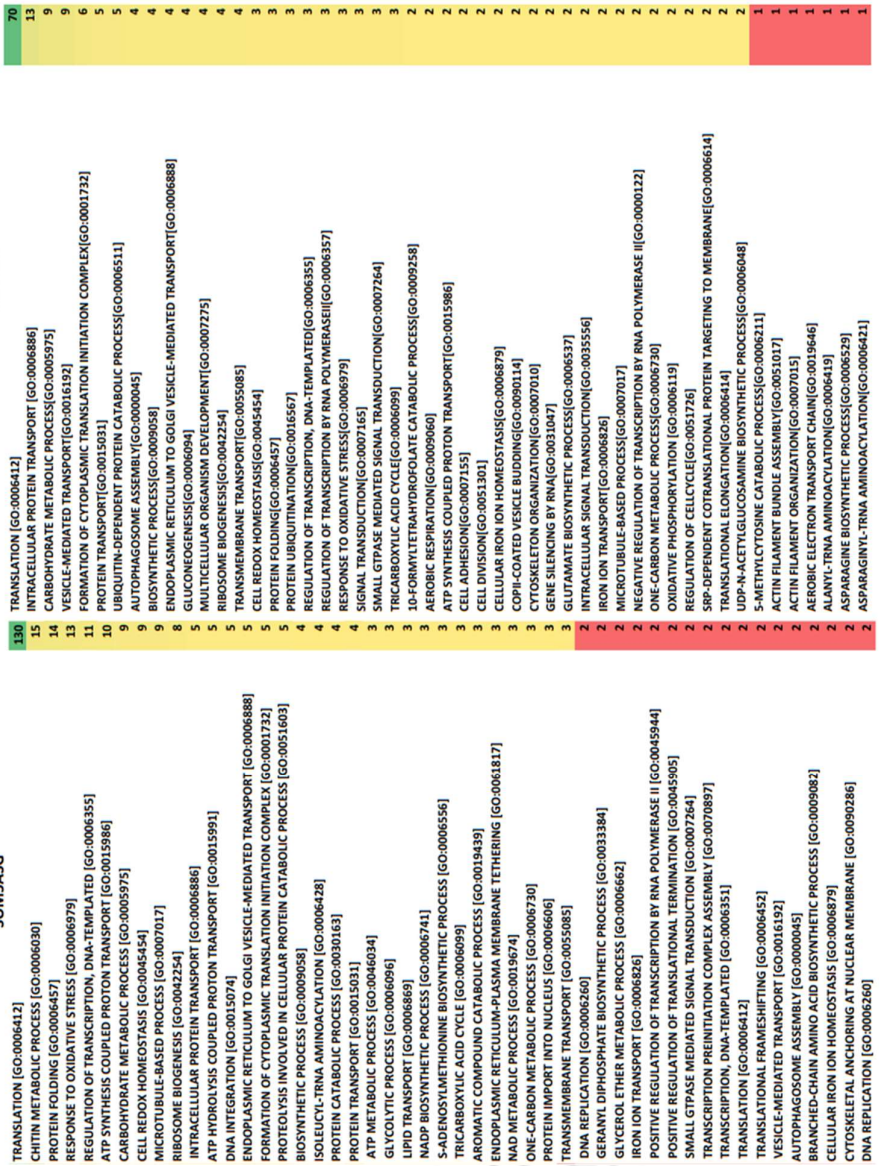


Fig. 4.4 A comparative heat map of the top fifty over-represented biological processes in anterior silk gland of 5th instar larvae of *A. assamensis* reared on *M. bombycina* (Som) and *L. citrata* (Mejankari)

Fig. 4.5 A comparative heat map of the top fifty over-represented biological processes in middle silk gland of 5th instar larvae of *A. assamensis* reared on *M. bombycina* (Som) and *L. citrata* (Mejankari)

Fig. 4.6 A comparative heat map of the top fifty over-represented biological processes in posterior silk gland of 5th instar larvae of *A. assamensis* reared on *M. bombycina* (Som) and *L. citrata* (Mejankari)

4.4 CONCLUSION

In the current study, we demonstrated how the silk gland's transcriptome profile varies with respect to developmental stage such as 4th and 5th instar as well as host plant such as *M. bombycina* and *L. citrata* using gene ontological information of the annotated transcripts. A few processes like chitin metabolism and transmembrane transport were expressed differentially with respect to host plant changes. Chitin metabolism was also expressed differentially with respect to the developmental stage or instar. Studies of such nature will be helpful in exploring host-plant and insect dynamics in future.

REFERENCE

1. Lindberg J, Lundeberg J. The plasticity of the mammalian transcriptome. *Genomics*. 2009;95:1-6. doi:10.1016/j.ygeno.2009.08.010

2. Kenkel CD, Matz M V. Gene expression plasticity as a mechanism of coral adaptation to a variable environment. *Nat Ecol Evol*. 2016;1(1):0014. doi:10.1038/s41559-016-0014

3. Christodoulides N, Van Dam AR, Peterson DA, et al. Gene expression plasticity across hosts of an invasive scale insect species. Doucet D, ed. *PLoS One*. 2017;12(5):e0176956. doi:10.1371/journal.pone.0176956

4. Yoder JB, Smith CI, Pellmyr O. How to become a yucca moth: Minimal trait evolution needed to establish the obligate pollination mutualism. *Biol J Linn Soc Lond*. 2010;100(4):847-855. doi:10.1111/j.1095-8312.2010.01478.x

5. Awmack CS, Leather SR. Host Plant Quality and Fecundity in Herbivorous Insects. *Annu Rev Entomol*. 2002;47(1):817-844. doi:10.1146/annurev.ento.47.091201.145300

6. Simon J-C, Alençon ', Guy E, et al. Genomics of adaptation to host-plants in herbivorous insects. doi:10.1093/bfgp/elv015

7. Nallu S, Hill JA, Don K, et al. The molecular genetic basis of herbivory between butterflies and their host plants. *Nat Ecol Evol*. 2018;2(9):1418-1427. doi:10.1038/s41559-018-0629-9

8. Ali JG, Agrawal AA. Specialist versus generalist insect herbivores and plant defense. *Trends Plant Sci*. 2012;17(5):293-302. doi:10.1016/j.tplants.2012.02.006

9. War AR, Paulraj MG, Ahmad T, et al. Mechanisms of plant defense against insect herbivores. *Plant Signal Behav*. 2012;7(10):1306-1320. doi:10.4161/psb.21663

10. Rashid War A, Kumar Taggar G, Hussain B, Sachdeva Taggar M, Nair RM, Sharma HC. Plant defense against herbivory and insect adaptations. *AoB Plants*. 2018;10(4). doi:10.1093/aobpla/ply037

11. Chutia P, Kumar R, Khanikar DP. Host plants relationship in terms of cocoon colour and compactness of eri silkworm *(Samia ricini*). https://www.researchtrend.net/bfij/bf12/55

12. Akai H. The ultrastructure and functions of the silk gland cells of *Bombyx mori*. 1984:323-364. doi:10.1007/978-1-4613-2715-8_9

13. Sehnal F, Sutherland T. Silks produced by insect labial glands. *Prion*. 2008;2(4):145-153.

14. Sehnal F, Akai H. Insect silk glands: their types, development and function, and effects of environmental factors and morphogenetic hormones on them. *Int J Insect Morphol Embryol*. 1990;19(2):79-132. doi:10.1016/0020-7322(90)90022-H

15. Jindra M, Palli SR, Riddiford LM. The Juvenile hormone signaling pathway in insect development. *Annu Rev Entomol*. 2013;58(1):181-204. doi:10.1146/annurev-ento-120811-153700

16. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

17. Krueger F. Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. 2015. https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/.

18. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

19. Song L, Florea L. Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *Gigascience*. 2015;4(1):48. doi:10.1186/s13742-015-0089-y

20. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41(Database issue):D590-6. doi:10.1093/nar/gks1219

21. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 2011;29(7):644-652. doi:10.1038/nbt.1883

22. Zdobnov EM, Tegenfeldt F, Kuznetsov D, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 2017;45(D1):D744-D749. doi:10.1093/nar/gkw1119

23. Waterhouse RM, Seppey M, Simão FA, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*. 2018;35(3):543-548. doi:10.1093/molbev/msx319

24. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12(1):323. doi:10.1186/1471-2105-12-323

25. Lattala GM, Kandukuru K, Gangupantula S, Mamillapalli A. Spermidine enhances the silk production by mulberry silkworm. *J Insect Sci*. 2014;14(1). doi:10.1093/jisesa/ieu069

26. Muthukrishnan S, Merzendorfer H, Arakane Y, Yang Q. Chitin metabolic pathways in insects and their regulation. In: *Extracellular Composite Matrices in Arthropods*. Cham: Springer International Publishing; 2016:31-65. doi:10.1007/978-3-319-40740-1_2

27. Dong Z, Zhao P, Zhang Y, et al. Analysis of proteome dynamics inside the silk gland lumen of *Bombyx mori. Sci Rep*. 2016;6(1):21158. doi:10.1038/srep21158

CHAPTER 5

MugaSeqDB, a database on *Antheraea assamensis* and its host plants

CHAPTER 5

MugaSeqDB, a database on *Antheraea assamensis* and its host plants

ABSTRACT

Here, we developed MugaSeqDB, a comprehensive, freely accessible database hosting the transcriptome data of muga silkworm and its two host plants, Som and Mejankari. It contains transcripts, predicted proteins, their respective functional and ontological annotations. It also provides secondary information on the three primary species like pest, pathogen and patents. Back-end of the database was developed using MySQL and phpMyAdmin while the front end was created using a combination of HTML, PHP and java scripts. The complete architecture was hosted at a Linux-based commercial server. Overall, MugaSeqDB embodies a platform-independent, user-friendly research tool that opens up a plethora of sequence data for exploration and constitutes an essential genomic resource to facilitate future studies on this endemic species.

5.1 INTRODUCTION

Muga silk is an important cultural and commercial commodity produced by the endemic lepidopteran “muga” silkworm, *Antheraea assamensis.* This unique lustrous golden yellow fabric contributes hugely towards the Indian sericulture industry. Its proteins, fibroin and sericin, have already been established as

successful biomaterials for biomedical applications [1,2]. Despite this extent of significance, genomic information on this important species was scarce until now, thus, acting as an impediment on possible research avenues like boosting silk yield or immunity by genetic improvement. In this study, we sequenced the transcriptome of muga silkworm (silk gland, alimentary canal and residual body for 4th as well as 5th instar larvae) and annotated transcripts using a combination UniProt, NCBI Protein, GO and Pfam databases (Chapter 2, 4 and 5). A similar strategy was also applied to the transcriptomes of the muga silkworm host plants, *M. bombycina* and *L. citrata* (Chapter 3). The raw sequencing reads were submitted to NCBI SRA database (see Chapter-2 and 3 for SRA IDs). However, it is only through sharing the transcripts and predicted proteins of these three species that we can contribute more meaningfully towards the serigenomic research scenario.

Next-generation sequencing technologies were introduced into scientific world almost a decade ago [3]. Given the tremendous amount of data generated in a single sequencing run, it is necessary to organize these data in a database and make it accessible to the public for facilitating further research. Some examples of species-specific databases include SilkTransDB (*Bombyx mori*), FlyBase (*Drosophila melanogaster*), etc [4–6].

Genome and transcriptome sequencing efforts have always contributed towards acceleration of scientific research. For e.g. the complete high-coverage draft genome of the domesticated mulberry silkworm, *B. mori,* was published in 2008. Since then, a myriad of research on the species has ensued ranging from transgenesis to genome editing, thus, giving it the reputation of being the

Lepidopteran model organism [7–9,10]. Similar or even more novelties can be expected from the lesser-studied semi- or undomesticated silkworms if the data from genomic studies become accessible to researchers.

Here, we developed MugaSeqDB, a comprehensive, freely accessible database hosting the transcriptome data of muga silkworm (*Antheraea assamensis)* and its two host plants, Som (*Machilus bombycina)* and Mejankari (*Litsea citrata)*. Its data structure, construction strategy and salient features are discussed below. This database comprises of essential genomic resources to facilitate future studies on this endemic species and its host plants. It is a platform-independent, user-friendly research tool that opens up a plethora of sequence data for future exploration.

5.2 MUGASEQDB CONSTRUCTION:

5.2.A Data type:

MugaSeqDB has two primary data sources- A. *De novo* transcriptome of *A. assamensis* and B. *De novo* transcriptome of host plants of *A. assamensis.* We derived the following information from each source-

A. Transcript sequences
B. Predicted proteins and antimicrobial peptides[#]
C. Functional annotations of transcripts
D. Gene Ontology
E. Functional Domains (Pfam)
F. Predicted Antimicrobial peptides

[#]To be made available in future

The database also had secondary data on the following -

A. Pest and pathogens of-
   a. Muga silkworm
   b. Host plants of muga silkworm

B. Patents applied or published
   a. Muga Silkworm
   b. Host plants of muga silkworm

5.2.B Construction of the database-

Web-based database application mostly consists of two components, namely, a database server and a web-based graphical user interface.

5.2.B.1 Database server:

The data described in section 5.2.A were initially prepared as csv (comma separated value) worksheets. Each entry was provided a unique identifier (MDB_AA#_001, MDB_MB#_001 or MDB_LC#_001). SW represented entries for muga silkworm while HP represented host plant entries. # was replaced with TR (transcript), PR (protein), PP (pest) and PT (patent) wherever applicable. For e.g. MDB_SWTR_001 denoted the Transcript 001 from muga silkworm while MDB_SWPR_001 denoted the predicted protein 001 from the same organism. MB and LC stood for *M. bombycina* and *L. citrata*.

The services of a commercial online hosting server, Arvixe, was used to host MugaSeqDB. We used the content administration tool, PhpMyAdmin and its in-built MySQL, an open-source relational database management system (RDBMS), for database construction. We first uploaded the .csv worksheets into MySQL which automatically creates a hierarchical database schematic for the input data.

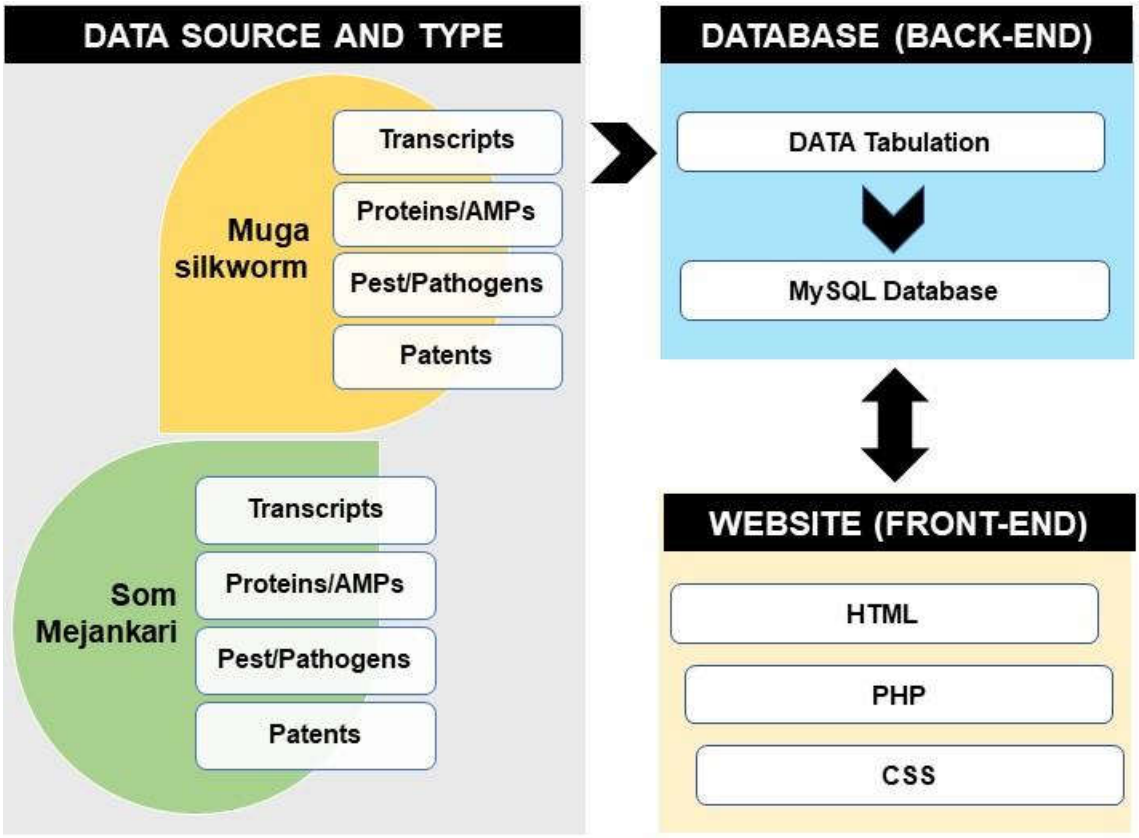


Fig. 5.1 The workflow and components involved in MugaSeqDB, namely, data source and type, back-end and front-end components

5.2.B.2 Web-based graphical user interface (Website)-

The website of MugaSeqDB was created using HTML, PHP and java scripts on the front-end with the MySQL database and Apache web server at the back end. The complete architecture resides upon Linux-based web host, Arvixe. HTML based web pages were designed using HTML text processors in cPanel. CSS scripts were edited as per the requirement for effective viewing and ease of access of the prospective users. The search engine and data browser for MugaSeqDB was developed using in-house php scripts. The website design was kept responsive and dynamic, thus aiding in hassle-free visualization of the site on different web browsers like Mozilla Firefox, Google Chrome, Internet Explorer

at different screen resolutions (e.g. in desktops, tablets or smartphones). Finally, a web-domain name was registered for the website (http://mugaseqdb.in) and connected to the website. The workflow for database construction and design has been diagrammatically represented in Fig. 5.1.

5.3 SALIENT FEATURES OF MUGASEQDB

MugaSeqDB is the first-ever database providing primary sequence information on the endemic silkworm, *A. assamensis* and its two host plants, *L. citrata* and *M. bombycina.* Hence, the most important feature of MugaSeqDB is the uniqueness of its data, i.e. transcripts and proteins, available in it. While raw sequencing reads of these three species (generated by this thesis) are only available in NCBI-SRA database (See Ch-2), the full-length assembled transcripts are available on MugaSeqDB. These transcripts have been exhaustively annotated using multiple databases like UniProt, NCBI Protein, GO, Pfam etc. All this information related to the transcripts are available in this database for further research. MugaSeqDB data also acts as a source of pest-pathogen as well as patent information on Muga silkworm, Som and Mejankari. A combination of all these information makes this database a one-stop free, online resource for students and enthusiasts.

MugaSeqDB's website design and its in-built tools are another set of salient features of the database (Fig. 5.2). Its dynamic, responsive design (described in materials and methods) makes it easily accessible in a user-friendly. It also contains an exclusive search engine that helps in easily finding the target data via keywords, sequence or MugaSeqDB Identifiers. Similar advantage is provided by the presence of a flexible data browser in the website. Finally, it has hyperlinks to

an array of tools for data analysis like ClustalO, NCBI Blast, Simple Phylogeny, etc.

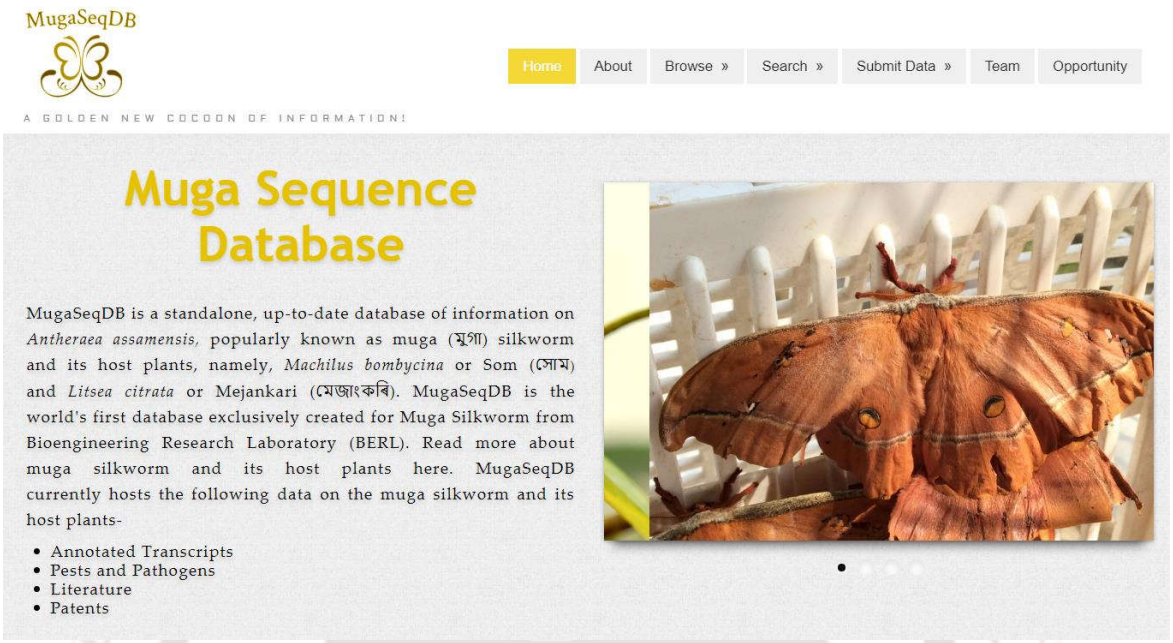


Fig. 5.2 Snapshot of the home page of MugaSeqDB

Other features of MugaSeqDB include the “Data submission”, “Feedback” and “Help” pages which make the database more amenable to use and contribute data from the users. The data submitted to the database will be scrutinized by the database administrator and published. This strategy is useful in keeping database up-to-date in terms of information as has been observed in the case of FlyBase [4].

5.4 CONCLUSION AND FUTURE PROSPECTS

MugaSeqDB is a one-stop database for information on muga silkworm (*A. assamensis*) and its two host plants, *L. citrata* and *M. bombycina.* It will act as a comprehensive, exhaustive resource of information for seri-researchers and enthusiasts. It’ll be key for dissemination of data on this endemic species of

silkworm. The in-built data submission by users will also help in enhancing the database.

The current information provided in the database is based upon de novo transcriptomes of muga silkworm and its host plant. In future, these data will be updated with more enriched genomic information as soon as their respective whole genomes are available. Data will also be updated on availability of any new information from its user-sourced data submissions or other secondary web-resources. Another aspect that will be improved further is the SEO (search engine optimization) of MugaSeqDB website for its improved visibility in popular search engines like Google or Yahoo.

REFERENCES

1. Kar S, Talukdar S, Pal S, Nayak S, Paranjape P, Kundu SC. Silk gland fibroin from indian muga silkworm *Antheraea assama* as potential biomaterial. *Tissue Eng Regen Med*. 2013;10(4):200-210. doi:10.1007/s13770-012-0008-6

2. Dutta S, Bharali R, Devi R, Devi D. Purification and characterization of glue like sericin protein from a wild silkworm *Antheraea assamensis* Helfer. 2012;1(2):229-233.

3. Kodama Y, Shumway M, Leinonen R. The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Res*. 2012;40(D1). doi:10.1093/nar/gkr854

4. Thurmond J, Goodman JL, Strelets VB, et al. FlyBase 2.0: the next generation. *Nucleic Acids Res*. 2019;47(D1):D759-D765. doi:10.1093/nar/gky1003

5. Singh D, Chetia H, Kabiraj D, et al. A comprehensive view of the web-resources related to sericulture. 2016;2016. https://academic.oup.com/database/article/doi/10.1093/database/baw086/2630457#87283070.

6. Li Y, Wang G, Tian J, et al. Transcriptome analysis of the silkworm (*Bombyx mori*) by high-throughput RNA sequencing. Gibas C, ed. *PLoS One*. 2012;7(8):e43713. doi:10.1371/journal.pone.0043713

7. The International Silkworm Genome Consortium. The genome of a lepidopteran model insect, the silkworm *Bombyx mori. Insect Biochem Mol Biol*. 2008;38(12):1036-1045. doi:10.1016/J.IBMB.2008.11.004

8. Dong Z, Dong F, Yu X, et al. Excision of nucleopolyhedrovirus form transgenic silkworm using the CRISPR/Cas9 System. *Front Microbiol*. 2018;9:209. doi:10.3389/fmicb.2018.00209

9. Cui Y, Zhu Y, Lin Y, et al. New insight into the mechanism underlying the silk gland biological process by knocking out fibroin heavy chain in the silkworm. *BMC Genomics*. 2018;19(1):215. doi:10.1186/s12864-018-4602-4

10. Xia Q, Li S, Feng Q. Advances in silkworm studies accelerated by the genome sequencing of *Bombyx mori*. *Annu Rev Entomol*. 2014;59(1):513-536. doi:10.1146/annurev-ento-011613-161940

CHAPTER 6

Comparative transportome study of Nosema, the causal organism of pebrine

CHAPTER 6

Comparative transportome study of Nosema, the causal organism of pebrine

A pre-print of this chapter has been published as:

Chetia H, Kabiraj D, Sharma S, et al. Comparative insights to the transportome of Nosema: a genus of parasitic microsporidians. bioRxiv, 2017; 110809. doi: 10.1101/110809

ABSTRACT

Nosema, a genus of intracellular parasitic microsporidia, causes pebrine disease in arthropods, including economically important insects like silkworms and honeybees. *Nosema* genomes are shaped by reductive evolution. They have lost some major metabolic pathways and depend on host-derived substrates. As an act of counterbalance, they have developed an array of transporter proteins that allow stealing biomolecules and metabolites from their hosts. Here, we have identified the complete transportome of four Nosema species, viz. *N. apis*, *N. bombycis*, *N. ceranae* and *N. antheraea.* Our results indicate that the transportomes of *N. apis, N. bombycis, N. ceranae* and *N. antheraea* have a dominant share of secondary carriers and primary active transporters. The comparatively rich and diverse transportome of *N. bombycis* indicates the role of transporters in its remarkable capability of host adaptation. We have further identified a set of twelve transporter families which seem to be of core

importance to the *Nosema* genus. This dataset includes ones that have a likely role in osmo-regulation, intra- and extra-cellular pH regulation, energy compensation and self-defense mechanism. We also identified a set of ten species-specific transporter families within the genus with possible implications in aiding species-specific adaptation. To our knowledge, this is the first ever intra-genus study on microsporidian transporters. Both these datasets constitute a valuable resource that can assist in development of inhibitor-based *Nosema* management strategies.

6.1 INTRODUCTION

Microsporidia is a group of unicellular obligate parasites which can infect a myriad of organisms including a few economically important insects like silkworms, honey bees etc. as well as humans [1,2]. Within this group, Nosematiidae is one of the most diverse of entomophagic microsporidia with ~150 genera reported so far [3]. The disease caused by *Nosema* is broadly called Pebrine, where destruction of insect midgut epithelial cells is a common pathological manifestation [4]. *Nosema* infect insects from different orders such as Hymenoptera (honey bees), Lepidoptera (silkworms), etc. Huge economic losses are incurred by countries of Europe, Asia and others every year due to mass death of these economically important host organisms which are utilized in sericulture, apiculture etc [5,6].

Microsporidia have highly reduced gene-dense genomes achieved via loss of the evolutionary sacrifice of the genes from many essential pathways (TCA cycle, metabolic pathways) to minimize biological complexity [7,8]. They possess an unstacked Golgi apparatus and a cryptic genome-less mitochondrion called mitosome. In order to compensate for their reduced metabolic capacity, they

utilize the host's inner metabolism to derive nutrition. They usually do this either by up-regulating the host's metabolic pathways via microsporidia-secreted factors and taking up nutrients from the host through their membrane-bound transporters [9]. So, the microsporidian transporters are an important component of their offense/defense. They span across membranes, facilitating exchange of substrates between host and microsporidia, thus helping in its sustenance within the host cytoplasm and bypass the energy-expensive biochemical pathways. An example of microsporidian transporters are the ATP transporters (NTT) that are ubiquitously present in all microsporidia known [10]. Genomic analyses have shown that despite undergoing genome reduction, some transporter proteins have been retained by microsporidia. These proteins usually transport crucial substrates like nucleotides, cations, sugars, ATP etc. and are possibly indispensable for the species [11]. This pathogenic advantage can be converted to a disadvantage if we can target and block the function of such crucial transporters. Deciphering the whole transportome of microsporidia can provide a head start to this venture.

Given the economic significance of silkworms as well as honeybees and the existing notoriety of their microsporidian pathogens discussed above, we focused to study the transportome of the Nosema genus. At present, the whole genome for three *Nosema* species, namely, *N. apis, N. bombycis* and *N. ceranae* is available in the NCBI Genome database (Table 6.1). The first genome to be published was that of *N. ceranae* in 2009 with a draft assembly size of 7.86 Mb [12]. *N. ceranae* has been found to be associated with "colony collapse disorder" in European honey bees, *Apis mellifera,* causing huge economic losses in apiculture sector. Since honey bees are principle

pollinators, their mass deaths have also affected agricultural productivity and sustainability. The second *Nosema* species with a sequenced whole genome is *N. bombycis* [13]. It has a GC-rich (~33%), gene-dense genome (~15.69 Mb) which is among the largest known *Nosema* genomes. Its aforementioned gene density has been attributed by abundant gene duplications, host-derived transposomal element proliferation and acquisition of bacterial genes via horizontal gene transfer. It is the causal organism of pebrine disease in domestic silkworms causing significant losses in sericulture industry [14]. *N. apis* is another microsporidian species responsible for death of European honey bees [15]. Apart from these three species, *N. antheraea* (also known as *N. pernyi*) which infects the wild and undomesticated tassar moth, *Antheraea pernyi*, has also been included in this study [13].

Here, we identified the total transportome of the four species of Nosema (*N. apis, N. bombycis*, *N. ceranae* and *N. antheraea*) and classified them into channels, secondary carriers and primary transporters based on the TC classification system [16]. Our study provided a snapshot of the core and species-specific transporters in *Nosema sp.s* and their potential roles in sustaining the obligate intracellular parasitic life of the species.

Table 6.1: Comparative genomic information of *Nosema apis*, *N. ceranae*, *N. bombycis* and *N. antheraea*

Microsporidia	Host (Order)	Genome size (mb)	NCBI ID	Reference
N. apis	Honeybees (Hymenoptera)	8.5	14500	15

N. ceranae	Honeybees (Hymenoptera)	7.86	931	12
N. bombycis	Silkworms (Lepidoptera)	15.69	11028	13
N. antheraea	Silkworms (Lepidoptera)	7.1	NA	13

6.2 MATERIAL AND METHODS

The whole proteome of *N. apis BRL01, N. bombycis CQ1 and N. ceranae BRL01* were retrieved from UniProt database [17]. Similarly, the draft proteome of *N. antheraea* was obtained from SilkPathDB [13]. The reference transporters were downloaded from TCDB (accessed January 2017). A local standalone blastP program was run using the TCDB dataset as a “blastdb” (database) and the four proteomes as query with an e-value cut off 0.001. The proteins with a percentage similarity of less than 30% and query coverage of 30 were discarded. Then, the proteins were further filtered out on the basis of presence or absence of transmembrane (TM) domains. The TM domains were detected using three different web-servers specifically used for TM prediction, namely, HMMTOP, TMHMMv2.0 and Phobius with default parameters [18–20]. Only those proteins whose TM domains were predicted by at least two of the three tools were selected for further analysis. Further, a two-way analysis of conserved domains of the filtered proteins was carried out using a local installation of InterProScan (version 5.17) and the web-server CD-Search based on CDD (Conserved Domain Database) [21,22]. Only those proteins whose predicted domains from

TCDB families matched with the predicted domains of InterProScan and CDD were retained as candidate transporters. These transporters were classified according the TCDB nomenclature (till the third position to indicate its respective family or superfamily). A flow chart depicting the method and online/offline tools implemented in this study is shown in Fig. 6.1.

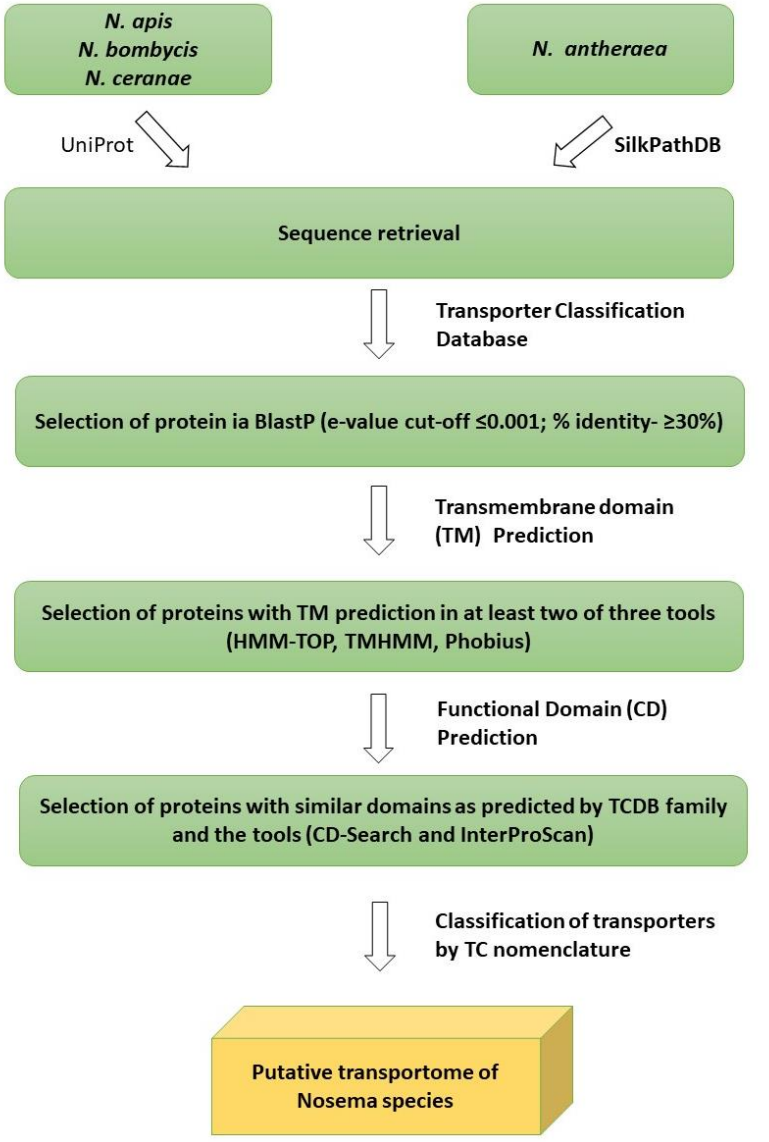


Fig. 6.1 Workflow followed to decipher the transportome of Nosema species

6.3 RESULTS AND DISCUSSION

Our analysis showed that the transportome of *N. apis*, *N. bombycis*, *N. ceranae* and *N. antheraea* consisted of 41, 78, 47 and 51 proteins respectively. Secondary carrier type facilitators (Class 2) were found to be the most abundant class of transporters in *N. apis, N. bombycis* and *N. ceranae* while primary transporters (Class 1) was found to be most abundant in *N. antheraea* (Fig. 6.2). Class 8 and 9 contains accessory proteins to other transporters and uncharacterized transporters, respectively, so they were not probed further in this study. The distribution of the transporter protein families and their substrates in *N. apis*, *N. bombycis*, *N. ceranae* and *N. antheraea* has been shown in Table 6.2, 6.3 and 6.4.

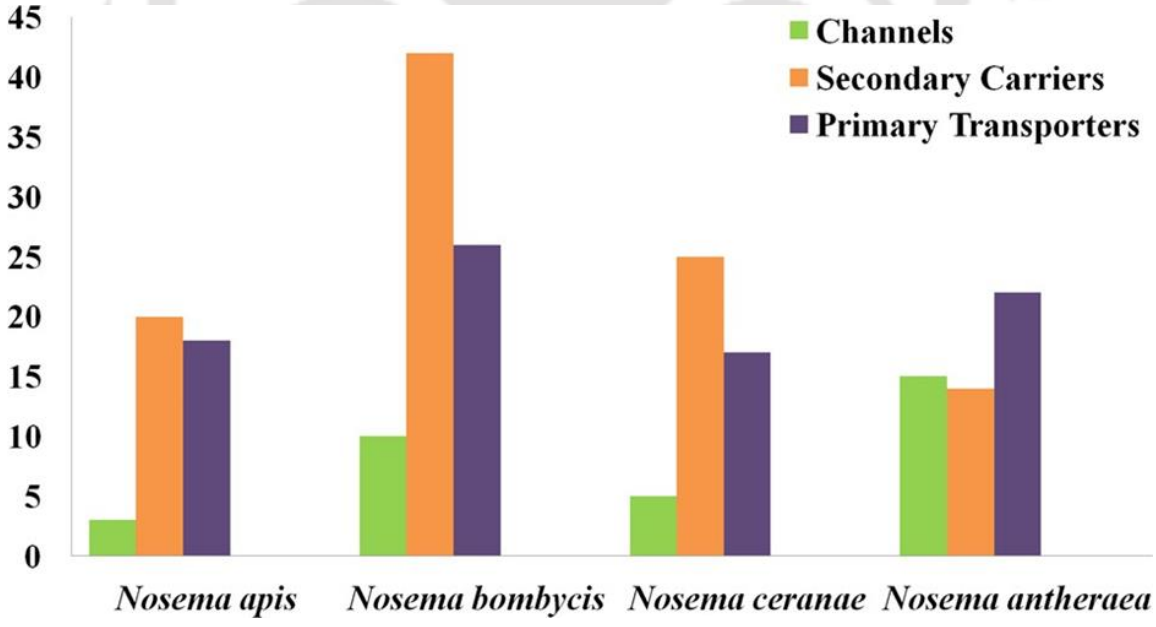


Figure 6.2- Class-wise distribution of the transportome of *Nosema apis*, *N. bombycis*, *N. ceranae* and *N. antheraea*

6.3A Class 1: Channels and Pores

Channels or pores which constitute the Class-1 of TCDB facilitate passive transport of substrates without the involvement of additional reaction and represents passive transport**.** Class 1 has channels mostly made up of α-helices or β-sheets spanning the lipid bilayer and forming a channel or pore in the organism's cell membrane to allow passage of solutes [16]. *Nosema* being a highly reduced obligate parasite lacks many basic biosynthetic pathways and relies upon its host for requirements like nucleotides which are building blocks of DNA and RNA [10]. Thus, passive transport, which is the dominant mode of transport for channels (Class 1) is an amenable way to import substrates.

Our study showed that a total of thirty-three proteins from the four *Nosema* species belonged to Class 1 out of which α-type channels were the most commonly present (Table 6.2). *N. antheraea* had the largest set of channels (15 no.s) among the four species and *N. apis* had the least (3 no.s). All four *Nosema* sp.s retained at least one transporter from the small conductance mechanosensitive ion channel (MscS) (1.A.23) and synaptosomal vesicle fusion pore (SVF-pore) family (1.F.1). MscS transporters are conserved within other microsporidian genomes like *E. cuniculi, E. bieneusi* and *A. algerae;* they represent a modest subset of *Nosema* transporters which are responsive to mechanical perturbations in the lipid bilayer, acting as mechanical switches [11].

SVF-Pore which was also present in all four *Nosema* species, is commonly found in yeasts and mammals and is involved in vesicle fusion. The presence of vesicles among microsporidia is not very common and intracellular sorting or transport of cargo proteins from the atypical microsporidian golgi complex

occurs by a mechanism that does not involve the participation of vesicles but rather tubular networks [23]. Expression of SVF-Pore proteins and avesicular transport shown in spores and consecutive intracellular stages have been established through RT-PCR in *Paranosema* genus which infects Orthoptera [24].

The conservation of both these families within four *Nosema* genomes indicates that these transporters might be essential for fulfilling the core features of a parasite life cycle such as avoiding mechanical distress or protein translocation [25]. These two families will be a part of the core transportome and have significance in the parasite's life cycle.

Four transporter families were found to be common in at least two *Nosema* species, namely, Non- selective cation channel-2 (NSCC2) family (1.A.15), Major Intrinsic Protein (MIP) family (1.A.8), CorA metal ion transporter (MIT) family (1.A.35) and Nuclear pore complex (NPC) (1.I.1). NSCC2 proteins are homologues of general protein secretory pathway in yeast microsomes and have been found to act as non-selective cation channels in mammalian cytoplasmic membranes [26]. MIP family transporters are aquaporins in native form, but are also capable of transporting carbohydrates, glycerol, urea, ion etc. by an energy- independent mechanism [27]. Both these α-type channels have been reported in other microsporidia like *Encephalitozoon cuniculi*, *Enterocytozoon bieneusi* and *Anncaliia algerae,* and hence, can be considered to play similar roles in *Nosema sps* [11]. Again, CorA transporters are similar to MscS transporters in substrate specificity, i.e., metal ions and this has been reported previously in *E. bieneusi* [28]. NPC homologues involved in transport across nuclear membrane were observed in *N. bombycis* and *N. antheraea* and

Table 6.2- Distribution of Class 1 transporters with their substrates for *Nosema apis (NAp), N. bombycis (NB), N. ceranae (NC)* and *N. antheraea (NAn)*

TC ID	Name	No. of Transporters				Substrates	Cellular Location
		NAp	NB	NC	NAn		
1.A.8	Major intrinsic protein (MIP) family	0	1	1	1	Water, Small carbohydrate (glycerol), Urea, NH3, CO2, H2O2 and ion	Plasma Membrane
1.A.15	Non-selective cation channel-2 (NSCC2) family	0	1	1	1	Na+ , K+ and Cs+	Endo membrane
1.A.23	Small conductance mechanosensitive ion channel (MscS)	2	3	1	4	Ions (K^+, Na^+, Ca^{2+} and Cl^-)	Plasma Membrane
1.A.35	CorA metal ion transporter (MIT)	0	0	1	1	Ions (Mg^{2+})	Plasma Membrane
1.B.12	Autotransporter-1 (AT-1)	0	1	0	0	Protein	Plasma Membrane
1.C.11.3	Hly III (Hly III) (Hly III)	0	0	0	1	Not Available	Not Available
1.F.1	Synaptosomal vesicle fusion pore (SVF-Pore)	1	3	1	2	Neurotransmitter, proteins, complex carbohydrates, small molecules such as ATP	Plasma membrane
1.I.1	Nuclear Pore Complex (NPC)	0	1	0	4	Small proteins	Nuclear membrane
1.M.1	Rz/Rz1 Spanin1 (Rz(1))	0	0	0	1	Hydrocarbon	Periplasm
	TOTAL = 33	**3**	**10**	**5**	**15**		

not in other two microsporidia. Their presence has been also been reported in developmental stages of microsporidia previously [29]. These channels are conserved in eukaryotes, comprise of a number of nucleoporins and are responsible for nuclear cytoplasmic exchange.

Apart from these, three species-specific Class 1 transporters (present only in one of the microsporidia) were identified, namely, AT-1 family in *N. bombycis*; Hly III Family (1.C.113) and Rz/Rz1 Spanin1 Family (1.M.1) in *N. antheraea* (Table 6.1). Both AT-1 and Rz(1) are involved in periplasmic transport of protein and hydrocarbons, respectively. The homologue of Hly family is a putative Hemolysin III-like protein which is currently uncharacterized.

Overall percentage of channels present within the four species is quite variable. Honeybee (Hymenoptera) infecting species had less channels than the ones infecting silkworms. This may be attributed to Hly, Rz(1) and AT-1 families and the larger set of homologues of NPC and MscS families in *N. antheraea* and *N. bombycis.* Our results pose a new question as to why or how does presence of more channels benefit the silkworm-infecting Nosema species, whose answer lies in future experimentations.

6.3B Class 2: Secondary carrier-type facilitators

Secondary carrier-type facilitators, also known as electrochemical-potential driven transporters, represent Class-2 of TCDB and employ a carrier-mediated process involving uniporters, symporters and antiporters for transport. We observed a total number of 101 transporters of *Nosema* belonging to Class 2 (Table 6.3). *N. bombycis* had the highest number of secondary transporters (forty-two) out of all the four species studied here. All the *Nosema* transporters

from Class 2 were found to be porters (uniporters, symporters and antiporters) and other sub-classes like ion-gradient driven energizers or trans-compartment lipid carriers were absent.

Out of the fifteen families, six were commonly found in all the four species, viz. Major facilitator superfamily (MFS) (2.A.1), Cation diffusion facilitator (CDF) superfamily (2.A.4), Drug/metabolite transporter (DMT) superfamily (2.A.7), ATP:ADP antiporter (AAA) (2.A.12), Proton dependent oligopeptide transporter (POT/PTR) (2.A.17) and Amino acid/auxin permease (AAAP) (2.A.18) families. It is known that all microsporidian genomes encode two or more MFS transporters that have specificity for sugars imported from hosts [10]. One of the MFS homologues belonging to sugar porter family (SP) (2.A.1.1) can take up environmental glucose to support its own metabolism. 11 homologues of this family were present in *N. bombycis*. Other MFS families observed in *Nosema,* namely, Proteobacterial intraphagosomal amino acid transporter (Pht) (2.A.1.53), Unidentified Major Facilitator-14 (UMF14) (2.A.1.65) and Drug-H+ antiporter (DHA1) (2.A.1.2), are also found in yeast and bacterial pathogens. The latter two have sequence similarities with multidrug-resistant transport proteins of yeast and *E. coli* as well as purine transporters [30]. The second conserved Class 2 family is CDF which is responsible for heavy metal ion efflux and mostly implicated in instances where human health and bioremediation is concerned [31]. It has also been reported to be involved in Co^{2+}-ion uptake in *Saccharomyces cerevisae* [32]*. E. cuniculi* and *E. bieneusi* comprise at least one transporter from this CDF family [28]. Again, DMT superfamily plays a role in host adaptation and is associated with transport of variety of metabolites (nucleotide, sugar etc.) from host cytoplasm [33]. Out of the 31 members of this superfamily,

Table 6.3- Distribution of Class 2 transporters with their substrates for *Nosema apis* (NAp), *Nosema bombycis* (NB), *N. ceranae* (NC) and *N. antheraea* (NAn)

TC ID	Name	No. of Transporters				Substrates	Cellular Location
		NAp	NB	NC	NAn		
2.A.1	Major facilitator superfamily (MFS)	3	11	4	5	Various (sugars, polyols, Krebs cycle and glycolytic metabolite, amino acids, peptides, osmolites, nucleoside, anions	Plasma Membrane
2.A.3	Amino acid- polyamine- organocation (APC)	0	4	0	0	Protein	Cytoplasmic side of transmembrane
2.A.4	Cation diffusion facilitator family (CDF)	2	1	1	1	Ions (heavy metals)	Endomembrane
2.A.5	Zinc (Zn^{2+})-iron (Fe^{2+}) permease (ZIP)	2	0	1	0	Ions (Fe2+ and Zn2+)	Plasma Membrane
2.A.6	Resistance- nodulation-cell division (RND)	0	1	1	0	Various (heavy metals, drugs, membrane, lipooligosaccharides, unfolded protein, lipids, sterols)	Endomembrane
2.A.7	Drug/metabolite transporter (DMT)	1	6	3	3	Various (heavy metals, multiple drugs, lipooligosaccharides, unfolded proteins, lipids, sterols)	Endomembrane (Periplasm)
2.A.12	ATP:ADP antiporter (AAA)	4	3	4	2	ATP/ADP	Plasma membrane
2.A.17	Proton-dependent oligopeptide transporter (POT/PTR)	1	1	1	1	Protein/Oligopeptides	Plasma membrane
2.A.18	Amino acid/auxin permease (AAAP)	4	10	7	1	Amino acids	Plasma membrane
2.A.36	Monovalent cation:proton antiporter-1 (CPA1)	0	0	1	0	Ions	Plasma membrane
2.A.38	K^+ transporter (Trk)	0	1	0	0	Drug/Metabolite	Plasma membrane
2.A.50	Glycerol uptake (GUP)	0	1	0	0	Others (Glycerol)	Plasma membrane
2.A.53	Sulfate permease (SulP)	2	2	1	0	Ions (SO42-)	Plasma Membrane
2.A.92	Choline transporter-like (CTL)	1	1	0	1	Others (Choline)	Plasma Membrane
2.A.94	Phosphate permease (Pho1)	0	0	1	0	Ions (Pi)	Plasma Membrane
	Total = 101	**20**	**42**	**25**	**14**		

four were found within *Nosema,* viz., Triose-phosphate Transporter (TPT) (acts as antiporter, exchanges organic phosphate ester for inorganic phosphate), UDP-Galactose:UMP Antiporter (UGA), UDP-N-Acetylglucosamine:UMP Antiporter (UAA) (both act as antiporter, exchanges a nucleotide-sugar for nucleotide) and NIPA Mg^{2+} Uptake Permease (NIPA) (mediates Mg^{2+} intake, causes a neurodegenerative disorder called hereditary spastic paraplegia in humans and has homologues in other eukaryotes). The fourth common transporter is the AAA family, which has been reported in other microsporidia as well as bacteria, fungi and plants [28,34]. It is known that microsporidia lack an ideal mitochondrion like structure and consequently, lacks the electron transport chain. Instead *Nosema* genus has acquired AAA transporters which can act as a part of the compensatory mechanism for the lack of the same and possibly have a role in importing ATP from host cytoplasm. AAAP family, another important transporter family, which is conserved across the *Nosema* genus, are associated with vacuolar bidirectional symport or antiport of various amino acids as well as ions (H^+, Na^+ etc.) in lower eukaryotes such as yeast [35,36]. Finally, the proton-dependent oligopeptide transporter (POT/PTR) family involved in substrate (small peptides, oligopeptides, proteins etc.) efflux coupled to H^+ antiport.

Apart from these core transporter families, four others were found in at least two of the *Nosema* species. Two of these families transport different types of ions, namely, ZIP family (2.A.5) involved in acquisition of metal ion, especially in intake and maintenance of homeostasis of Zn^{2+} and SulP family involved in inorganic anion uptake. Other than these ion transporters, Resistance-nodulation-cell division (RND) superfamily (2.A.6), which is homologous to solute carrier family of mammals and can transport a range of substrates like peptides, amino acids

(Histidine), nitrates etc. and the choline transporter-like (CTL) family (2.A.92), a solute carrier family for choline, were also identified.

Five species-specific Class2 transporters were also identified in this process. Of these five, four families were present in *N. bombycis,* namely, amino acid-polyamine-organocation (APC) (2.A.3) (functions as arginine/ornithine or cystine/glutamate antiporter to maintain cellular redox balance as well as cysteine/glutathione levels), monovalent cation:proton antiporter-1 (CPA1) family (2.A.36) involved in Na^+:H^+ exchange, K^+ transporter (TrK) (2.A.38) (K^+:H^+ symport) and glycerol uptake (GUP) (2.A.50) (implicated in glycerol and amino acid uptake in bacteria, yeast etc.) [37]. *N. ceranae* retained the fifth species-specific secondary facilitator family, namely, phosphate permease (Pho1) (2.A.94) which carries out inorganic phosphate transport in plants.

As discussed above, MFS, DMT and AAAP families were conserved across the four species under study and have also been reported in other microsporidian species. Interestingly, the number of homologues from these three classes were much greater in *N. bombycis*, leading to a spike in the total Class 2 transportome of the species. There are two possible explanations for this disparity. Firstly, the genome of *N. bombycis* is larger than the rest due to events of duplication and horizontal gene transfer. Secondly, it is able to infect multiple silkworm species and existence of a better substrate uptake mechanism, in the form of greater number of secondary facilitators, may contribute towards its broader host specificity.

6.3C Class 3: Primary active transporters

Primary active transporters of Class 3 drive transport of a solute against a

concentration gradient using a primary source of energy. Around 32-37 primary transporters have been found to be present in other microsporidia (*E. cuniculi, E. bieneusi, A. algerae*) [28]. We identified eighty-three primary transporters in our subset of *Nosema* genus (Table-6.4). All of these homologues drive the active uptake and/or extrusion of a solute or solutes via hydrolysis of diphosphate bond of inorganic pyrophosphate, ATP, or nucleoside triphosphate. Oxidation-reduction, methylation or decarboxylation driven transporters were not found in the studied *Nosema* genus.

As per our observation, the largest group of Class 3 transporters were the ATP-binding cassette (ABC) transporters (3.A.1) superfamily, conserved across *Nosema* as well as other genera of microsporidia like *Encephalitozoon, Enterocytozoon* and *Anncaliia* genera [11]. ABC transporters are associated with transport of various substrates like peptides, lipid, ions, drugs etc.

A total of thirty-six ABC transporters have been identified in *Nosema*, whereas previous studies showed the presence of four sub-families in *E. cuniculi* [38]. Within the ABC superfamily, Heavy metal transporter (HMT) family (also known as ABCB) (3.A.1.210) was the most abundant (21 no.s in the four *Nosema sp.s*), followed by Eye pigment precursor transporter (EPP) family (also known as ABCG) (3.A.1.204) (10 no.s in the four *Nosema sp.s*). Heavy metal transporters have been observed in *E. cuniculi* and *E. intestinalis* previously [38]. Similarity between these putative transporters and that of yeast ATM1 protein which is a prototype for this subfamily suggests that they might be carrying out similar function in Fe–S cluster export. A related hypothesis proposes the role of cryptic mitochondria of microsporidia in iron-sulfur biogenesis [9]. Similarly,

Table 6.4- Distribution of Class 3 transporters with their substrates for *Nosema apis (NAp), Nosema bombycis (NB), N. ceranae (NC)* and *N. antheraea (NAn)*

TC ID	Name	No. of Transporters				Substrates	Cellular Location
		NAp	NB	NC	NAn		
3.A.1	ATP binding cassette (ABC) superfamily	6	10	9	11	Various	Plasma Membrane/Mitosome
3.A.2	H+- or Na+- translocating F- type, V-type and A-type ATPase (F-ATPase) superfamily	3	4	1	3	Ions (H^+/ Na^+)	Endomembrane and Plasma membrane
3.A.3	P-type ATPase superfamily	4	4	5	3	Cations	Endo and Plasma membrane
3.A.5	General secretory pathway (Sec)	4	5	1	2	Unfolded protein	Endomembrane
3.A.15	Outer membrane protein secreting main terminal branch (MTB)	0	1	0	0	DNA/Proteins	Plasma membrane
3.A.16	Endoplasmic reticular retro-translocon (ER-RT)	1	1	0	2	Protein	Endo membrane (endoplasmic reticulum)
3.A.20	Peroxisomal Protein Important (PPI)	0	1	1	0	Protein	Endo Membrane (Peroxisome)
3.A.25	Symbiont- specific ERAD-like Machinery (SELMA)	0	0	0	1	(Pre-) Protein	Outer Membrane of plastids
Total = 101		**18**	**26**	**17**	**22**		

EPP family of transporters have been identified previously in *E. cuniculi* which is speculated to involved in guanine and tryptophan transport in *Drosophila melanogaster* [39]. Other than ABC superfamily, the classes 3.A.2 (H+- or Na+-translocating F-type, V-type and A-type ATPase (F-ATPase)), 3.A.3 (P-type ATPase (P-ATPase)) and 3.A.5 (General secretory pathway (Sec)) families/superfamilies were conserved across the four *Nosema* species. F-ATPase superfamily is conserved from bacteria to eukaryotes and is associated with vacuolar transport of H^+ and Na^+. Only V-type ATPases were identified in *Nosema* genus. P-ATPases identified in this study were the ones involved in phospholipid translocation and cation transportation [40]. Again, homologues to yeast Sec-SRP complex (3.A.5.8) and mammalian Sec-SRP complex were observed in the fourth conserved and abundant transporter family in Class 3, i.e., Sec family, including Sec61 (α and γ) subunits and sec63 translocase subunits. Sec61 complex is the general secretory (Sec) pathway for protein secretion into ER with its subunits Sec61α, Sec61β, and Sec61γ. This complex has been identified previously in *N. bombycis*, *E. cuniculi* and *Antonospora locustae* [41].

Two other families from Class 3 were found in at least two of the four organisms studied here, namely, Endoplasmic reticular retrotranslocon (ER-RT) (3.A.16) and Peroxisomal Protein Importer (PPI) (3.A.20) families. ER-RT transport proteins are abundant in *N. antheraea* and are supposedly associated with ER-associated degradation system which involves translocation of misfolded proteins from ER to cytoplasm for degradation. Microsporidia is known to lack any organelle like peroxisomes and the role of PPI transporters (3.A.20) needs in-depth analysis to understand the functions and relevance to its survival or other functions. Two species-specific unique transporter families from Class 3 were

also found in *N. bombycis* and *N. antheraea*, indicating that it still retains several transporter genes which are absent in its *Nosema* counterparts. The outer membrane protein secreting main terminal branch (MTB) family (3.A.15) which is present only in *N. bombycis* and it shares 71% sequence identity with the pulF protein of *Klebsiella pneumoniae*. Presence of *K. pneumonia* in silkworm gut has been reported previously; the presence of this unique transporter family with bacterial-origin in *N. bombycis* is a probable case of horizontal gene transfer due to co-localization at a shared niche. The acquisition of this transporter could probably aid *N. bombycis* in DNA and protein transport. Another species-specific transporter family found only in *N. antheraea* was the symbiont-specific ERAD-like Machinery (SELMA) (3.A.25) which transports nucleus-encoded pre-protein complexes in human pathogens, *Plasmodium falciparum* and *Toxoplasma gondii.*

6.4 CONSERVED AND UNIQUE TRANSPORTERS IN *NOSEMA*: HOW ARE THEY RELEVANT?

Microsporidia, being an obligate parasite with highly reduced gene content, is incapable of carrying out numerous processes that free-living organisms like yeast, bacteria etc. can carry out. These include Kreb's cycle, electron transport chain, nucleotide synthesis, etc. With genome reduction via gene loss, the transportome diversity of a microsporidia should ideally decrease. Despite this, a set of core transporters like permeases allowing passage of sugar, glucose, metal ions viz. Mg^{2+}, Ca^{2+}; polyamine transporters; ion transporters with metal ion specificity; sulfate transporters etc. are commonly retained by any microsporidia [2]. *Nosema,* like any other microsporidia, is also dependent on the host cytoplasmic contents for survival and proliferation. To obtain a plausible

view of the core transporters of *Nosema* genus, we compared the transportome of *N. apis, N. bombycis, N. ceranae* and *N. antheraea* (Fig. 6.3). Then, we compared the existing data on reviewed and unreviewed microsporidian transporters from UniProt to the complete set of *Nosema* transporters. Our study revealed that the *Nosema* genus has a core set of transporter families conserved among them (Figure 6.3 and 6.4). This set of twelve transporter families (ABC, DMT and MFS superfamily, F-ATPase, P-ATPase, Sec, MscS, CDF, SVF-Pore, POT/PTR, AAA and AAAP family) can be perceived to be crucial for a typical microsporidian life (Figure 6.4). Except the SVF-Pore family, all others have been reported in at least one microsporidian species other than those in *Nosema* genus. Three of these families are involved in ion transport,

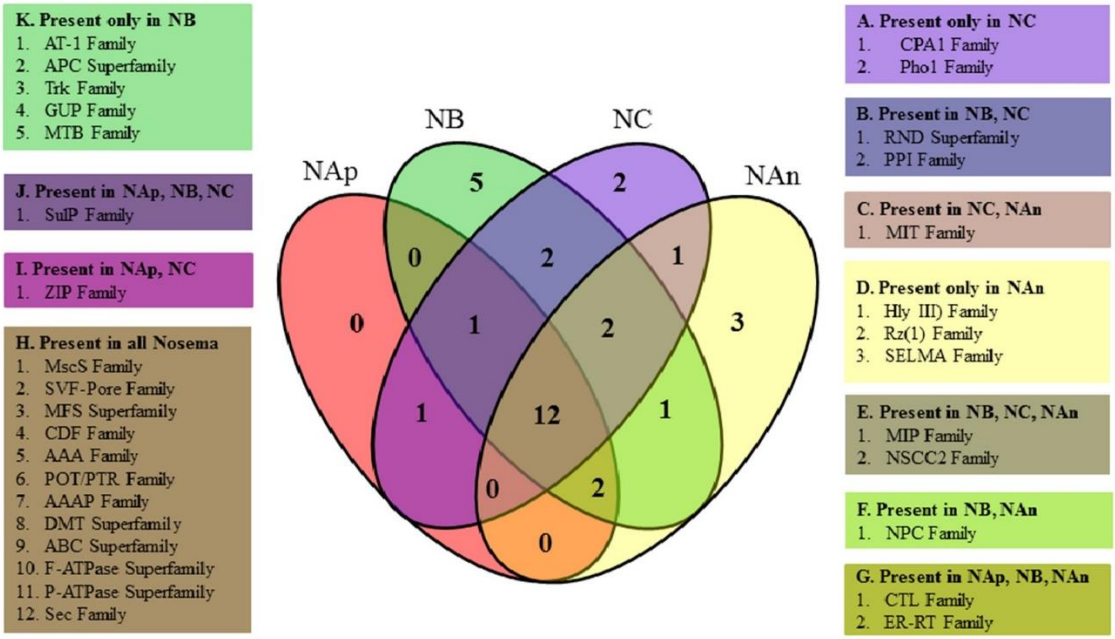


Figure 6.3- Transporter family distribution among the *Nosema* genus: Venn diagram showing shared and unique transporter families among four *Nosema* species, viz. NAp- *Nosema apis,* NB- *N. bombycis,* NC- *N. ceranae,* NAn- *N. antheraea*

namely, the energy-independent MscS family and the energy-dependent F and P-ATPase families. Recent research on microsporidia like *T. hominis*, *N. parisii* and *E. cuniculi* has reported that microsporidian MscS have two distinct origins, one is eukaryotic MscS1 proteins and another is bacterial-like microsporidian MscS2 [42–44]. MscS proteins can regulate osmotic homeostasis at the cell surface during different life stages by opening or closing a channel permeable to water and small ions in response to mechanical deformation of the cell membrane, such as that caused by physical or osmotic pressure [25]. It is to be noted that increase or decrease osmotic pressure in an intracellular parasite can be related to increase or decrease in membrane tension to speed up or delay its egress from a host cell [45]. This means that MscS family could have a probable role in *Nosema* infection process. Again, the V-type ATPase (from F-ATPase family) and the P-ATPases mediate ATP-dependent H^+/Na^- ion transport and have been widely observed in other microsporidian species across different genera such as *Anncaliia, Encephalitozoon, Enterocytozoon* etc [28]. Both these families are capable of cation efflux and can be speculated to be related to pH regulation, i.e, maintenance of resting pH and recovery from pH dysregulation inside the gut of a host organism by H^+ extradition. This phenomenon has been experimentally observed previously in protozoa such as *Plasmodium falciparum* [46]. Similar to *P. falciparum*, the hosts for *Nosema* genus are also insects and microsporidian spore germination occurs in various parts of the midgut or gut epithelial cells in insects of Hymenoptera and Lepidoptera [2]. Presence of these transporters probably helps *Nosema* sp.s in cation uptake, not only for nutritional or metabolic purposes but also survival amidst its hosts. Similar roles may also be played by the other core ion

transporter family, CDF and as well as less conserved SulP, NSCC2 etc.

The MFS superfamily is one of the predominant transporter families found in microsporidia and study of the *Nosema* genus provided us a similar observation. All known microsporidian genomes encode three or more MFS transporters with specificity for sugars likely to be acquired from hosts [42]. Microsporidian genome retains the pathways like glycolysis, pentose phosphate pathway, etc. but expression of the same varies from species to species and also from the habitat nature of the parasite (aquatic/terrestrial) [47]. In two of the *Nosema* species (*N. apis* and *N. ceranae*) studied here, it has been observed that sugar utilization increases along with spore counts within the host, resulting in energetic stress to the host that translates into several abnormal behavioral manifestations [48]. These studies, combined with our results portray the possibility that pathways for sugar uptake and utilization for ATP synthesis could still be functional within the *Nosema* genus (except *N. ceranae* which lacks this gene according to genome analysis). Presence of a terminal electron acceptor enzyme found in mitosomes, Alternative oxidase, has been hypothesized to be present in some microsporidia, aiding in glycolytic pathway of energy generation [49]. Other than sugar, MFS family also transports other substrates like polyols, amino acids, osmolytes, drugs, neurotransmitters, Krebs cycle metabolites, phosphorylated glycolytic intermediates, inorganic anions, etc. making it an inevitable part of the *Nosema* as well other microsporidian life cycle. Another substrate importing transporter family which was less conserved among the studied genus was the MIP family.

Among the highly conserved families, Sec, POT/PTR, SVF-Pore families were

involved in polypeptide/protein transport. Homologues of Sec family have been reported in microsporidia previously [28]. It was hypothesized that microsporidia exploit a protein transport and secretion machinery which is similar to yeast or mammalian cell [2]. However, even this machinery has undergone severe loss of non-essential genes. *E. cuniculi* genome reportedly, has all the components of the Sec61 translocon channel along with other associated proteins like Sec62-63, Hsp70 etc. which make translocation of proteins into ER possible [8]. Our analysis discovered homologues of Sec61 (α and γ) subunits and sec63

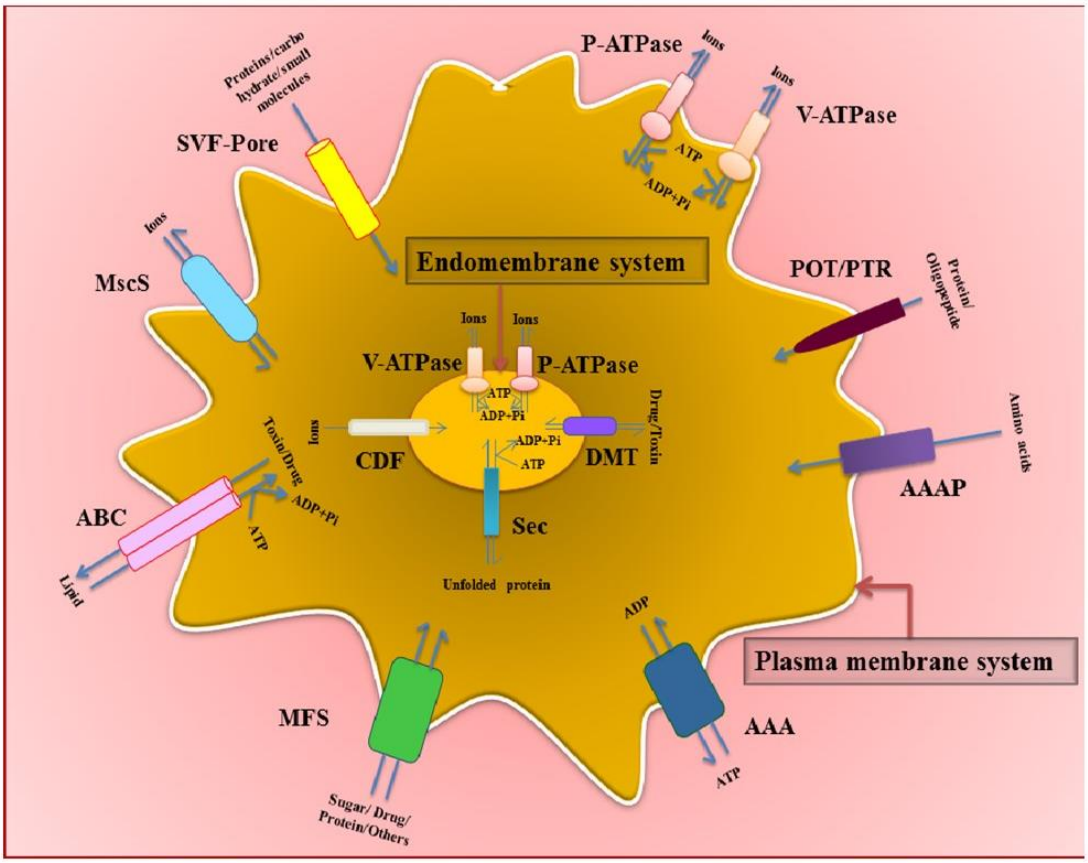


Figure 6.4- Diagrammatic representation of a typical *Nosema* cell with the core set of transporters conserved within the *Nosema* genus *(Note- The images used to denote the various type of transporter families doesn't imply their actual structure within membrane)*

translocase subunits and high conserved nature of these homologues across the four species of *Nosema* genera suggests that this pathway might be functional in *Nosema*, carrying out co-translational translocation of polypeptide chains into the ER lumen. Another complimentary mechanism serving this purpose is the SNARE protein mediated vesicular fusion. Homologues of the SVF-Pore family (1.F.1.1) that consists of proteins from the SNARE complex was conserved across *Nosema* (during our study) and *E. cuniculi* (from the UniProt database). Apart from the functions mentioned above, these transporter families might also be involved in other forms of intracellular transport and exocytosis.

Multiple studies have suggested that the *de novo* synthesis pathway for amino acids is lost in microsporidians, making them partially dependent on host-derived amino acids and partially on self-mediated inter-conversion [42]. In such a case, transporters with differential specificity for differently charged amino acids like the ones in AAAP family can be of immense importance for the organism. Apart from the functionalities of this family discussed earlier, these core transporters also have a role in the pathogenicity of *Nosema.* Usage of host-derived amino acids disrupts its levels within the host, indirectly putting it in energetic stress. The life cycle of *Nosema* demands usage of considerable amounts of amino acids, for e.g. during spore wall or polar tube formation. Whether the levels of amino acid depletion in host are relative to the rate of infection by *Nosema* can be experimentally determined, as has been done in case of sugar usage; this will provide another measure of energetic stress that this intracellular parasite places on its insect hosts.

As for the most important energy component of any organism, i.e., ATP, microsporidia are dependent on hosts (although presence of an alternative glycolytic pathway has been hinted previously) and the conserved AAA family is capable of serving this purpose for *Nosema* and its microsporidian counterparts. Regarded as a genomic hallmark for microsporidia, these transporters have similarities to bacteria and are speculated to have been acquired by horizontal gene transfer led by co-existence. As discussed in earlier, AAA proteins can exchange ATP for ADP, gaining energy in form of an extra pyrophosphate bond. These ATP exchanges along with that of sugar and amino acids further distresses host metabolism. Therefore, these transporters have been used as a target of gene silencing to reduce parasite load in Nosemosis, leading to favorable changes in host physiology [50]. The putative transporters discovered in this study can be further characterized and experimentally tested as novel gene silencing targets. Again, establishing the importance of DMT superfamily within *Nosema* genus and microsporidia, as a whole, can be related to the myriad of substrates it deals with. For example- Mg^{2+} ion transported by NIPA family can be utilized by *Nosema* augmenting the adherence of spore wall to host cell surface as observed in *E. cuniculi* glycosaminoglycans [51]. Endosomal transporter families were also found amongst other members of DMT superfamily, including those for golgi apparatus. However, these organelles are highly reduced or absent in microsporidia and more evidence on their presence is required before arriving at any conclusion. Finally, the ABC superfamily, discussed in Section 3.3, is found in many other microsporidia and the present study concurs with this fact. The presence of this family in the microsporidian genomes, like *E. cuniculi*, have been confirmed but with unknown specificity [38].

ABC transporters, reportedly act as exporters in certain parasitic protists (the classification that *Nosema* previously belonged to) but they might have a role to play in import or retrieval of host nutrients as well; its presence in *Nosema* and other microsporidia indicates the same. Its involvement in unidirectional transportation of substrates including extrusion of anti- parasitic molecules within eukaryotic parasites as a measure of self-defense is a possibility.

From the standpoint of an individual species, *N. bombycis* appears to be the most endowed or rather, most evolved microsporidian from *Nosema* genus. It has a greater number of transporter and probably have an advantage over the other three species studied here in terms of survival, pathogenicity and proliferation. This might be the reason behind the broad range of domesticated and wild Lepidopteran hosts that it can infect; susceptible insect families include Bombycidae, Noctuidae, Pieridae, Arctiidae and Crambidae [52,53]. *N. bombycis* also has the largest subset of unique, species-specific transporters, which could be result of adaptive evolution or horizontal gene acquisition from co-existing gut microbes [2]. It is closely followed by *N. antheraea* (51 no.s) in terms of transporter count. The host of *N. antheraea* is a wild silkworm species, *A. pernyi,* unlike the hosts of *N. apis* and *N. ceranae* species studied here (Table-6.1). Like *N. bombycis,* this organism also has three unique transporter families which have not been reported in other microsporidians till now.

Observation of the transporters according to host specificity shows that the share of ion transporters differs between the honey bee and silkworm parasites (Refer Table 6.2, 6.3 and 6.4). Out of the four observed ion transporter families in *N. apis*, three are conserved while one is common with *N. ceranae* only.

However, unique ion-transporter families discovered in *N. bombycis* points out that it may involve these transporters in pH or osmotic stress regulation, thus conferring it some added advantage in its survival within the cytoplasm of an array of insect hosts. The absence of same in *N. apis* may confer a disadvantage in its host adaptation abilities.

Close observation of transporter distribution across the four species shows a pattern where *N. apis, N. bombycis and N. ceranae* has a smaller number of porters/channels (Class 1) and a greater number of secondary carriers (Class 2) (Figure 6.2). However, *N. antheraea* transportome shows a reverse pattern. Within the core set of ten transporter families in *N. antheraea,* number of transporter proteins is reduced in AAA, AAAP, MFS etc. Again, the number of homologues of NPC family within *N. antheraea* is high, increasing the share of channels/porters within its transportome. It is somehow related to the fact that *Nosema* uses host nucleus as its developmental niche [9]. Again, *N. antheraea* has the highest number of primary transporters (22). How does a highly host-dependent, energy-efficient intracellular parasite like *N. antheraea* keep up with the energetic costs for Class 3 when the number of ATP transporter is as low as two (Table-6.3)? Assuming that all the deduced transporters of Class 3 are functional, a striking hypothesis can be made from the above observations- *N. antheraea* may have compensated for the dearth of Class 2 transporters by scaling-up the transporters from Class 1 and 3. From Table-6.1, it is apparent that *N. antheraea* has the smallest genome among four organisms and among other genes, it may have shed the genes for secondary carrier facilitators. In terms of species-specific transporters, *N. antheraea* and *N. bombycis* are richer than *N. apis* and *N. ceranae.* However, the ones exclusively found in *N.*

antheraea like Rz(1) or Hly III have not been reported earlier and requires further analysis. The species-specific transporters of Nosema should also be investigated further (Fig. 6.4). Our observations are largely based on the outputs of an *in-silico* pipeline and requires greater in-depth experiments before garnering any conclusive view.

6.5 CONCLUSION

The current scenario of apiculture and sericulture is threatened by microsporidians of *Nosema* genus and new strategies are required to tackle these parasites. Since microsporidia are entirely dependent for some crucial substrates like ATP, sugar, nucleotides etc. on its host, the transporters for these substrates can act as critical components of a pest management strategy. The broad spectrum of transporter proteins within the *Nosema* genus identified by us using available data can act as a valuable resource for future studies. This includes understanding microsporidian biology, inner mechanism and its relation to host variability.

REFERENCE

1. Silveira H, Canning EU, Shadduck JA. Experimental infection of athymic mice with the human microsporidian *Nosema corneum*. Parasitology. 2009;107(05):489. doi:10.1017/S0031182000068062

2. Weiss LM, Becnel JJ. Microsporidia: pathogens of opportunity. Wiley; 2014.

3. Wittner M. The microsporidia and microsporidiosis. (Weiss LM, Wittner M, eds.). American Society of Microbiology; 1999. doi:10.1128/9781555818227

4. Hges M, Martín-Hernández R, Meana A. *Nosema ceranae* in Europe: an emergent type C nosemosis. Apidologie. 2010;41(3):375-392. doi:10.1051/apido/2010019

5. Jeffree EP, Allen DM. The influence of colony size and of nosema disease on the rate of population loss in honey bee colonies in winter. J Econ Entomol. 1956;49(6):831-834. doi:10.1093/jee/49.6.831

6. Farrar Cl. Nosema losses in package bees as related to queen supersedure and honey yields. J Econ Entomol. 1947;40(3):333-338. http://www.ncbi.nlm.nih.gov/pubmed/20264495.

7. Keeling PJ, Corradi N. Shrink it or lose it: balancing loss of function with shrinking genomes in the microsporidia. Virulence. 2(1):67-70. http://www.ncbi.nlm.nih.gov/pubmed/21217203.

8. Katinka MD, Duprat S, Cornillot E, et al. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature. 2001;414(6862):450-453. doi:10.1038/35106579

9. Corradi N. Microsporidia: Eukaryotic Intracellular parasites shaped by gene loss and horizontal gene transfers. Annu Rev Microbiol. 2015;69:167-183. doi:10.1146/annurev-micro-091014-104136

10. Heinz E, Hacker C, Dean P, et al. Plasma membrane-located purine

nucleotide transport proteins are key components for host exploitation by microsporidian intracellular parasites. PLoS Pathog. 2014;10(12):e1004547. doi:10.1371/journal.ppat.1004547

11. Peyretaillade E, Boucher D, Parisot N, et al. Exploiting the architecture and the features of the microsporidian genomes to investigate diversity and impact of these parasites on ecosystems. Heredity (Edinb). 2015;114(5):441-449. doi:10.1038/hdy.2014.78

12. Cornman RS, Chen YP, Schatz MC, et al. Genomic analyses of the microsporidian *Nosema ceranae*, an emergent pathogen of honey bees. PLoS Pathog. 2009;5(6):e1000466. doi:10.1371/journal.ppat.1000466

13. Pan G, Xu J, Li T, et al. Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. BMC Genomics. 2013;14(1):186. doi:10.1186/1471-2164-14-186

14. Ishihara R, Fujiwara T. The spread of pebrine within a colony of the silkworm, *Bombyx mori* (Linnaeus). J Invertebr Pathol. 1965;7(2):126-131. doi:10.1016/0022-2011(65)90023-6

15. Chen Y ping, Pettis JS, Zhao Y, et al. Genome sequencing and comparative genomics of honey bee microsporidia, *Nosema apis* reveal novel insights into host-parasite interactions. BMC Genomics. 2013;14(1):451. doi:10.1186/1471-2164-14-451

16. Saier MH, Tran C V, Barabote RD. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res. 2006;34(Database issue):D181-6. doi:10.1093/nar/gkj001

17. UniProt Consortium TU. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2012;40 (Database issue):D71-5. doi:10.1093/nar/gkr981

18. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 2001;305(3):567-580.

doi:10.1006/jmbi.2000.4315

19. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001;17(9):849-850.

20. Käll L, Krogh A, Sonnhammer ELL. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res. 2007;35(Web Server issue):W429-32. doi:10.1093/nar/gkm256

21. Jones P, Binns D, Chang HY, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-1240. doi:10.1093/bioinformatics/btu031

22. Marchler-Bauer A, Derbyshire MK, Gonzales NR, et al. CDD: NCBI's conserved domain database. Nucleic Acids Res. 2014;43 (Database issue):D222-6. doi:10.1093/nar/gku1221

23. Beznoussenko GV, Dolgikh VV, Seliverstova EV, et al. Analogs of the Golgi complex in microsporidia: structure and avesicular mechanisms of function. J Cell Sci. 2007;120(Pt 7):1288-1298. doi:10.1242/jcs.03402

24. Dolgikh VV., Senderski IV., Pavlova OA, Beznoussenko GV. Expression of vesicular transport genes in avisicular cells of microsporidia *Paranosema* (*Antonospora*) *locustae*. Cell tissue biol. 2010;4(2):136-142. doi:10.1134/S1990519X10020033

25. Nakjang S, Williams TA, Heinz E, et al. Reduction and expansion in microsporidian genome evolution: new insights from comparative genomics. Genome Biol Evol. 2013;5(12):2285-2303. doi:10.1093/gbe/evt184

26. Tyerman SD. Nonselective cation channels. multiple functions and commonalities. Plant Physiol. 2002;128(2):327-328. doi:10.1104/pp.900021

27. Bienert GP, Schüssler MD, Jahn TP. Metalloids: essential, beneficial or toxic? Major intrinsic proteins sort it out. Trends Biochem Sci. 2008;33(1):20-26. doi:10.1016/j.tibs.2007.10.004

28. Peyretaillade E, Parisot N, Polonais V, et al. Annotation of microsporidian genomes using transcriptional signals. Nat Commun. 2012;3:1137. doi:10.1038/ncomms2156

29. Liu T. A freeze-etching study on the nuclear envelope during development in microsporidian *Thelohania bracteata* (Strickland, 1913). J Parasitol. 1972. http://www.jstor.org/stable/3278157.

30. Goffeau A, Park J, Paulsen IT, et al. Multidrug-resistant transport proteins in yeast: complete inventory and phylogenetic characterization of yeast open reading frames with the major facilitator superfamily. Yeast. 1997;13(1):43-54. doi:10.1002/(SICI)1097-0061(199701)13:1<43::AID-YEA56>3.0.CO;2-J

31. Nies DH. Efflux-mediated heavy metal resistance in prokaryotes. FEMS Microbiol Rev. 2003;27(2-3):313-339.

32. Conklin DS, McMaster JA, Culbertson MR, Kung C. COT1, a gene involved in cobalt accumulation in *Saccharomyces cerevisiae*. Mol Cell Biol. 1992;12(9):3678-3688.

33. Jack DL, Yang NM, Saier MH. The drug/metabolite transporter superfamily. Eur J Biochem. 2001;268(13):3620-3639. http://www.ncbi.nlm.nih.gov /pubmed/11432728.

34. Tjaden J, Winkler HH, Schwöppe C, Van Der Laan M, Möhlmann T, Neuhaus HE. Two nucleotide transport proteins in *Chlamydia trachomatis*, one for net nucleoside triphosphate uptake and the other for transport of energy. J Bacteriol. 1999;181(4):1196-1202.

35. Russnak R, Konczal D, McIntire SL. A family of yeast proteins mediating bidirectional vacuolar amino acid transport. J Biol Chem. 2001;276(26):23849-23857. doi:10.1074/jbc.M008028200

36. Chardwiriyapreecha S, Mukaiyama H, Sekito T, Iwaki T, Takegawa K, Kakinuma Y. Avt5p is required for vacuolar uptake of amino acids in the fission yeast *Schizosaccharomyces pombe*. FEBS Lett. 2010;584(11):2339-2345. doi:10.1016/j.febslet.2010.04.012

37. Saier MH. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev. 2000;64(2):354-411. doi:10.1128/MMBR.64.2.354-411.2000

38. Cornillot E, Metenier G, Vivares CP, Dassa E. Comparative analysis of sequences encoding ABC systems in the genome of the microsporidian *Encephalitozoon cuniculi*. FEMS Microbiol Lett. 2002;210(1):39-47.

39. Ewart GD, Cannell D, Cox GB, Howells AJ. Mutational analysis of the traffic ATPase (ABC) transporters involved in uptake of eye pigment precursors in *Drosophila melanogaster*. Implications for structure-function relationships. J Biol Chem. 1994;269(14):10370-10377. http://www.ncbi.nlm.nih.gov/pubmed/8144619

40. Alder-Baerens N, Lisman Q, Luong L, Pomorski T, Holthuis JCM. Loss of P4 ATPases Drs2p and Dnf3p disrupts aminophospholipid transport and asymmetry in yeast post-Golgi secretory vesicles. Mol Biol Cell. 2006;17(4):1632-1642. doi:10.1091/mbc.E05-10-0912

41. Wu Z, Li Y, Pan G, et al. A complete Sec61 complex in *Nosema bombycis* and its comparative genomics analyses. J Eukaryot Microbiol. 2007;54(4):379-380. doi:10.1111/j.1550-7408.2007.00272.x

42. Heinz E, Williams TA, Nakjang S, et al. The genome of the obligate intracellular parasite *Trachipleistophora hominis*: New insights into microsporidian genome dynamics and reductive evolution. PLoS Pathog. 2012;8(10):e1002979. doi:10.1371/journal.ppat.1002979

43. Cuomo CA, Desjardins CA, Bakowski MA, et al. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. Genome Res. 2012;22(12):2478-2488. doi:10.1101/gr.142802.112

44. Grisdale CJ, Bowers LC, Didier ES, Fast NM. Transcriptome analysis of the parasite *Encephalitozoon cuniculi*: an in-depth examination of pre-mRNA splicing in a reduced eukaryote. BMC Genomics. 2013;14(1):207. doi:10.1186/1471-2164-14-207

45. Lavine MD, Arrizabalaga G. Exit from host cells by the pathogenic parasite *Toxoplasma gondii* does not require motility. Eukaryot Cell. 2008;7(1):131-140. doi:10.1128/EC.00301-07

46. Saliba KJ, Kirk K. pH Regulation in the intracellular malaria parasite, *Plasmodium falciparum*: H^+ extrusion via a V-type H^+-ATPase. J Biol Chem. 1999;274(47):33213-33219. doi:10.1074/jbc.274.47.33213

47. Undeen AH, Vander Meer RK. Microsporidian intrasporal sugars and their role in germination. J Invertebr Pathol. 1999;73(3):294-302. doi:10.1006/jipa.1998.4834

48. Martín-Hernández R, Higes M, Sagastume S, et al. Microsporidia infection impacts the host cell's cycle and reduces host cell apoptosis. PLoS One. 2017;12(2):e0170183. doi:10.1371/journal.pone.0170183

49. Williams BAP, Elliot C, Burri L, et al. A broad distribution of the alternative oxidase in microsporidian parasites. PLoS Pathog. 2010;6(2):e1000761. doi:10.1371/journal.ppat.1000761

50. Paldi N, Glick E, Oliva M, et al. Effective gene silencing in a microsporidian parasite associated with honeybee (*Apis mellifera*) colony declines. Appl Environ Microbiol. 2010;76(17):5960-5964. doi:10.1128/AEM.01067-10

51. Southern TR, Jolly CE, Russell Hayman J. Augmentation of microsporidia adherence and host cell infection by divalent cations. FEMS Microbiol Lett. 2006;260(2):143-149. doi:10.1111/j.1574-6968.2006.00288.x

52. Kashkarova LF, Khakhanov AI. Range of the hosts of the causative agent of pébrine (*Nosema bombycis*) in the mulberry silkworm. Parazitologiia. 14(2):164-167. http://www.ncbi.nlm.nih.gov/pubmed/6769083.

53. Kudo R, DeCoursey J. Experimental infection of *Hyphantria cunea* with *Nosema bombycis*. J Parasitol. 1940. http://www.jstor.org/stable/3272378.

CHAPTER 7

Summary and Future Prospects

CHAPTER 7

Summary and Future Prospects

The specialty of *Antheraea assamensis* (muga silkworm) is that it can produce golden colored silk and is endemic to Assam and the adjoining hilly areas of Northeast India. We reported here- the transcriptomes of muga silkworm *()* and its two host plants, namely, *Machilus bombycina* (Som) and *Litsea citrata* (Mejankari), for the first time ever. We also presented the changes occurring across the transcriptome of different tissues of muga silkworm with respect to host plant and development. In addition, we constructed a biological database for dissemination of the data generated in this study. Finally, we elucidated the transportome of Nosema, a group of microsporidia, which causes microsporidiosis/pebrine in muga silkworm. A chapter-wise summary is provided below-

7.1 *De novo* transcriptome of *Antheraea assamensis* (muga silkworm)

-In this chapter, we reported the transcriptome of *A. assamensis* (muga silkworm) using high-throughput sequencing of three of its tissues (alimentary canal, silk gland and residual body) from its 5th instar larvae for the first time ever. The de novo strategy for assembly was employed due to lack of a reference whole genome sequence. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were annotated using a combination of

databases- UniProt, NCBI-NR (Non-redundant), Pfam, GO, COG and KEGG. Analysis of the resultant transcriptome lead to identification of differentially expressed candidate genes involved in silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. A set of candidate antimicrobial peptides of *A. assamensis* with antifungal, antibacterial, antiviral and antiparasitic potential were also identified. Finally, the transcriptome was validated by quantitative real-time PCR (qPCR) amplification of eight random candidate transcripts.

7.2 *De novo* transcriptome of two *A. assamensis* host plants: *Machilus bombycina* (som) and *Litsea citrata* (mejankari)

- Som (*Machilus bombycina*) and Mejankari (*Litsea citrata*) are one of the primary and secondary host plants of muga silkworm, respectively. Here, we reported the *de novo* transcriptomes of both these host plants, thus identifying 55,400 and 1,38,690 transcripts, respectively. ~50% transcripts in both the transcriptomes were annotated using a combination of databases (UniProt Viridiplantae, NCBI NR, Pfam, MetaCyc and GO). We also identified the putative transcripts related to plant immune system, namely, glucosinolate-myrosinase pathway (which is a herbivore defense system of plants) and antimicrobial peptides. We were able to identify almost homologs of almost all the enzymes which mediate glucosinolate biosynthesis and activation using long-chain aliphatic and aromatic amino acid precursors in both the host plant transcriptomes. We were also able to identify a myriad of potential peptides with specific- and broad-spectrum antimicrobial activity, chiefly, against bacteria, fungi and virus. Our findings generated a novel resource of sequence data on these two host plants from Lauraceae family and

also provided a foundation for future studies on plant defense for benefit of the sericulture industry.

7.3 Transcriptome profile in *A. assamensis* with respect to host plant and silk gland development

A. assamensis is an economically significant silkworm due to its ability to produce golden silk and its endemic nature. Dearth of sequence information on this species has hindered the scientists and indigenous seri-rearer communities of India for long. In this study, we sequenced the *de novo* transcriptomes of 5th instar larvae of *A. assamensis* reared on two of its host plants, *Litsea citrata* and *Machilus bombycina*. We also sequenced the *de novo* transcriptomes of 4th instar larvae of *A. assamensis* reared on *M. bombycina.* Using the data generated in this study, we reconstructed the transcriptome for *A. assamensis,* identified the top most expressed transcripts in each tissue and observed how biological processes associated with each tissue varies with respect to host plant and larval development (4th instar and 5th instar). We found that translation was the most unanimous process expressed in each tissue of silk gland of *A. assamensis* regardless of the developmental stage or host plant. Other than this process, other processes like oxidative stress management, redox homeostasis, transcriptional regulation, etc. had variable representation across different stages. Analysis of these patterns showed how the transcriptional profile of *A. assamensis* silk gland can vary in different anatomical sections. Analysis of specific pathways like silk synthesis, ecdysteroid and juvenile hormone synthesis are underway as a follow-up of this study.

7.4 MugaSeqDB, a database on *A. assamensis* and its associated host plants

-Here, we developed MugaSeqDB, a comprehensive, freely accessible database hosting the transcriptome data of muga silkworm (*A. assamensis)* and its two host plants, Som *(M. bombycina)* and Mejankari (*L. citrata)*. This database hosts transcripts, predicted proteins, their respective functional and ontological annotations for these three species. Additionally, it provides secondary information on pest, pathogen and patents of the muga silkworm and its host plants. A combination of MySQL and phpMyAdmin was utilized to develop its back-end while the front end was created using a combination of HTML, php and java scripts. The complete architecture was hosted at a Linux-based commercial server. Features like search, browse, download, secondary database cross-linking, patent information, informative help pages and scope for user data submission has been incorporated in MugaSeqDB. The ultimate goal of this database will be to perform as a one-stop database for information on muga silkworm (*A. assamensis*) and other species associated with it. This database is now available online as an open-access resource at http://mugaseqdb.in.

7.5 Comparative transportome of Nosema, the causal organism of pebrine

-*Nosema* is a genus of intracellular parasitic microsporidia which causes pebrine disease in arthropods, including economically important silkworms and honeybees. *Nosema* have gene-poor genomes shaped by loss of the metabolic pathways, as a consequence of continued dependence on host-derived

substrates. As an act of counterbalance, they have developed an array of transporter proteins that allow stealing from their hosts. In the present study, we predicted the putative transportomes of four Nosema species, viz. *Nosema apis*, *N. bombycis*, *N. ceranae* and *N. antheraea.* Our results indicated that the transportomes of *N. apis, N. bombycis, N. ceranae* and *N. antheraea* have a dominant share of secondary carriers and primary active transporters. The comparatively rich and diverse transportome of *N. bombycis* indicates the role of transporters in its remarkable capability of host adaptation. We identified a set of twelve transporter families core to the *Nosema* genus with possible role in osmo-regulation, intra- and extra-cellular pH regulation, energy compensation and self-defense mechanism. We also identified a set of ten species-specific transporter families within Nosema which may be involved in species-specific host adaptations. Both the core and species-specific transporter proteins of Nosema constituted a valuable resource that will come handy in development of inhibitor-based *Nosema* management strategies in future and thereby, help the sericulture and apiculture scenario of the industrial world.

Future Prospects:

- The whole genome sequencing of *A. assamensis, M. bombycina* and *L. citrata* will be able to generate an even greater genomic resource for potential seribiologists. Other host plant resources for muga silkworm, namely, Soalu (*L. polyantha)* and Dighloti (*L. salicifolia*) are also apt for whole genome sequencing.

- An improved MugaSeqDB v2.0 will be published with enriched genomic information, once the whole genome sequences of the three species or any other related species becomes available.

- Similarly, other pests and pathogen of this silkworm should be sequenced to identify their weaknesses and exploit those for biological control. The transporter proteins identified in Nosema must be functionally validated followed by knock-out studies to test for possible molecular targets for pebrine management in commercial insects.

- Domestication of *A. assamensis* will be a breakthrough for not only the sericulture industry but also biomedical and other basic biological sciences.

- Application of our transcriptome data for improving the yield and nutritional value of the host plants' foliage can prove to be beneficial to the sericulture industry as well.

Dr. Hasnahana Chetia

Bioengineering Research Laboratory (BERL),
Dept. of Biosciences and Bioengineering,
Indian Institute of Technology Guwahati, India-781039

Personal Profile

Date of birth	6th February, 1989
Gender	Female
Nationality	Indian
Languages	English, Assamese, Hindi and Spanish
Marital status	Unmarried
Present Address	Bioengineering Research Laboratory (BERL) Department of Biosciences and Bioengineering O-Block, Academic Complex, Indian Institute of Technology Guwahati (IITG) Guwahati-781039, Assam, India
Email	hasnahana@iitg.ac.in, hasche1989@gmail.com
Phone no:	+91 801 101 4417

Education

Degree	University/Institute	Specialization	Year	Marks	Division
Doctor of Philosophy	Indian Institute of Technology Guwahati, Assam, India	Biosciences and Bioengineering	2014-2019	-	-
Master of Science	Gauhati University, Guwahati, Assam, India	Biotechnology	2010-2012	8.94/10	First
Bachelor of Science	North Eastern Hill University, Shillong, Meghalaya, India	Biotechnology	2007-2010	7.2/10	First

Research experience:

- Doctoral research on Genomics and Transcriptomics at the Bioengineering Research Laboratory (BERL), Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati
  - Doctoral Supervisor: Prof. Utpal Bora
  - Thesis title: Understanding molecular aspects of *Antheraea assamensis*

- Bioinformatics Trainee at the Bioinformatics Infrastructure Facility (BIF), Dept. of Zoological Sciences, Gauhati University, Assam, India
    - Supervisor: Prof. Dhirendra Kumar Sharma
    - Key subject areas: Molecular docking and homology modelling of key proteins related to pathogenesis, ageing and nerve growth.

- Master's Dissertation at the Department of Biotechnology, Gauhati University, Assam, India
    - Dissertation Supervisor: Prof. Rupjyoti Bharali
    - Dissertation Title: A comparative study of antioxidant and phytochemical content in the aqueous and ethanolic extracts of an indigenous fruit of North-East India, *Phyllanthus acidus.*

Academic Responsibilities

- Teaching and Laboratory Assistant (TA) for BT 503 (Advanced Genetic Engineering) [Instructors- Prof. Utpal Bora and Prof. Ranjan Tamuli, Dept. of BSBE, IITG]

List of Publications (in chronological order)

- Kabiraj D, Chetia H, Nath A, Sharma P, Singh D, Mosahari PV, Dutta P, Neog K and Bora U. 2019. Mitogenome-wise codon usage pattern from comparative analysis of the first mitogenome of *Blepharipa* sp. (Muga uzifly) with other Oestroid flies *(submitted).*

- **Chetia H,** Kabiraj D, Bharali B, Ojha S, Barkataki MP, Saikia D, Singh T, Mosahari PV, Sharma P, Bora U. 2019. Exploring the benefits of endophytic fungi via omics, p. 51–81. In. Springer, Cham.

- **Chetia H**, Kabiraj D, Singh D, Mosahari PV, Das S, Sharma P, Neog K, Sharma S, Jayaprakash P, Bora U. 2017. *De novo* transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer). Gene 611.

- Singh D, Kabiraj D, Sharma P, **Chetia H**, Mosahari PV, Neog K, Bora U. 2017. The mitochondrial genome of muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. PLoS One 12.

- **Chetia H**, Kabiraj D, Sharma S, Bora U. 2017. Comparative insights to the transportome of Nosema: a genus of parasitic microsporidians. bioRxiv 110809.

- Kabiraj D, Kalita J, **Chetia H**, Singh D, Bora U. 2015. Expanding the

frontiers of rice research through omics. Assam Sc. Soc. Vo. p. 1-28.

- Kumar A, **Chetia H**, Sharma S, Kabiraj D, Talukdar NC, Bora U. 2015. Curcumin Resource Database. Database 2015:bav070.

- Singh D, **Chetia H**, Kabiraj D, Sharma S, Kumar A, Sharma P, Deka M, Bora U. 2016. A comprehensive view of the web-resources related to sericulture. Database 2016:baw086.

- Ojha S, Singh D, Sett A, **Chetia H**, Kabiraj D, Bora U. 2018. Nanotechnology in Crop Protection. Nanomater Plants, Algae, Microorg 345–391.

- Mosahari PV, Singh D, Kalita JJ, Sharma P, **Chetia H**, Kabiraj D, Mahanta C, Bora U, Singh D, Kalita JJ, Sharma P, Chetia H, Kabiraj D, Mahanta C, Bora U. 2018. Nanotoxicity: Impact on Health and Environment, p. 21–46. In Environmental Toxicity of Nanomaterials. CRC Press.

- **Chetia H**, Sarma R, Verma A, Kumar Sharma D. 2014. Comparative modelling, characterization and molecular dynamics study of trypanothione reductase from *Leishmania donovani*. Int J Sci Eng Res 5.

- **Chetia H**, Kumar Sharma D, Sarma R, Verma A. 2014. An *in silico* approach to discover potential inhibitors against multi-drug resistant bacteria producing New-Delhi metallo-β-lactamase 1 (NDM-1) enzyme. Int J Pharm Pharm Sci 6:299–303.

- Sarma R, Verma A, **Chetia H**, Sharma DK. 2013. 3D structure prediction of aging related proteins of *Silurana tropicalis* Gray, 1864. J Pharm Res 7:762–765.

- Verma A, Sharma DK, Sarma R, **Chetia H**, Saikia J. 2013. Comparative insights using the molecular homology model of BDNF (Brain-Derived Neurotrophic Factor) of *Varanus komodoensis* and the known NGF (Nerve Growth Factor) structure of *Naja atra*. Bioinformation 9:755–8.

Awards and Fellowships

- Qualified the Graduate Aptitude Test in Engineering (GATE) in Biotechnology 2013 with All India Rank- 504 (Fellowship availed from Ministry of Human Resource Development (MHRD) of India from 2014-2019).

- Qualified the Graduate Aptitude Test in Engineering (GATE) in Biotechnology 2012 with All India Rank- 453 (Fellowship not availed).

- Awarded partial grant to attend NextGen Genomics, Biology, Bioinformatics and Technologies (NGBT) Conference in Bhubaneswar, Odisha, India.

Workshops

- Workshop on "Understanding Human Disease and Improving Human Health Using Genomics-Driven Approaches" co-organized by National Institute of Biomedical Genomics (NIBMG), Kalyani & Tezpur University at Tezpur University from 9th to 13th May, 2016.

- Advanced Workshop on "Understanding Human Disease and Improving Human Health Using Genomics-Driven Approaches" at National Institute of Biomedical Genomics (NIBMG), Kalyani, West Bengal from 27th February to 10th March, 2017.

- Hands-on Bioinformatics Workshop on "Biological Data Analysis using R-Statistics package" at CSIR-North East Institute of Science and Technology (CSIR-NEIST) Jorhat from 5th to 6th January, 2016.

- "North-East Winter School on Human Genetics 2016-Genetic Analyses of Complex Traits" organized jointly by Dibrugarh University and Indian Statistical Institute Kolkata from 21st to 22nd December, 2016.

- BIOCONVERSE 2018 (One day workshop of Wildlife ecology and Seri bioresources) organized by Directorate of Sericulture (BTC) & College of Veterinary Sciences (AAU), Khanapara at Manas National Park & BTAD, Assam from 30th Jan- 01st Feb 2018.

Conferences, Seminar and Symposia

- "International symposium on biodiversity and biobanking (BIODIVERSE) 2018" with Association for Promotion of DNA Fingerprinting and Other DNA Technologies (ADNAT) at IIT Guwahati from 27th to 29th January, 2018.

- One-day Capacity Building Workshop-cum-Brainstorming Meeting on "River ecosystems and fresh water biodiversity research (REFRESH) 2018" at IIT Guwahati on 2nd February, 2018.

- "Nextgen Genomics, Biology, Bioinformatics and Technologies Conference (NGBT 2017)" at Bhubaneswar, Odisha from 02nd to 4th October, 2017. (Partial grant awarded)

- Conference on "Exploitation of Seribiodiverity for Novel Product Development" at IIT Guwahati from 29th to 30th November, 2014.

Gene 611 (2017) 54–65

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

De novo transcriptome of the muga silkworm, *Antheraea assamensis* (Helfer)

Hasnahana Chetia [a], Debajyoti Kabiraj [a], Deepika Singh [a], Ponnala Vimal Mosahari [b], Suradip Das [a], Pragya Sharma [c], Kartik Neog [d], Swagata Sharma [a], P. Jayaprakash [e], Utpal Bora [a,b,*]

[a] Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam 781039, India
[b] Centre for the Environment, Indian Institute of Technology Guwahati, Assam 781039, India
[c] Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati 781014, Assam, India
[d] Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, 785700 Jorhat, Assam, India
[e] Central Silk Board (CSB), Bangalore 506068, Karnataka, India

ARTICLE INFO

Article history:
Received 20 October 2016
Received in revised form 29 January 2017
Accepted 15 February 2017
Available online 17 February 2017

Keywords:
Alimentary canal
Antimicrobial peptide
Lepidoptera
Machilus bombycina
Next generation sequencing
Residual body
Saturniidae
Sericin
Silk gland
Silk gland factor

ABSTRACT

Antheraea assamensis (Lepidoptera: Saturniidae), is a semi-domesticated silkworm known to be endemic to Assam and the adjoining hilly areas of Northeast India. It is the only producer of a unique, commercially important variety of golden silk called "muga silk". Herein, we report the *de novo* transcriptome of *A. assamensis* reared on *Machilus bombycina* leaves for the first time. Short reads generated by high throughput sequencing of cDNA libraries from multiple tissues, viz. alimentary canal, silk gland and residual body of the 5th instar of muga silkworm were assembled into transcripts via a *de novo* assembly pipeline followed by functional annotation and classification. A total of 1,21,433 transcripts were generated from ~231 million raw reads of which ~74% (89,583) were either allocated a functional annotation or categorized under Pfam/COG/KEGG categories. Identification of differentially expressed transcripts and their comparative sequence analysis revealed candidate genes related to silk synthesis, viz. silk gland factor-1 and 3, sericin-like transcript, etc. with conserved forkhead, homeo- and POU domains. Several candidate anti-microbial peptides which may have potential anti-bacterial, anti-fungal or anti-parasitic activity in *A. assamensis* were also identified. T/A and AT/TA were predicted to be the most abundant mono- and di-nucleotide simple sequence repeat markers in the transcriptome. Transcriptome validation was carried out by quantitative real-time PCR (qPCR) amplification of eight transcripts. The resources generated by this study will expand the periphery of existing genomic data on *A. assamensis* facilitating future in-depth studies on its unknown aspects.

1. Introduction

Silk is an important cultural and commercial fibre largely obtained from silkworms. Some silkworms have been domesticated by humans over a period of time to exploit their potential in textiles. Still, majority of them remain semi-domesticated or wild. The silk proteins- fibroin and sericin of domesticated mulberry silkworm, *Bombyx mori* (Family: Bombycidae) has been extensively used for tissue engineering applications (Das et al., 2014). Research on these has been further accelerated by discovery of its complete genome and transcriptome (Li et al., 2012; The International Silkworm Genome Consortium, 2008). Similar as well as novel research applications can also be expected from biomaterials of other less-studied semi- or undomesticated silkworms. The dearth of information and hindrances in domestication of these silkworms currently restricts their usage in such applications. One such semi-domesticated silkworm is *Antheraea assamensis* (Helfer), also known as the "muga silkworm". *A. assamensis* ($n = 15$) is a multivoltine, polyphagus silkworm classified under the order - Lepidoptera and family - Saturniidae. It is mostly endemic to the Brahmaputra valley of Assam and adjoining hilly areas of Northeast India (Tikader et al., 2013). Being the sole producer of globally acclaimed "muga silk", a lustrous golden yellow fabric, makes *A. assamensis* one of the most important components of the Assamese silk industry and it hugely contributes towards employment generation in North-Eastern India. The unique quality of this silk-based textile earned it a geographical indication tag

Abbreviations: AaCbp, *Antheraea assamensis* Carotenoid binding protein; AaFhc, *Antheraea assamensis* Fibroin heavy chain; AC, Alimentary Canal; AMP, Anti-microbial peptides; CEG, Core eukaryotic genes; COG, Cluster of Orthologous genes; EST, Expressed Sequence Tag; GO, Gene Ontology; KEGG, Kyoto Encyclopaedia of Genes and Genomes; MISA, MIcroSAtellite Identification Tool; PNTAa, Putative novel transcripts of *Antheraea assamensis*; POU, Pit-Oct-Unc; qPCR, Quantitative real time PCR; RB, Residual body; SG, Silk gland; SRA, Short Read Archive; SSR, Simple sequence repeats.

* Corresponding author at: Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India.
E-mail addresses: ubora@iitg.ernet.in, drutpalbora@gmail.com (U. Bora).

http://dx.doi.org/10.1016/j.gene.2017.02.021
0378-1119/

Comparative insights to the transportome of *Nosema*: a genus of parasitic microsporidians

Hasnahana Chetia[1,§], Debajyoti Kabiraj[1,§], Swagata Sharma[1,§], Utpal Bora[1, 2*]

[1] Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam-781039, India.

[2] Institutional Biotech Hub, Centre for the Environment, Indian Institute of Technology Guwahati (IITG), Assam 781039, India

Corresponding Author

Prof. Utpal Bora, Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam – 781039, India.

Email: ubora@iitg.ernet.in

Phone : +913612582215, Fax: +913612582249

§- These authors contributed equally to this work.

Abstract

Nosema, a genus of parasitic microsporidia, causes pebrine disease in arthropods, including economically important silkworms and honeybees. *Nosema* have gene-poor genomes shaped by loss of the metabolic pathways, as a consequence of continued dependence on host-derived substrates. As an act of counterbalance, they have developed an array of transporter proteins that allow stealing from their hosts. Here, we have identified the core set of twelve transporter families present in *Nosema* genus, viz. *N. apis*, *N. bombycis*, *N. ceranae* and *N. antheraea* through *in silico* pipeline. Transportomes of *N. apis, N. bombycis, N. ceranae* and *N. antheraea* have a dominant share of secondary carriers and primary active transporters. The comparatively rich and diverse transportome of *N. bombycis* indicates the role of transporters in its remarkable capability of host adaptation. The core set of transporter families of *Nosema* includes ones that have a likely role in osmo-regulation, intra- and extra-cellular pH regulation, energy compensation and self-defence mechanism. This study has also revealed a set of ten species-specific transporter families within the genus. To our knowledge, this is the first ever intra-genus study on microsporidian transporters. Both these datasets constitutes a valuable resource that can aid in development of inhibitor-based *Nosema* management strategies.

Keywords: honeybee, Nosemosis, *Nosema apis*, *Nosema bombycis*, *Nosema ceranae*, *Nosema antheraea*, pebrine, silkworm, transporter, TCDB

[1] Introduction

Microsporidia is a specific group of unicellular obligate parasites or hyperparasites which can infect a myriad of organisms including a few economically important insects like silkworms, honey bees etc. as well as humans **(Silveira *et al.*, 2009; Weiss and Becnel, 2014)**. They generally have highly reduced gene-dense genomes achieved via evolutionary sacrifice of the genes of many essential pathways (TCA cycle, metabolic pathways) to minimize biological complexity **(Keeling and Corradi, 2011; Katinka *et al.*, 2001)**. Microsporidia possess an unstacked Golgi apparatus and a cryptic genome-less mitochondria called mitosome. In order to compensate for their reduced metabolic capacity, they utilize the host's inner metabolism to derive nutrition. They usually do this either by up-regulating the host's metabolic pathways via microsporidia-secreted factors and taking up nutrients from the host by their membrane-bound

Chapter 4
Exploring the Benefits of Endophytic Fungi via Omics

Hasnahana Chetia, Debajyoti Kabiraj, Biju Bharali, Sunita Ojha, Manash Pratim Barkataki, Dharitri Saikia, Tinka Singh, Ponnala Vimal Mosahari, Pragya Sharma, and Utpal Bora

4.1 Introduction

4.1.1 What Are Endophytic Fungi?

AU1 AU2 The term "endophyte" stems from the two Greek words—*endon* = within and *phyte* = plant. Thus, an endophyte is an organism surviving within a host plant without deleterious consequences. In 1809, a German botanist, Johann Heinrich Friedrich Link termed them as "Entophytae" which constituted of moderately parasitic species living within plants in 1809. In 1991, Orlando Petrini re-termed these organisms as "endophytes". However, modern studies have shown that endophytes colonize plant tissues without detriment to the host and at most times, rather acts as a benefactor. An endophyte may be bacteria, fungi, algae or oomycetes. Here, we will focus on endophytic fungi only.

Endophytic fungi are a diverse group of fungi living a primarily asymptomatic life within almost every terrestrial plant lineage available in natural and anthropogenic habitats. Fossil records indicate that most of the plants, if not all, have been associated with endophytic fungi for >400 myr (million years). They are capable of

H. Chetia · D. Kabiraj · B. Bharali · S. Ojha · M. P. Barkataki · D. Saikia · T. Singh · P. V. Mosahari
Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam, India

P. Sharma
Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati, Assam, India

U. Bora (✉)
Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam, India

Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam, India

B. P. Singh (ed.), *Advances in Endophytic Fungal Research*, Fungal Biology,
https://doi.org/10.1007/978-3-030-03589-1_4

PLOS ONE

RESEARCH ARTICLE

The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects

Deepika Singh[1,2], Debajyoti Kabiraj[1], Pragya Sharma[3], Hasnahana Chetia[1], Ponnala Vimal Mosahari[2], Kartik Neog[4], Utpal Bora[1,2]*

1 Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Assam, India, **2** Centre for the Environment, Indian Institute of Technology Guwahati, Assam, India, **3** Department of Bioengineering and Technology, Gauhati University Institute of Science and Technology (GUIST), Gauhati University, Guwahati, Assam, India, **4** Biotechnology Section, Central Muga Eri Research & Training Institute (CMER&TI), Lahdoigarh, Jorhat, Assam, India

* ubora@iitg.ernet.in, ubora@rediffmail.com

Citation: Singh D, Kabiraj D, Sharma P, Chetia H, Mosahari PV, Neog K, et al. (2017) The mitochondrial genome of Muga silkworm (*Antheraea assamensis*) and its comparative analysis with other lepidopteran insects. PLoS ONE 12(11): e0188077. https://doi.org/10.1371/journal.pone.0188077

Editor: Daniel Doucet, Natural Resources Canada, CANADA

Received: June 28, 2017

Accepted: October 31, 2017

Published: November 15, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The full annotated mitogenome sequence and SRA data of A. assamensis submitted to NCBI GenBank are available under the accession numbers KU379695 and SRR3948351, respectively.

Funding: The authors thank the Department of Biotechnology, Govt. of India, New Delhi for supporting the research through the UXCEL project

Abstract

Muga (*Antheraea assamensis*) is an economically important silkmoth endemic to the states of Assam and Meghalaya in India and is the producer of the strongest known commercial silk. However, there is a scarcity of genomic and proteomic data for understanding the organism at a molecular level. Our present study is on decoding the complete mitochondrial genome (mitogenome) of *A. assamensis* using next generation sequencing technology and comparing it with other available lepidopteran mitogenomes. Mitogenome of *A. assamensis* is an AT rich circular molecule of 15,272 bp (A+T content ~80.2%). It contains 37 genes comprising of 13 protein coding genes (PCGs), 22 tRNA and 2 rRNA genes along with a 328 bp long control region. Its typical $tRNA^{Met}$-$tRNA^{Ile}$-$tRNA^{Gln}$ arrangement differed from ancestral insects ($tRNA^{Ile}$-$tRNA^{Gln}$-$tRNA^{Met}$). Two PCGs *cox1* and *cox2* were found to have CGA and GTG as start codons, respectively as reported in some lepidopterans. Interestingly, *nad4l* gene showed higher transversion mutations at intra-species than inter-species level. All PCGs evolved under strong purifying selection with highest evolutionary rates observed for *atp8* gene while lowest for *cox1* gene. We observed the typical clover-leaf shaped secondary structures of tRNAs with a few exceptions in case of $tRNA^{Ser1}$ and $tRNA^{Tyr}$ where stable DHU and TΨC loop were absent. A significant number of mismatches (35) were found to spread over 19 tRNA structures. The control region of mitogenome contained a six bp (CTTAGA/G) deletion atypical of other *Antheraea* species and lacked tandem repeats. Phylogenetic position of *A. assamensis* was consistent with the traditional taxonomic classification of Saturniidae. The complete annotated mitogenome is available in GenBank (Accession No. KU379695). To the best of our knowledge, this is the first report on complete mitogenome of *A. assamensis*.

DATABASE
The Journal of Biological Databases and Curation

Database, 2015, 1–6
doi: 10.1093/database/bav070
Original article

Original article

Curcumin Resource Database

Anil Kumar[1,2], Hasnahana Chetia[1], Swagata Sharma[1], Debajyoti Kabiraj[1], Narayan Chandra Talukdar[3,*] and Utpal Bora[1,4,*]

[1]Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati (IITG), Assam 781039, India, [2]Centre for Biological Sciences (Bioinformatics), Central University of South Bihar (CUSB), Patna 800014, India, [3]Institute of Advanced Studies on Science and Technology (IASST) Boragaon, Guwahati, Assam 781035, India and [4]Institutional Biotech Hub, Centre for the Environment, Indian Institute of Technology Guwahati (IITG), Assam 781039, India

*Corresponding author: Tel: +91 361 2582215; Fax: +91 361 2582249; Email: ubora@iitg.ernet.in, ubora@rediffmail.com

Correspondence may also be addressed to Narayan Chandra Talukdar. Tel: +91 361 2273058; Fax: +91 361 2273062; Email: nctalukdar@yahoo.com

Citation details: Kumar,A., Chetia,H., Sharma,S., *et al.* Curcumin Resource Database. *Database* (2015) Vol. 2015: article ID bav070; doi:10.1093/database/bav070

Received 29 April 2015; Revised 10 June 2015; Accepted 26 June 2015

Abstract

Curcumin is one of the most intensively studied diarylheptanoid, *Curcuma longa* being its principal producer. This apart, a class of promising curcumin analogs has been generated in laboratories, aptly named as Curcuminoids which are showing huge potential in the fields of medicine, food technology, etc. The lack of a universal source of data on curcumin as well as curcuminoids has been felt by the curcumin research community for long. Hence, in an attempt to address this stumbling block, we have developed Curcumin Resource Database (CRDB) that aims to perform as a gateway-cum-repository to access all relevant data and related information on curcumin and its analogs. Currently, this database encompasses 1186 curcumin analogs, 195 molecular targets, 9075 peer reviewed publications, 489 patents and 176 varieties of *C. longa* obtained by extensive data mining and careful curation from numerous sources. Each data entry is identified by a unique CRDB ID (identifier). Furnished with a user-friendly web interface and in-built search engine, CRDB provides well-curated and cross-referenced information that are hyperlinked with external sources. CRDB is expected to be highly useful to the researchers working on structure as well as ligand-based molecular design of curcumin analogs.

Database URL: http://www.crdb.in

Introduction

Curcumin (diferuloylmethane) is a hydrophobic polyphenol derived from rhizome of the perennial herb turmeric (*Curcuma longa*) which belongs to the ginger family (Zingiberaceae) native to tropical South Asia (1). Numerous traditional usage of turmeric is described in

DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–31
doi: 10.1093/database/baw086
Review

Review

A comprehensive view of the web-resources related to sericulture

Deepika Singh[1], Hasnahana Chetia[1], Debajyoti Kabiraj[1], Swagata Sharma[1], Anil Kumar[2], Pragya Sharma[3], Manab Deka[3] and Utpal Bora[1,4,5,*]

[1]Bioengineering Research Laboratory, Department of Biosciences and Bioengineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India, [2]Centre for Biological Sciences (Bioinformatics), Central University of South Bihar (CUSB), Patna 800014, India, [3]Department of Bioengineering & Technology, Gauhati University Institute of Science & Technology, Gauhati University, Guwahati, Assam 781014, India, [4]Centre for the Environment, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India and [5]Mugagen Laboratories Pvt. Ltd, Technology Incubation Centre, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

*Corresponding Author: Tel: +913612582215; Fax: +913612582249; Email: ubora@iitg.ernet.in; ubora@rediffmail.com

Citation details: Singh,D., Chetia,H., Kabiraj,D. *et al.* A comprehensive view of the current web-resources in sericulture and related fields. *Database* (2016) Vol. 2016: article ID baw086; doi:10.1093/database/baw086

Received 21 January 2016; Revised 25 April 2016; Accepted 2 May 2016

Abstract

Recent progress in the field of sequencing and analysis has led to a tremendous spike in data and the development of data science tools. One of the outcomes of this scientific progress is development of numerous databases which are gaining popularity in all disciplines of biology including sericulture. As economically important organism, silkworms are studied extensively for their numerous applications in the field of textiles, biomaterials, biomimetics, etc. Similarly, host plants, pests, pathogens, etc. are also being probed to understand the seri-resources more efficiently. These studies have led to the generation of numerous seri-related databases which are extremely helpful for the scientific community. In this article, we have reviewed all the available online resources on silkworm and its related organisms, including databases as well as informative websites. We have studied their basic features and impact on research through citation count analysis, finally discussing the role of emerging sequencing and analysis technologies in the field of seri-data science. As an outcome of this review, a web portal named SeriPort, has been created which will act as an index for the various sericulture-related databases and web resources available in cyberspace.

Database URL: http://www.seriport.in/